Metabarcoding and metagenomic approaches to decipher microbial communities in suboxic environments

Guillaume Reboul

Thèse de doctorat

NNT: 2020UPASL041

Metabarcoding and metagenomic approaches to decipher microbial communities in suboxic environments

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°577 Structure et dynamique des systèmes vivants (SDSV)

Spécialité de doctorat: Sciences de la vie et de la santé

Unité de recherche: Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique et Evolution, 91405, Orsay, France.

Référent: : Faculté des sciences d'Orsay

Thèse présentée et soutenue en visioconférence totale, le 27 Novembre 2020, par

Guillaume REBOUL

Composition du jury:

Olivier Lespinet Professeur, Université Paris-Saclay (UMR 9198)	Président
Emmanuelle Gérard Ingénieure de Recherche HDR, Institut De Physique Du Globe De Paris (IPGP)	Rapportrice
Rohit Ghai Directeur de recherche, Institute of Hydrobiologie (CAS), Tchéquie	Rapporteur
Hélène Agogué Chargée de Recherche, CNRS - Université de La Rochelle (UMR 7266)	Examinatrice
Loïs Maignien Maître de conférences, CNRS - Université de Bretagne Occidentale (UMR 6197)	Examinateur
Purificación Lopez-Garcia Directrice de recherche, CNRS - Université Paris-Sud (UMR 8079)	Directrice de thèse
David Moreira Directeur de recherche, CNRS - Université Paris-Sud (UMR 8079)	Co-Directeur de thèse

ACKNOWLEDGEMENTS

I'm very much aware of the fact that writing the thesis is not the end of the process to became PhD and I will still need to pass the (confined) upcoming PhD thesis defense but I would like to thank many humans being for the past four years anyway.

Jury members

First of all, I would like to thank all the jury members for their time and especially the ones who accepted to review my thesis. In that matter, thank you **Emmanuelle Gérard and Rohit Ghai** as well as **Olivier Lespinet, Hélène Agogué and Lois Maignien**.

Also, thank you **Laura Eme** for being the technical support of the thesis *via* visio at the last minute.

The bosses

Of course, I would like to thank **Puri and David** for the opportunity of doing the PhD in their team. I remember applying for the Dallol subject and losing to Jodie and how I was very surprised that you offered me to join anyway on your own funding because you believed I could be an asset for the lab. Also, during my PhD you gave me the opportunity to participate in a sampling trip,

do molecular biology, manage students, participate at international conferences and this means a lot to me. Even if things have sometimes been complicated, I am grateful for what you have done for me. Spending 4 years in your team has been a rewarding experience and allowed me to have my second taste of science and research after my sandwich course at CEA. I opened my mind and learned more about the field of research and met many interesting and nice people. I hope you are not or not too disappointed about this decision to hire me and I wish you the best scientific future!

Office 208

Naturally, people from the office 208 come right after. I spent at least 2 years with all of you and 4 with some. These were very nice years made of laughs, help, discussion and parties.

The "old guys" **Guifre "papito" and Rafael**, thanks for welcoming me and the others and for helping us with initial paperwork and advice on how to handle the thesis. Good luck for your promising scientific careers!

The "new guys" **Fabian and Thomas**, we had very nice 2 years together. You are both very kind and I enjoyed hanging out with you very much. **Thomas**, your generosity is outstanding and I hope that your PhD video will be as good as ours because you deserve it. Thanks for being the scriptwriter and movie director of the lab lately. **Fabian**, I know it's not always been easy for you but I'm sure you gonna get stronger through this! Keep the socks on, these are your Samson's hair! Thanks also for the poop experiment which ended in a mysterious way but was very funny !

Then, the other 3 I spent 4 years with.

Gwendou, thanks a lot for all daily help and support and also great food! Thanks for having participated in opening my mind on numerous subjects through critical discussions. Thanks for your patience at home and at work and I really believe your hard work will pay off soon!

Jodie, it all started by a rivalry and you won the Dallol subject and experienced many good and sometimes bad things with it. Anyway I was glad I could participate to your thesis as a collaborator still and hope you don't only keep in mind the headaches about learning Bash, Python or Awk. For sure I'll also remember the dramas and all other epic situations you put yourself into!

Luis, I would never thank you enough for when you stepped up for us on the boat. You were

there when I needed and we had crazy moments in the office, which were also sometimes very needed! I'm really proud of you and what you have accomplished and are still accomplishing as a researcher. You have successfully become a great mature scientist and I wish you all the best for the future.

Other DEEM/ESE members

Naoji, thanks a lot for your kindness and your smile. Our chess and shogi sessions as well as the petanque games we played are very good memories for me! You truly put yourself out there to learn about French culture! I'm very happy to have met you and I hope you will be happy back in Vancouver! For sure, I'm gonna visit you and keep trying to teach you how to use your mouth for noises/sounds!

Miguel, je te suis reconnaissant pour tellement de choses et la plupart sont si abstraites que c'est difficile de les retranscrire. Je tiens simplement à te remercier pour tous tes conseils, ton temps pour discuter ou m'aider ou encore ta patience à m'expliquer plusieurs fois les mêmes concepts biologiques. Tu as été très important dans les derniers mois de cette thèse et vraiment, merci beaucoup pour cela. Je te souhaite plein de réussite dans ta vie Parisienne !

Ludwig, nous avons beaucoup ri dans le bureau "d'en bas" quand tu croyais encore que tu étais jeune, avant donc que tu n'aménages à l'étage. Ces moments m'ont vraiment permis de me sentir bien au sein de l'équipe et je te remercie pour cela. Merci aussi pour ta gentillesse et le temps que tu passes pour les autres quand ils en ont vraiment besoin. J'ai eu plusieurs fois besoin de discuter et j'ai pu te trouver sans soucis avec une vision bienveillante.

Pao, je ne sais pas si tu sais à quel point j'admire ton dévouement pour l'équipe et le laboratoire. Tu es pour moi une clé de voûte encore indispensable au bon déroulement de nombreux projets. Ta nature généreuse est un atout formidable dans le monde sans merci de la recherche scientifique. Merci encore et bravo !

Seif, je suis vraiment content de t'avoir rencontré, merci de m'avoir motivé pour le rugby. Tu es une des personnes les plus agréables et souriantes qu'il m'a été possible de rencontrer dans ma vie. J'espère que tout ira bien pour toi et continu de répandre ta bonne humeur et ta joie de

vivre.

I also want to thank **Philippe and Ana** for technical discussions and being the local bioinformatician help when needed. Good luck to **Britt and Jasmin** for your PhD and thanks for the moments we had together already. I hope this COVID pandemia will end so you can truly enjoy this period. And thank you **Maria** for being there when I have needed to talk, and for being so nice with the people around you. You are on the track of becoming the next anchor of the team!

External Collaborators

Some of the work in this thesis was done in collaboration with external (from the DEEM team) scientists. I would like to thank **Albert** for his patience and sympathy and **Natasha** who invited the team to sample lake Baikal and managed all the administrative part with the Russian government. This was my first sampling trip and also a very nice opportunity for me to travel to Russia for the first time !

Administrative help

Cette thèse aurait sûrement été plus compliquée voire n'aurait pas eu lieu sans l'aide administrative de **Nathalie Lecat, Emmanuelle Jestin et Sandrine Le Bihan**. Merci à toutes les trois pour votre patience et vos réponses rapides à mes angoisses administratives !

PUC(ANAM)

À tous les PUCistes, vous n'avez jamais participé à ma thèse scientifiquement parlant mais faire partie du club et m'éclater à chaque week-end de championnat, à l'entraînement ou sur WhatsApp avec vous m'a fait beaucoup de bien dans les moments les plus compliqués de cette thèse. Merci particulièrement à certains d'entre vous, qui m'ont aidé, consolé, appris et fait confiance tant au floorball que pour la vie en général et avec lesquels j'ai aussi passé de vrais bons moments d'amitié en dehors des terrains ! Donc merci **Eric, JE, Sid, Fab, Loïk, Clément, Patrick, Toko, Oliv, Marie, Charlène, JB, Quentin, les Juliens AL** et j'en oublie sûrement et je m'en excuse par avance ! Je ne sais pas encore trop de quoi mon avenir sera fait, mais une chose est sûre, le PUC ne sera jamais oublié et j'y resterai aussi longtemps que possible en tant que

joueur, entraîneur, membre du bureau ou simple bénévole tant que cette ambiance et cet esprit d'équipe continuera à y régner !

Les amis

Je ne peux pas ne pas faire un petit mot pour vous les Bros du master. Ces 4 dernières années ont été mouvementées pour beaucoup d'entre nous et ça ne nous a pas permis de nous voir souvent mais nos liens sont solides et bien entretenus par de bons moments via les messages, les appels ou les évènements ponctuels ! Savoir que vous êtes tous, là, prêt à aider, discuter, réconforter est un réel plus pour des épreuves-périodes comme une thèse ! Merci donc **Charlie, Clément, Thib, Ben, Laura et Marion** !

Binwei, tu es certainement le plus ancien de mes amis. C'est vrai qu'on s'était un peu perdus de vue avec mes études à droite et à gauche mais je suis vraiment très content qu'on arrive à se voir plus depuis que je suis en région parisienne. Merci pour les discussions et l'aide pour la moto et les apparts.

Robin alias "Robinou" ou "lapin". Discuter avec toi et écouter tes discours anarchiques est toujours un plaisir autour d'un bon verre. Tu es sûrement l'un des plus fiables dans mon entourage et pour cela tu as ma gratitude !

Max et Gwendou, ça pour le coup c'était pas prévu. Merci encore à vous les bretons (oui ce n'est pas évident mais des fois il faut bien ravaler sa fierté de Normand) de m'avoir accueilli alors que je me retrouvais seul dans une situation compliquée. J'ai (re)découvert tellement de choses en vivant avec vous, ça change tout quand on a envie de rentrer à la maison après une dure journée ! **Max**, tu es parfois un peu mystérieux et énigmatique mais j'aime ton honnêteté et ta franchise. Je tiens beaucoup à nos soirées jeux de société ou les incontournables parties de Worms qui ont rythmées ma vie ces deux dernières années !

J'en profite pour remercier **Rémi et Virginie** pour toutes ces soirées jeux et/ou fou rires en France et en Bulgarie, ce sont ces moments là qui me permette de me vider la tête quand j'ai pas sport !

Thanks also to **Fanny, Marco and Sam** who accepted me at their place for the entire first confinement during which we shared sport sessions and good food as well as hard working sessions

and fun !

La famille

Ces quatre dernières années ont vu notre famille globalement se ressouder et se rapprocher peu à peu et je suis très content de cela. C'est bien à nous les "cousins" de faire aussi cet effort parfois et je pense qu'on est bien parti.

Plus spécialement, merci **papa** pour ton soutient sans faille. Tu n'es pas le plus expressif mais je sais que je peux toujours compter sur toi quoi qu'il arrive et c'est très important pour moi.

Simon, même si des fois nos points de vue divergent et nos caractères s'entrechoquent, je suis toujours très content de te voir et pouvoir échanger avec toi. Tu as beaucoup mûri récemment et je trouve que tu as très bien géré ces dernières années et que tu es sur une bonne pente ! Félicitation pour ton CDI !

Justin, notre écart d'âge n'est pas si important tellement je te trouve intéressé et renseigné sur des sujets importants et "adultes". J'ai l'impression récemment d'avoir enfin pu être un "vrai" grand frère pour toi et cela me fait extrêmement plaisir.

Quoi qu'il advienne, je souhaite que cela soit écrit noir sur blanc, **papa, Simon et Justin**, je suis très fier de vous pour vos combats et victoires personnelles récentes. Je vous aime fort. Continuez !

Ralitsa

Тук няма да коментирам много, това е предимно шега за тези, които ще копират/поставят това българско изречение в преводач! Хванах те!

Както и да е, ти си знаеш колко важна беше твоята подкепа, било то техническа и морална, особено през последните няколко месеца от тази докторантура. Ти си много важна в моя живот и се надявам, че животът ни заедно, който започна съвсем скоро, да продължи за доста дълго.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER

1

INTRODUCTION

The focus of the work over the course this PhD was on the use of bioinformatics approaches in order to describe and characterize environmental systems. In this chapter, I will first provide historical background with an ecological perspective of Microbiology and Bioinformatics, the two major research fields used in this PhD research project. Then, I will describe the extraordinary environments on which I had the opportunity to apply these tools: Movile cave and lake Baikal. Finally, I will summarize the main goals I addressed during this thesis.

1.1 "Who's there?" It's microbes!

Today, Microbiology is a very broad field, but at the beginning, it was the study of microbial diversity and the evolution and life cycle of these micro cells. The initial focus was mainly in the context of medicine, studying pathogenic microorganisms to prevent infectious diseases from

spreading in humans. More recently, the focus of Microbiology has expanded to also include the perspective of Ecology (the "study of the house" from house-*oikos* and the suffix study-*logia*) besides a physiological one, investigating the role of these forms of life on our planet by describing their presence and absence in various environments and analysing their interactions with each other or with macro forms of life like plants and animals.

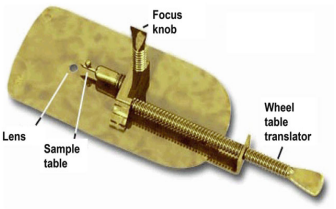
1.1.1 Microbiology

The word “Microbiology" itself is composed of three Greek words: small-*mïkros*, life-*bios*, and study-*logia*, and as a scientific discipline, Microbiology studies the forms of life we can’t see with the naked human eye: the microorganisms. Microbial organisms have been explored through the use of microscopes from the mid-seventeenth century onwards. The pioneers of microbiology and the first observers and describers of microbial cells were Robert Hooke (1635-1703), Antonie van Leeuwenhoek (1632–1723) and Louis Joblot (1645–1723) (Caumette et al., 2015). In 1655 Robert Hooke published his book *Micrographia* where he described what can be seen through the lens of his microscope prototype (see Figure 1.1a). His drawings of plants and insects as seen under the microscope were pioneering, overshadowed only by the first sketches of previously unseen microorganisms (fungus and probably protozoa). Yet the resolution of this early microscope did not allow him to see bacteria and other smaller microorganisms which remained hidden from view. In addition to his merit in being vastly influential and setting the stage for Microbiology, Hooke also coined the term *cell* in reference to the cells of an honeycomb, which plant cells reminded him of.

The first to see and describe single-cell organisms was Antonie van Leeuwenhoek in 1674. His work was the beginning of current protistology, a nowadays important scientific field of biology. In 1676, using the limit of his microscope magnification capacity he described bacterial cells for the first time and estimated that the volume of thousands of those cells would be equal to a small grain of sand. Through these discoveries and many others, Antonie van Leeuwenhoek is considered the *father of Microbiology*.

A contemporary of Antonie van Leeuwenhoek, Louis Joblot was a pioneer of Microbiology

(a) *Microscope manufactured by Christopher White of London for Robert Hooke. Hooke is believed to have used this microscope for the observations that formed the basis of Micrographia*



(b) *Microscope used by Antonie van Leeuwenhoek for its discovery of protists and bacteria.*

(c) *Microscope used by Louis Joblot for his discoveries.*

Figure 1.1 – Images of the first microscope prototypes used for microbiology. (sources : (a) `lensonleeuwenhoek.net`, (b) adapted from Caumette et al. (2015) and (c) `https://micro.magnet.fsu.edu/primer/museum/joblot1718.html`)

who is somewhat neglected in microbiology history today, likely because his work was published later (1718). While his descriptions and drawings were considered as excellent for the time (Lechevalier, 1976) and he improved microscopy techniques, Joblot's landmark contribution was his experiments in opposition to the spontaneous generation hypothesis for microbial forms of life. In this, he was far ahead of his contemporaries.

Until the mid-nineteenth century microbiology had not undergone any major transformations mainly due to the limited tools. It was with the work of Ferdinand Cohn, Robert Koch and Luis Pasteur that isolation and cultivation approaches were improved. This helped them to discover many new lineages in their research on pathogenic bacteria, also decisively disproving the spontaneous generation hypothesis which was still a subject of contention at the time.

1.1.2 Microbial diversity

Life on Earth, in the form of microbial cells, appeared 3.8-3.9 billion and microbes have been the most abundant form of life ever since. Today, the number of microbial species is still subject to debate but earth home at least 10^6 (Louca et al., 2019) and maximum 10^{14} (Lennon, Locey, 2020) prokaryotic cells according to the most recent estimations. In any above cases, microbial cells (protists included) are still found key players in the functioning of our ecosystems.

Microorganisms are mainly unicellular, ubiquitous and spanning across what is called the tree of life. In 1937, Edouard Chatton coined the terms *prokaryotes* and *eukaryotes* based on, respectively, the absence or presence of a nucleus in the observed cell and then divided the world of microorganisms in two. Later, in 1962, Stanier, Niel van (1962) described for the first time many of the molecular differences between viruses, bacteria and protists (microbial eukaryotes) and confirmed Chatton convictions.

1.1.2.1 Prokaryotes

Prokaryotes are microorganisms in which all machinery reactions such as translation and traduction processes happen directly in the cytoplasm, without any organelles like nucleus. Prokaryotes, also called the 'unseen majority' (Whitman et al., 1998), are divided into the two primary domains of life - Archaea and Bacteria, both composed exclusively of unicellular organisms. Bacteria have been known since their discovery by Antonie van Leeuwenhoek in the 1670s, but Archaea were only discovered 300 years later by Carl Woese and George Fox in 1977 (Woese, Fox, 1977) through rRNA gene analyses (see Section 1.2.2.2).

Bacteria In recent years, the bacterial world has also been subject to breakthrough studies. Hug et al. (2016) studied conserved proteins and retained a set of 16 ribosomal genes[1] in order to build a phylogenomic alignment and infer the tree of life. This allowed them to considerably revise the tree of life, adding a vast expansion further highlighting the predominance of bacterial

[1] ribosomal genes are involved in the translation machinery and thus are good single-copy phylogenetic marker genes candidates

lineages compared to archaeal and eukaryotic ones (see Figure 1.2). Genome sequences of isolated or cultured representative genomes are still lacking for many of these major bacterial lineages, especially in the newly described group Candidate Phyla Radiation (CPR (Brown et al., 2015)). Since, Parks et al. (2018, 2020) proposed another classification based on genome phylogeny alone, which is controversial. Indeed, the authors implemented a standardized taxonomy for bacteria solely based on inferred concatenated protein phylogeny trees. Therefore, with each update of the database, the taxonomy is likely to change according to the added taxa. Another potential issue with this is that in this proposed classification, morphological and metabolic traits, which have often served as basis for the initial description of already known groups, would not be taken into account. This could create confusion as the same taxonomic name could end up representing possibly completely different taxa.

Archaea In the past 5 to 10 years, the phylogeny of prokaryotic domains has encountered significant changes with a deep impact on our understanding of these domains of life. Before 2013, archaea were mostly divided into 2 groups, the TACK superphylum (Guy, Ettema, 2011) and Euryarchaeota. Then, in 2013, Rinke et al. (2013) coined the DPANN superphylum regrouping many deep-branching archaeal lineages which did not belong to either of the primary first groups. Since then, many new lineages has been added to this group (Castelle et al., 2015) and the monophyly of this superphylum is still actively debated (Dombrowski et al., 2019). In 2015, Spang et al. (2015) described a novel candidate archaeal phylum: the Lokiarchaeota. Certain eukaryotic signature proteins could be found in the genomes of this novel phylum, therefore placing it as a monophyletic group at the base of the eukaryotes in the tree of life. Following this discovery, many other new archaeal lineages and phyla closely related to eukaryotes were found and they were regrouped two years later by Zaremba-Niedzwiedzka et al. (2017) into the Asgard superphylum (all members are named after Norse mythology). Before earlier this year (2020), all evidence of this superphylum was inferred from big datasets without cultured representatives and its validity and placement in the tree of life was therefore questioned (Da Cunha et al., 2017, 2018; Williams et al., 2020). Since January 2020 and the publication by Imachi et al. (2020),

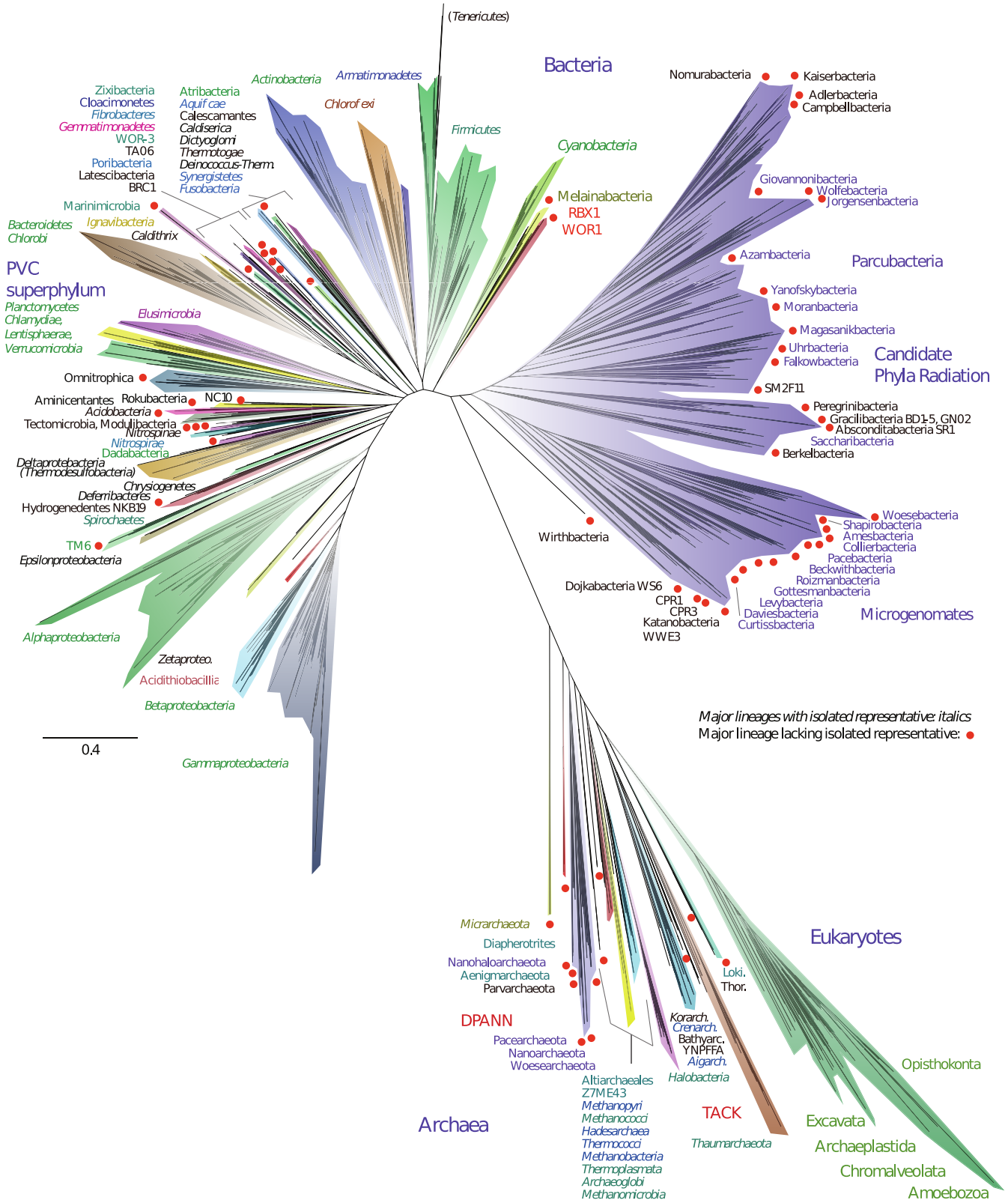
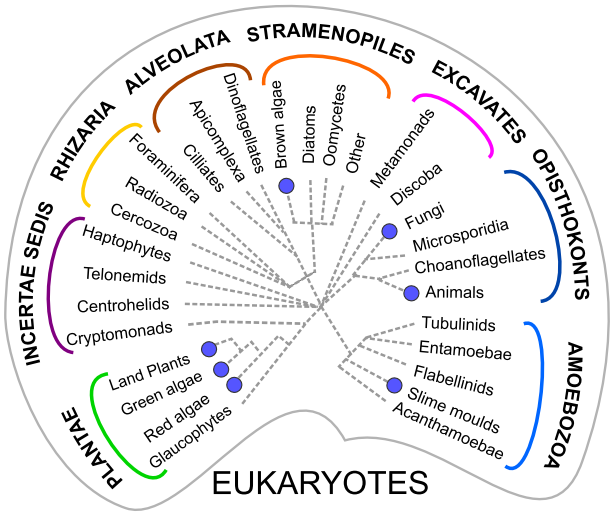


Figure 1.2 – Metagenomic tree of life using a set of ribosomal proteins and displaying 92 named bacterial phyla, 26 archaeal phyla and all five of the Eukaryotic supergroups. (source: adapted from Hug et al. (2016))

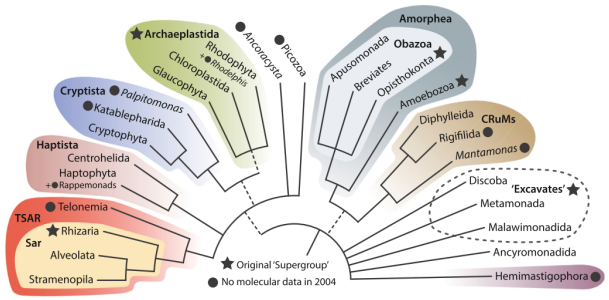
there is now a cultivated representative of the Asgard superphylum, opening new possibilities to study hypotheses about eukaryogenesis, the origin of eukaryotes from prokaryote symbioses reviewed in López-García, Moreira (2020).

1.1.2.2 Eukaryotes

The secondary domain of life, Eucarya (cells with organelles like nucleus) is also mainly composed of various unicellular microorganisms named protists (Figure 1.3a;Kazamia et al. (2016)) even though species of Metazoan (animals), plants and Fungi are described more extensively (Burki, 2014; Burki et al., 2019). The eukaryotic tree of life (eToL) remains debated (see Figure 1.3b) as data are missing for some under-studied protist taxa (Sibbald, Archibald, 2017), leaving supergroups of the eToL not significantly supported (Burki et al., 2019)(see Figure 1.3b).



(a) Unicellularity in Eukaryotic tree of life



(b) Consensus tree from eukaryotic phylogenomics

Figure 1.3 – (a) A schematic diagram showing the major group of eukaryotes and how unicellular organisms dominate this eukaryotic tree of life. In 2016, the positions of the Haptophytes, Telonemids, Cryptomonads and Centrohelids remained uncertain (Incertae sedis). Multicellularity groups are highlighted with filled circles. (b) Consensus of eukaryotic phylogenomic trees. The colors represent the nowadays *'supergroups'*. The tree shows unresolved branching orders and monophyly uncertainties of some groups by the use of multifurcations and dashed lines. (source: (a) adapted from Kazamia et al. (2016); (b) adapted from Burki et al. (2019))

Protists are infamous for causing diseases (Sibbald, Archibald, 2017) but the vast majority of protists fill ecological roles (Geisen et al., 2015). Indeed, despite receiving relatively little attention in the field of microbial ecology compared to prokaryotes, as pointed out by Caron

et al. (2009)'s perspective *Protists are microbes too: a perspective*, protists are key players in the ecosystems, either as autotrophs (primary producers of food for the ecosystem; (Field et al., 1998; Ynalvez et al., 2018)), heterotrophs (consumers of environmental molecules or cells; (Glücksman et al., 2010)) or both (mixotrophs) (see Section 1.2.2.2 for more details). For example, aquatic photosynthetic protists *i.e.* autotrophic algae are responsible for half of the carbon fixed through photosynthesis every year on Earth as estimated by Ynalvez et al. (2018). Also, it was long thought that horizontal gene transfer (HGT) plays only an adaptive role for prokaryotic species to their environment but protists have been identified to do so as well (Eme et al. (2017); Leger et al. (2018) and Yubuki et al. (2020) in Appendix E).

1.1.3 Microbial ecology

All these life forms do not evolve and develop in isolation – microbes (protists and prokaryotic microorganisms) coexist in complex habitats. The first steps in investigating this and the field of microbial ecology started with the exceptional fifty years of work of Sergei Winogradsky in the end of the 19th century (Dworkin, 2012; Caumette et al., 2015). In 1887, he was the first to characterize the chemolithotrophy (energy metabolism through oxidation of inorganic substances) in bacteria oxidizing hydrogen sulfide (H_2S) to sulfuric acid (H_2SO_4). Three years later, while working on denitrification, he succeeded to grow microbes without any organic inputs to the media and observed an increasing amount of organic molecules as the colony grew, which led him to conclude that *"a complete synthesis of organic material by the action of living organisms has been accomplished on our planet independent of solar energy"*. Autotrophy *i.e.* producing complex organic compounds from simple carbon sources such as CO_2 was discovered. Other contributions of Winogradsky includethe description of the nitrification process involving two bacterial species he isolated from field samples, and the development of the direct method for studying the microbiology of the soil, an important milestone in the microbial ecology heritage (Madigan et al., 2015). Winogradsky was a pioneer and in a way, the first microbial ecologist in the sense that he tried to understand the microorganisms' role in their environments.

As previously introduced, microbial ecology is the science studying microorganisms and their

interactions with the environment and between each other. Interactions between the microbial species themselves or with macro-organisms are referred to as biotic, while interactions with physical and chemical components of the microbial community habitat as abiotic. In broad terms, abiotic interactions are key to a microorganism's metabolism, cell structure, physiology and overall to its survival in the environment. Biotic interactions, on the other hand, are the mediators in community functioning and the ecosystem at large. Microbes are ubiquitous on Earth habitats: they can thrive in the metazoan gut, plant leaves as well as extreme environments such as boiling hot springs, permafrost, very acidic environments (near pH 0), salt saturated brines and even environments contaminated with radionuclides or heavy metals. The limits of life are the subject of debates, and only a specific setting of multiple extreme conditions has been recently shown to be life-free (Belilla et al., 2019). The diversity of microbial habitats translates to a metabolic and ecological diversity of microorganisms in them. This diversity is the subject of exploration of microbial ecology.

1.1.3.1 Metabolic strategies

Every living cell is constituted of 7 major elements which are essential: Carbon (C, $\sim$50%), Oxygen (O, $\sim$17%), Nitrogen (N, $\sim$13%), Hydrogen (H, $\sim$8%), Phosphate (P, $\sim$3%), Sulfur (S, $\sim$2%) and Selenium (Se, $<$0.01%)(percentages of dry weight from Madigan et al. (2015); Fagerbakke et al. (1996)). These as well as other elements a given cell may require need to be extracted from the environment, and the process of incorporating outside elements within the cell is termed metabolic assimilation. Using the elements collected from nature, cells can produce more complex molecules.

As carbon is the very basis of organic molecules, an ecosystem is dependent on its carbon sources (although in environments with ubiquitous carbon the limiting factors of cell growth may be other nutrients such as N and P (Elser et al., 2007)). To produce their cell material, microorganisms can obtain carbon from inorganic sources (autotrophs) or organic sources (heterotrophs) in process termed carbon assimilation.

Autotrophs, also called primary producers, are critical to a thriving ecosystem. These mi-

croorganisms assimilate carbon using inorganic carbon sources such as carbon dioxide (CO_2). Six carbon fixation pathways have been identified so far (see Thauer (2007); Berg (2011) for reviews): the Calvin-Benson reductive pentose phosphate cycle (Calvin cycle), Reductive citric acid cycle or Arnon-Buchanan cycle (rTCA cycle), Reductive acetyl-CoA pathway or Wood-Ljungdahl pathway (WLP pathway), 3-Hydroxypropionate bi-cycle or Fuchs-Holo bi-cycle (HP bi-cycle), 3-hydroxypropionate/4-hydroxybutyrate cycle (HP/HB cycle) and the dicarboxylate/4-Hydroxybutyrate cycle (DC/HB cycle). While all of these major pathways have already been studied extensively because of their importance in ecosystems (Santoro et al., 2013; Baltar, Herndl, 2019), novel variants are still being discovered (Assié et al., 2020; Rubin-Blum et al., 2019; Mall et al., 2018; Nunoura et al., 2018). The more complex carbon molecules synthesized by autotrophs serve as organic carbon source for heterotrophic organisms. Heterotrophs are thus consumers that need to assimilate carbon from organic molecules. To do so, they can be osmotrophs, absorbing carbohydrates, fats, and proteins available in the environment (produced by the autotrophs) or they can feed on other cells (autotrophs or other heterotrophs) by phagocytosis. Phagocytosis is a conserved eukaryotic feature but has also been found very recently in the prokaryotic phylum Planctomycetes (Shiratori et al., 2019). Mixotrophs are organisms which can be autotrophs and heterotrophs. These are mainly primarily photosynthetic algae that consume other cells through phagocytosis in certain conditions.

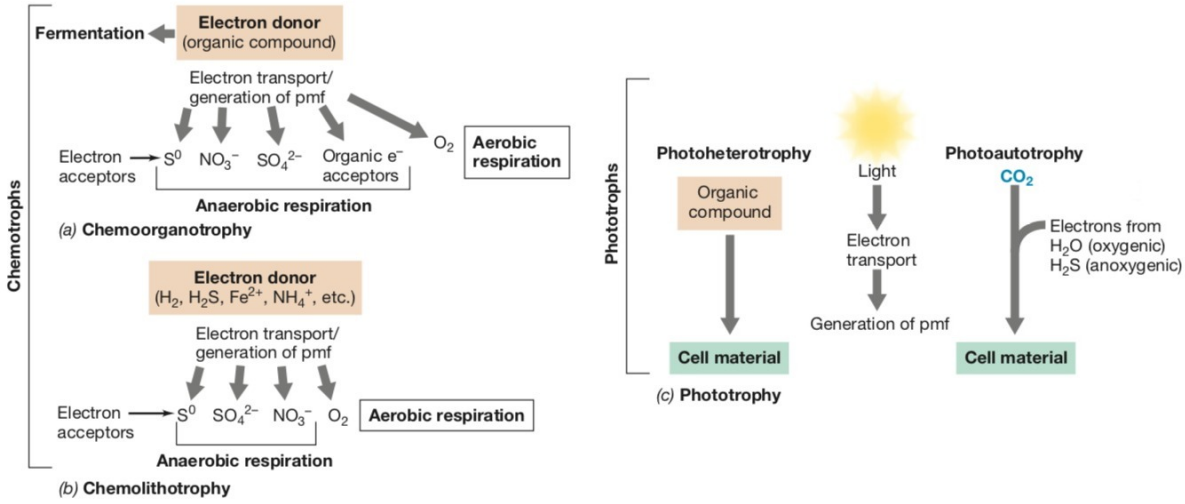


Figure 1.4 – A schematic diagram showing the different metabolic classes found in microorganisms. (source: adapted from Madigan et al. (2015))

Regardless of how a microorganism assimilates carbon, it would require energy for cell biosynthesis and maintenance. Therefore, orthogonal to carbon fixation strategy, organisms can be phototrophs or chemotrophs depending on how and what they oxidize to produce and store energy. Phototrophs are organisms which convert light into chemically-stored energy. In microbes, this is achieved through pigments present in the cells. The most common and well-known microbial phototrophs are cyanobacteria and algae, which oxidize water and produce oxygen through the oxygenic photosynthesis pathway, but *anoxygenic* photosynthesis also exists in some species which oxidize other chemical substances (e.g. green sulfur bacteria that oxidize hydrogen sulfide (H_2S)). Instead of light, chemotrophs derive energy from chemical reactions through catabolic pathways. Chemotrophs are divided into two sub-classes depending on the type of chemical compounds used for energy: Chemoorganotrophs oxidize organic molecules such as sugars, while Chemolithotrophs oxidize inorganic substances such as NH_3 or H_2S. The latter group has a limited selection of exploitable inorganic compounds, which can often be byproducts of metabolic pathways employed by the former group. Both of these processes are taking place over catabolic pathways comprising many enzymatic reactions to finally oxidize a compound (organic for Chemoorganotrophs, inorganic for Chemolithotrophs) and gain energy, producing an energy-rich molecule such as Adenosine Triphosphate (ATP), which can be used by the cell to drive processes required for life. Some of these reactions can be pathway-specific, with the enzymes exclusively adapted to catalyse a reaction exclusive to the pathway. Consequently, such specific enzymes could be used to predict presence or absence of a specific metabolic trait in the genome sequence of a species.

1.1.3.2 The case of upper layer sediments

All the above processes can take place under aerobic (presence of O_2 in the environment) or anaerobic (absence of O_2 in the environment) conditions. Aerobic conditions are the most commonly known because the ecosystems with available O_2 are easier to access and therefore to study. Also, growth under aerobic conditions is favored in the presence of O_2 as the O_2/H_2O oxidation–reduction (redox) couple has the highest standard electrode potential E'0, *i.e.*

power/inertia to oxidize or reduce, with high positive values meaning a strong inertia for oxidation. When coupled with the oxidation of complex organic molecules (*i.e.*, aerobic respiration (Figure 1.4)) the reaction releases a considerable amount of energy. However, in absence of oxygen, microbial life uses other redox couples to achieve anaerobic respiration even if those yield less energy.

Upper layer sediments are interesting and complex ecosystems. In oligotrophic water bodies, sediments are usually composed of two habitats, one at the top of the sediment which is still aerobic or micro-aerobic and the second, just below, in anoxic conditions. Moreover, sediments are the reservoir of sinking organic materials and the place of decomposition of organic matters. Therefore, the role of microbial communities in the nutrient cycle involves many different classes of energy transformation and a wide diversity of microorganisms (Orsi, 2018).

1.2 And Bioinformatics arises

Bioinformatic may directly or indirectly inherit from the early field of computational biology. Computational biology started in late 1950s and early 1960s mainly with the work of Margaret Oakley Dayhoff (and colleagues). Dayhoff developed programs in the FORTRAN programming language to reconstruct the complete amino-acid sequences of proteins based on overlaps of partial sequences (Dayhoff, Ledley, 1962; Dayhoff, 1965), a precursor to “assembly", a step widely used nowadays with DNA datasets. In this work, to reduce computational load she established the one letter amino-acid alphabet. Dayhoff and colleagues rapidly reached the conclusion that for sequencing with this methodology to be useful, one needed reference sequences to compare the result to. Therefore, they created the first database: the *Atlas of Protein Sequence and Structure* (Strasser, 2010) which was later made available online (Dayhoff et al., 1981; Orcutt et al., 1983) and eventually became the PIR database (Barker et al., 1991). They developed protocols to compare sequences and eventually, following the work of other early molecular evolutionists (Needleman, Wunsch, 1970), developed matrices of amino-acid substitutions (point accepted mutation, PAM) which score sequence alignments by penalizing incongruent substitu-

tions (where an amino acid being substituted by another amino acid with vastly different structure or chemical properties) more than minor substitutions (two similar amino acids). These tools are still used today, with some improvements (Lambert et al., 2005).

While computational biology was blooming, the term 'bioinformatics' was coined in 1970 by Dutch scientists Hesper and Hogeweg to mean 'the study of informatic processes in biotic systems' (Hesper, Hogeweg, 1970; Hogeweg, 2011). Their idea was to study the processes of information processing, accumulation, transmission and interpretation happening in living systems in order to better understand their functioning. However, when Frederick Sanger succeeded to sequence DNA in the 70s, all techniques previously developed for peptide comparisons could be also applied for nucleic acids (Sanger et al., 1977). With these abilities of sequencing and comparing datasets on a routine basis, as well as the hypothesis of molecular clock[2] and the rise of evolutionists, bioinformatics started to refer to the treatment of these molecular datasets using modern tools like computers and available sequence databases for sequence alignment, comparison and phylogeny (Hagen, 2000, 2001).

Nowadays, the term bioinformatics refers to an interdisciplinary research field involving theoretical molecular biology, method and software development, computational biology, computer science as well as mathematics and statistics. In other words, bioinformatics can describe everything from the computation of a new software to the use of this software for biological interpretation, especially on molecular datasets. Even though the definition of bioinformatics or bioinformatician can be debated (Vincent, Charette, 2015; Smith, 2015, 2018), most of them study macromolecules-based datasets, the building blocks of molecular biology.

1.2.1 Molecular biology

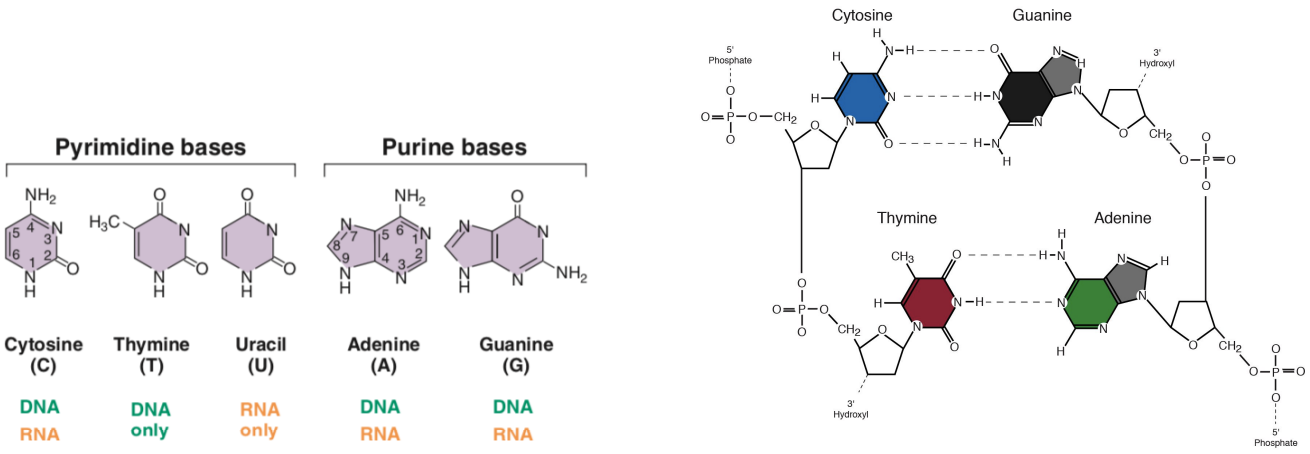
For 301 years, from the first report of bacterial cells through a microscope by van Leeuwenhoek in 1676 and the first DNA sequencing technique by Sanger et al. (1977), microbial communities were only studied in reference to their morphological traits, supported media for growth, and other visual features. Thanks to advances in molecular biology, we are now able to sequence more,

[2]the "molecular clock" refers to the concept of using the mutation rate to deduce time when life forms diverged

faster and for cheaper (Goodwin et al., 2016). These advances allow us to posit and address new hypotheses on the microbial world, its evolution, the diverse roles of microorganisms on Earth's ecosystems and their interactions with macroorganisms. Hereafter, I will describe nucleic acids and proteins – the key pillars of molecular biology – along with the central dogma of life which links them together.

1.2.1.1 Nucleic acids - DNA and RNA

For a brief recount of the key discoveries of nucleic acids, we need to go back quickly at the second half of 19th century. A curious acidic substance 'nuclein' (later called DNA) was first isolated from cell nuclei in 1869 by Friedrich Miescher (Dahm, 2005), gaining the attention of many scientists. Afterwards, Albrecht Kossel isolated and described the five organic compounds composing nucleic acids (later known as the five nucleobases) and coined the terms: adenine, cytosine, guanine, thymine, and uracil. Then, at the beginning the 20th century, Phoebus Levene discovered that nucleobases are linked together through a pentose sugar and a phosphate chain. He coined this structure (phosphate–sugar–nucleobase) nucleotide and formulated the "tetranucleotide hypothesis" about the structure of the DNA molecule as a ring of four nucleotides linked together through their phosphate groups, which remained the predominant view for decades.



(a) *Nucleobases.* (b) *Nucleotides.*

Figure 1.5 – Illustration of nucleobases (left) and nuclotides (right) structures. (sources : (a) Madigan et al. (2015) (b) `https://knowgenetics.org/nucleotides-and-bases/`)

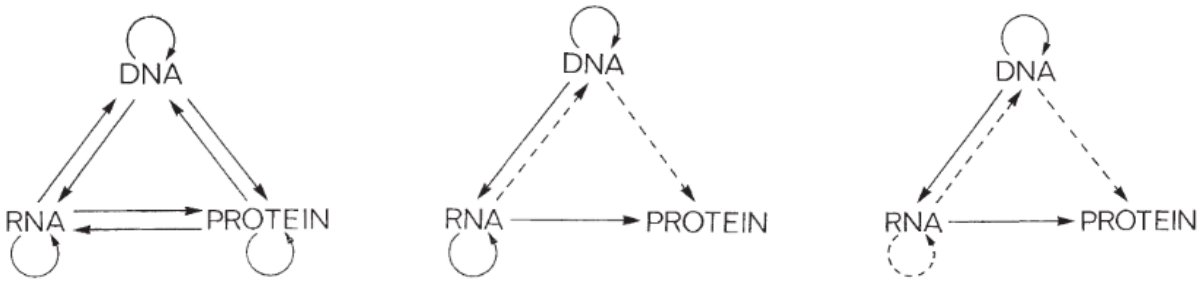
In a very interesting review, Allen (1941) proposed to substitute the terms “plant nucleic acids" and “animal nucleic acids" to RNA and DNA respectively because of their carbohydrate difference: *d*-ribose and *d*-2-deoxyribose, and posed unanswered questions about the structure of DNA, notably in what order would the four nucleotides be in the tetranucleotide ring and how would the proposed tetranucleotides be linked together in a polymerized molecule. This proposed DNA structure with equal proportions of the four nucleotides (and thus of guanine (G), adenine (A), cytosine (C), and thymine (T)) was disproved by Erwin Chargaff who, in the late 1940s, showed that in a cell, there is a 1:1 ratio between the amount of G and C and between the amount of A and T, and that the relative amounts of A, G, C, T bases vary between species. The double helix structure of DNA was famously discovered by Rosalind Franklin, James D. Watson, and Francis Crick in 1953 (Watson, Crick, 1953). It revealed the reason behind Chargaff’s rules: nucleobases are paired together through hydrogen bonds, forming a base pair (G—C or A–T), a fundamental unit of DNA. This important discovery was a milestone advance of our understanding of life.

1.2.1.2 Proteins

Proteins are key to most cell functions as they catalyze reactions (in the case of enzymes), transport molecules, react to stimuli, provide structure within the cells. In the early 20th century, Emil Fischer put forth the view of proteins as polypeptides: compounds formed by linked amino acids[3]. At the beginning of the 1950s, Frederick Sanger successfully determined the amino acid sequence of the two chains of bovine insulin. Thus, he proved that proteins consisted of linear amino acid polymers. This understanding guided Francis Crick to formulate the sequence hypothesis during a lecture in 1957, published in 1958, later clarified and updated it in scientific article Crick (1970) illustrating the concept (see Figure 1.6).

Indeed, we now know that the amino acid sequence is determined by the genetic code. DNA is transcribed into messenger RNA (mRNA), which is then translated into proteins by the ribonucleoprotein complex macromolecule called ribosome. This translation mechanism is present in

[3]The majority of the twenty common amino acids were discovered at the beginning of the 19th century but the last one, threonine, was discovered in 1935 by William Cumming Rose.



(a) *All possibility hypothesis* (b) *Crick central dogma in 1957-1958* (c) *Crick central dogma in 1970*

Figure 1.6 – Figures published by Crick to illustrate his central dogma of molecular biology, showing the possible, impossible and unusual way for information to be transferred in molecular biology. Solid arrows indicate probable transfers of information, dotted arrows indicate improbably but possible transfers, and a lack of an arrow indicates impossible transfers. In the final version of the central dogma (c), DNA can replicate, DNA can be transcribed into RNA, and RNA can be translated into proteins (black arrows). While other ways of transfer remain possible, Crick was clear that once the information has reached the protein level, it cannot go back (note absence of arrows starting from proteins in (b) and (c)), which is, among other reasons, a consequence of the redundancy of the genetic code shown by Bernfield, Nirenberg (1965). (source : Crick (1970).

all living organisms and also in some viruses and organelles such as mitochondria and plastids. Quickly, mRNA sequences are read per group of three nucleotides called codons as proved by the experiment in 1961 conducted by Crick et al. (1961). Each of these mRNA codons encodes for an amino acid. The encoding mecanism is as follow: for each mRNA codon the ribosome recruits the corresponding tRNA anticodon (RNA triplet site of small RNA polymers named transfer RNA) used as a tag for the corresponding tRNA cargo amino acid as discovered for the first time by Robert W. Holley in 1965 (Holley et al., 1965). Each tRNA encodes for only one anticodon and its corresponding amino acid as shown, also in 1964, by Nirenberg, Leder (1964) in which they revealed the sequences of 54 existing codons out of the $4^3 = 64$ possible permutations. This was later called the genetic code (see Table 1.1). As these codons encode amino acids, of which only around twenty exist, redundancy is present in the genetic code. Codons which are interchangeable in the sense of producing the same exact amino acid were coined synonymous codons and are the key to the concept of codon usage bias. Also, within the 64 combinations, three are stop codons, which are recognised by a release factor protein rather than a tRNA anticodon, which causes the ribosome to release the finished peptide. The protein structure is determined by the

amino acid sequence forming the protein as chemical interactions cause it to fold in particular shapes. The first known protein structures were those of hemoglobin and myoglobin determined in 1954 by Max Perutz (Perutz, 1954) and in 1958 by Sir John Cowdery Kendrew (Kendrew et al., 1958), respectively. Wetlaufer introduced the concept of protein domains which were defined as stable units of protein structures in 1973 (Wetlaufer, 1973).

	U			C			A			G			
U	UUU	Phe	[F]	UCU	Ser	[S]	UAU	Tyr	[Y]	UGU	Cys	[C]	U
	UUC	Phe	[F]	UCC	Ser	[S]	UAC	Tyr	[Y]	UGC	Cys	[C]	C
	UUA	Leu	[L]	UCA	Ser	[S]	UAA	*STOP*		UGA	*STOP*		A
	UUG	Leu	[L]	UCG	Ser	[S]	UAG	*STOP*		UGG	Trp	[W]	G
C	CUU	Leu	[L]	CCU	Pro	[P]	CAU	His	[H]	CGU	Arg	[R]	U
	CUC	Leu	[L]	CCC	Pro	[P]	CAC	His	[H]	CGC	Arg	[R]	C
	CUA	Leu	[L]	CCA	Pro	[P]	CAA	Gln	[Q]	CGA	Arg	[R]	A
	CUG	Leu	[L]	CCG	Pro	[P]	CAG	Gln	[Q]	CGG	Arg	[R]	G
A	AUU	Ile	[I]	ACU	Thr	[T]	AAU	Asn	[N]	AGU	Ser	[S]	U
	AUC	Ile	[I]	ACC	Thr	[T]	AAC	Asn	[N]	AGC	Ser	[S]	C
	AUA	Ile	[I]	ACA	Thr	[T]	AAA	Lys	[K]	AGA	Arg	[R]	A
	AUG	Met	[M]	ACG	Thr	[T]	AAG	Lys	[K]	AGG	Arg	[R]	G
G	GUU	Val	[V]	GCU	Ala	[A]	GAU	Asp	[D]	GGU	Gly	[G]	U
	GUC	Val	[V]	GCC	Ala	[A]	GAC	Asp	[D]	GGC	Gly	[G]	C
	GUA	Val	[V]	GCA	Ala	[A]	GAA	Glu	[E]	GGA	Gly	[G]	A
	GUG	Val	[V]	GCG	Ala	[A]	GAG	Glu	[E]	GGG	Gly	[G]	G

Table 1.1 – The Genetic Code

1.2.2 Bioinformatics applied to microbial diversity and microbial ecology

After the first fully sequenced genome (bacteriophage) (Sanger et al., 1977), the development and popularization of DNA sequencing methods gained momentum, opening new points of views on microbial ecology. Hypotheses flourished, especially on the metabolic potential and ecological place of microorganisms. Indeed, many of today's discoveries and ongoing analyses regard the interactions between microorganisms within ecosystems (horizontal gene transfer, symbiosis, etc..) or between microbial communities and their environment or how they adapt to it (genome size, intron size, GC content, etc..).

1.2.2.1 DNA sequencing technologies history

First generation: Sanger based sequencing As described above, sequences started in earnest with the work of Frederick Sanger. He won his first chemistry Nobel Prize by sequencing the first protein chains in the early 1950s. Then, after dedicating research to RNA sequencing he shifted the focus of his lab to DNA sequencing. Sanger was a pioneer and in Sanger et al. (1965) presented the first widely used protocol for DNA sequencing. Years later, he was the first to publish a complete genome – the bacteriophage phi X 174, a single-strand DNA virus of *E.coli* (Sanger et al., 1977). The same year he published the most important discovery in the era of sequencing: the 'chain-termination' technique (Sanger et al., 1977) involving deoxyribose nucleotide triphosphate (dNTPs) which led to faster and more efficient Sanger sequencing and ultimately to automatisation. His important advances in the field of molecular biology earned him a second chemistry Nobel Prize in 1980. His technique was taken up and improved upon by others and, after almost 15 years of hard work the first large scale DNA-sequencer was released (Hunkapiller et al., 1991). Then, along with some refinement and upgrade of the methodology, the first major advances using this technology were published, among them the first free-living bacterial genome *Haemophilus influenzae* (Fleischmann et al., 1995) followed by many others and the human genome in 2001 (Craig Venter et al., 2001; Lander et al., 2001).

Second generation or next-generation sequencing (NGS): light-based sequencing Sequencers of the second generation replaced the dNTPs or oligonucleotides and the gel electrophoresis output taking advantage of advances in luminescence and fluorescence (Nyrén, Lundin, 1985; Nyrén, 1987) and improvements in the Polymerase Chain Reaction (PCR; Mullis et al. (1986); Saiki et al. (1988)). This light-based technique called pyrosequencing consists of a two-step process involving the enzymes ATP sulfurylase followed by luciferase after each iterative incorporation of nucleotide by the DNA polymerase (see Figure 1.7) (Hyman, 1988). For each incorporated nucleotide, one molecule of pyrophosphate (PPi) is released during the polymerisation; this molecule is detected as a luminous signal proportional to the number of (identical) bases incorporated. Note that the four types of nucleotides are washed and reintroduced

one after the other in consecutive batches as this technique does not discriminate between the nucleotides themselves. One major drawback of the technique is the detection of homopolymer. The difference of luminescence between n and $n+1$ (for $n > 3$) nucleotides incorporated could not be well detected. Pyrosequencing technique was licensed in 2005 by a biotechnology company named *454 Life Sciences* (later *Roche*) and was the first great commercial success in sequencing technology. Indeed, it was the first time that parallel sequencing was developed through the use of beads and adapter sequences undergoing water-in-oil emulsion PCR (emPCR). This allowed clonal DNA population on each bead and then the amplification process within each droplet. The pyrosequencing platform was discontinued by the *Roche* company in 2013 when no longer competitive.

In parallel to pyrosequencing, a different approach also based on light emission (fluorescent dNTPs) was developed by the company *Solexa* (later acquired by *Illumina*) to parallelize the sequencing processes. Instead of using emPCR like pyrosequencing, this technique employed DNA attached to a solid flow cell (through adapters) followed by a solid phase PCR to create clonal DNA clusters. This method allowed for the first time paired-end sequencing outputs (as the reverse strand is also sequenced), which is useful for cross-checking or increasing the read size. Very promising technology in 2008 (Hall, 2007), Illumina methodology has since been published and undergone many improvements and is nowadays the most used sequencing technology (Figure 1.8), replacing pyrosequencing for massive medium-short read sequencing projects (Heydari et al., 2017) (up to 300bp in 2020 on the MiSeq machines). Its main attribute is the low cost (Escobar-Zepeda et al., 2015) and the very high number of sequences sequenced at the same time in addition to the paired-end option. The main drawback in addition to the read length are the unpredictable substitution errors; however, this error rate is relatively low ($\sim$0.1%) and errors are mostly identified using the automated quality score of the nucleotide (Shendure, Ji, 2008).

Other 'wash and scan' methodologies exist, most of them based on light detection like pyrosequencing and Illumina: AB SOLID and related Polonator (Shendure et al., 2005) and HeliScope(Harris et al., 2008). The first non light-based sequencing approach and the first step to

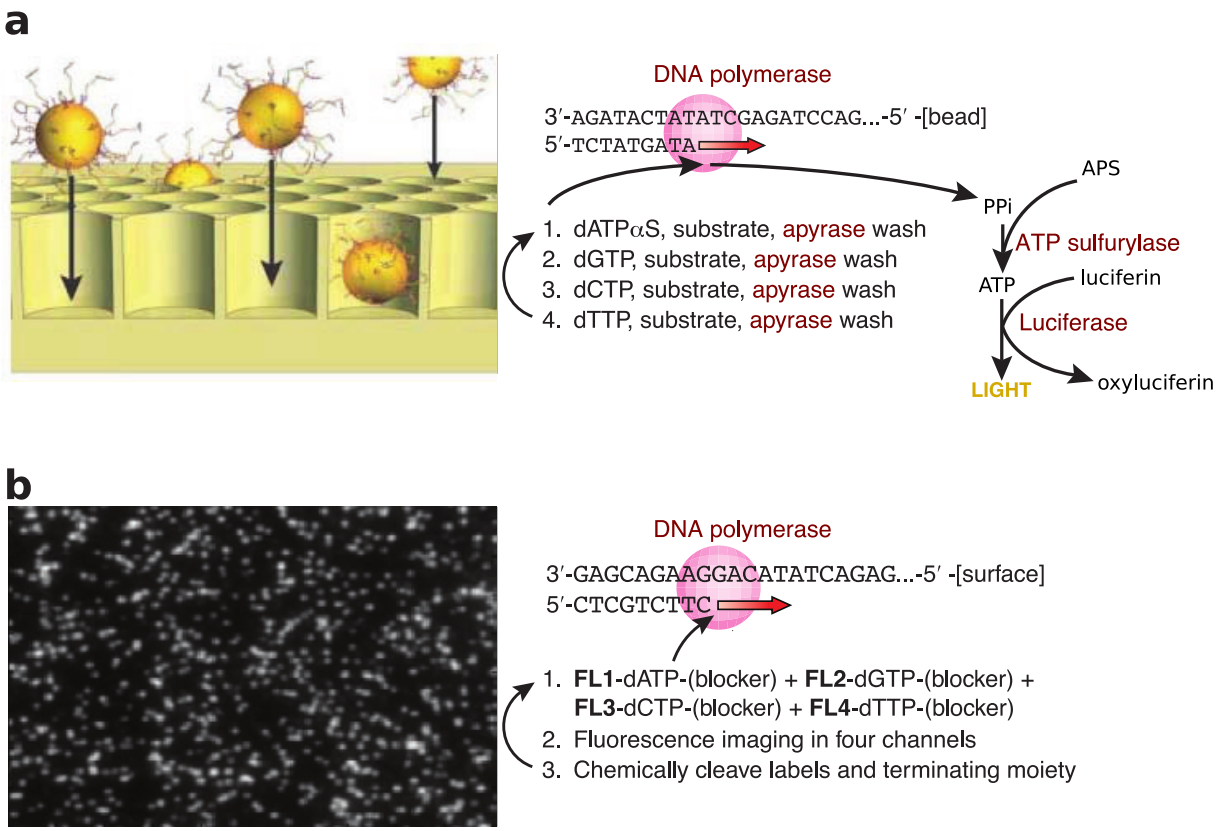


Figure 1.7 – **a.** Pyrosequencing technology in which many enzymes are involved in order to produce light from luminescence source. Each dNTP is incorporated in a batch and the excess is wiped out by an enzyme wash before the next (different) dNTP batch. During the incorporation phase, if the DNA polymerase add dNTPs to the sequence, it releases a pyrophosphate (PPi) which is detected as a luminescence signal. **b.** Solexa/Illumina technology in which fluorescence is captured by a laser after polymerization. In each step, all types of dNTP are incorporated at the same time, each dNTP kind linked to a different fluorochrome and a polymerase blocker structure; a chemical wash takes place in between steps. (source : adapted from Shendure, Ji (2008)

the third generation was released and called Ion Personal Genome Machine system (Ion PGM or 'ion torrent'), detecting pH variations after the release of protons during polymerization (Rothberg et al., 2011). However, these alternatives, extensively reviewed by Shendure, Ji (2008); Metzker (2010); Heather, Chain (2016); Garrido-Cardenas et al. (2017), are not as widely used as pyrosequencing or Illumina and were not used during the work of this thesis and are therefore outside the scope of this manuscript. In parallel of the second generation of sequencers, many tools and advances in protocols were developed, highlighting the first great era of bioinformatics and sequencing technologies (Shendure, Ji, 2008), especially when applied to microbial ecology (Boughner, Singh, 2016).

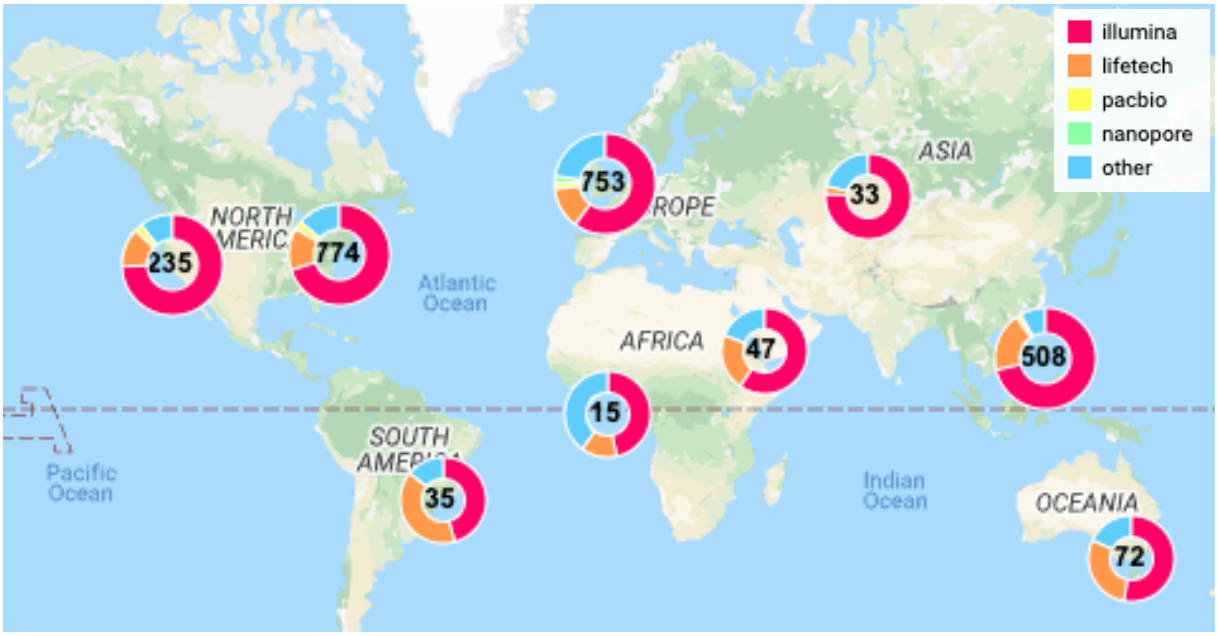


Figure 1.8 – Map of the World High-throughput Sequencers centers highlighting the most used sequencing companies technologies nowadays. (source : `http://enseqlopedia.com/ngs-mapped/` ; accessed 06/08/2020)

Third generation: single molecule-based sequencing (SMS) The main improvement of this generation is the protocols free of 'wash-and-scan'- and PCR, which allows them to target single DNA molecules (Dijk van et al., 2018; Schadt et al., 2010; Heather, Chain, 2016; Garrido-Cardenas et al., 2017) and produce (very) long reads (up to thousands of nucleotides). The major drawback is the high error rate (∼15%), which has prompted recent improvements by either reducing the error rate directly (∼3%) or by sequencing of the same fragments multiple times to correct for errors *post-hoc* (Dijk van et al., 2018). The two technologies which are available to produce long-read sequencing results are *Pacific Bioscience* (Eid et al., 2009) and *Oxford Nanopore* (Mikheyev, Tin, 2014). Long-read sequencing opens new doors and its applications and impact in the field is reviewed in detail in Dijk van et al. (2018).

1.2.2.2 Metabarcoding: a marker gene approach applied at the community scale

Metabarcoding is a culture-independent approach which is nowadays commonly used to study microbial diversity in ecosystems. It consists in employing a barcoding approach to an entire community (Figure 1.9).

What is a barcoding approach? The principle is best expressed by the following Carl Woese's sentence: *To determine relationships covering the entire spectrum of extant living systems, one optimally needs a molecule of appropriately broad distribution* (Woese, Fox, 1977).

Of course, reality is more complex than theory. Ideally, a good marker molecule should be: **i)** very ubiquitous, single-copy and referenced (there should be a reference database of this DNA fragment with identified classified species); **ii)** unlikely to be subject to horizontal gene transfer (HTG); **iii)** of medium/short length to be compatible with the first and second generation of sequencers (up to 1500bp for Sanger and 500bp for NGS, see Section 1.2.2.1); **iv)** with a well-conserved DNA sequence in terms of nucleotide identity across the chosen taxonomical-level species (in order to design universal primers with ideally identical affinity for every species) but at the same **iv)** with DNA sequence containing a variable region in order to discriminate between the taxonomic groups. The realisation of these requirements together is a challenging task on which biologists and phylogeneticists have been working for decades.

The foundations of metabarcoding were set with the work of Carl Woese and collaborators. First, Sogin et al. (1972) identified DNA sequences involved in the translation apparatus as promising marker genes. Indeed, they reviewed the previously reported characteristics of the ribosome molecules and concluded that **i)** the translation machinery is likely to be present in every organisms and therefore can be considered universal; **ii)** this machinery is likely to have been inherited going back to the Last Universal Common Ancestor (LUCA) and has been diverging in every lineage since and **iii)** the protein structures of its component parts change relatively slowly and thus the rRNA sequences are likely to be highly conserved. They chose the 5S rRNA molecule because it was easier to isolate at the time, short and sequence-able and known as other 5S sequences had been published few years earlier(Forget, Weissman, 1967; Brownlee et al., 1967; DuBuy, Weissman, 1971). Sogin et al. (1972) used oligomer distributions (*k*-mer frequencies) to compare their 5S rRNA gene sequences as proper sequencing techniques became available only few years later (Section 1.2.2.1). Three years later, Woese et al. (1975) published a very important study on the conservation of the 16S rRNA gene primary structure and argued that these RNA are directly involved in the ribosomal function, which would explain their low vari-

ability as they are keep their primary functions within the translation machinery. The next year, Woese et al. (1976) adapted the Sanger (Sanger et al., 1965) method for RNA sequencing and sequenced up to 1500–3000 nucleotides. Thanks to this technical feat, Fox et al. (1977) argued that 16S rRNA molecules were the most suitable for classifying the prokaryotes. Indeed, 16S rRNA is longer than the 5S (1600 nucleotides over 120 nucleotides) but it is much easier to sequence than the 23S rRNA (3300 nucleotides) with the modified Sanger technique (Woese et al., 1976). Moreover, 16S rRNA was still part of the translation apparatus and therefore ubiquitous, containing highly conserved regions as well as hypervariable regions. However, the sequencing technique was very time consuming and required a real expertise.

Another achievement of Fox et al. (1977) was to define for the first time a similarity percentage between two 16S rRNA sequences. Similarity percentages became important tools in studying microbial diversity in ecosystems, and as techniques improved, a question was raised: what is the appropriate threshold to use to infer taxonomy classification? Wayne et al. (1987) were the first to introduce a clear protocol and threshold to infer hierarchical taxonomy classification from the DNA-DNA hybridization (DDH) similarity. But as protocols to sequence the 16S rRNA gene improved the first small rRNA databases were published, applying 16S rRNA gene sequence analysis to identify and classify strains against databases proved very useful(Fox et al., 1992; Amann et al., 1995). Comparing with DDH results, Stackebrandt, Goebel (1994) set the sequence identity threshold for 16S rRNA similarity at 97% to define species. Thresholding the 16S rRNA sequence identity proved to be a good proxy for inferring prokaryote taxonomy and the 16S rRNA sequence similarity percentage threshold was raised to 98.7-99% for prokaryotic species identification (Rosselló-Mora, Amann, 2001; Stackebrandt, Jonas, 2006; Chan et al., 2012; Kim et al., 2014; Edgar, 2018).

The first major finding using a barcoding approach was indisputably the discovery of the archaeal domain of life (Woese, Fox, 1977). Woese, Fox (1977) found that the comparison of the 16S and the 18S rRNA genes revealed 3 distinct groups and not two as expected initially. Nonetheless, the scarce taxon sampling due to sequence availability in 1977 needed the sequencing revolution and the work of Sanger to improve enough his sequencing technique. Bar-

coding applied as an exploratory technique of microbial ecology in the field was first achieved in 1990: building up on work by Pace et al. (1986), cultivation-independent approaches revealed the diversity of bacterioplankton in the Sargasso Sea using 16S rRNA (Giovannoni et al., 1990). For 18S rRNA-based studies the first applications followed 10 years later (López-García et al., 2001; Moon-Van Der Staay et al., 2001) with a major impact on the field of protistology (see Moreira, López-García (2002) for a review).

Since then, as technology developed making it possible to sequence more, faster and for cheaper prices. Metabarcoding methodology was developed and benefited from sequencing directly from field samples without having to culture populations (cultivation- independent approaches). Below, I detail the main processes and drawbacks of each step in the metabarcoding approach (Figure 1.9).

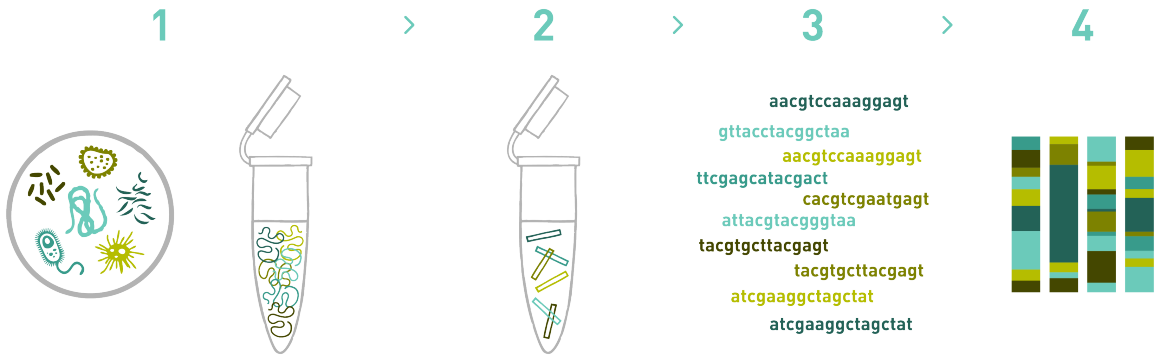


Figure 1.9 – Metabarcoding general steps, see text for details. (source : https://www.allgenetics.eu/services/genomic-solutions-for-your-company/ecotoxicology/characterisation-of-microbial-communities-by-DNA-metabarcoding ; accessed 12/09/2020)

Step 0: sample collection The very first step is the sampling campaign. Sampling multiple biological replicates can avoid sampling biases, but unfortunately due to cost many studies do not perform them, risking to stumble upon a non-representative sample due to chance (Zinger et al., 2019). One should always be aware of the possible bias from the sample collection processes and interpret result with caution. When applicable and doable, field negative controls are also good practice, which can allow one to discard field contaminants(Belilla et al., 2019).

Step 1: DNA extraction Multiple protocols exists for extracting the DNA from environmental samples (Kamble, Singh, 2020). The key steps in extraction kits are the microbial cell lysis and then the purification of the DNA to isolate the nucleic acids and precipitating them in a suitable buffer solution. On first glance, this step may seem simple and easy but many possible artefacts have to be taken into account. The cell lysis can be performed enzymatically, chemically, mechanically or through a combination of these three procedures. The reason why this step can be a hurdle is because while the primary goal for metabarcoding is to highlight the true diversity, some cells are harder to break than others, especially eukaryotes. An extraction step that is too mild would not result in some cells not releasing their DNA, while being too rough risks to break some DNA fragments apart (which is especially problematic for metabarcoding using SMS sequencing technologies which specializes in very long fragments). Either way, a part of the community in the sample is lost and not studied. The purification is less problematic as biochemical companies have developed methodologies including buffers and reactants to suit most microorganisms and DNA types. Nonetheless, there will always be at least a small fraction of the community that ends up underrepresented. One solution is to perform this step multiple times and produce technical replicates (Zinger et al., 2019).

Step 2: PCR amplification The PCR amplification is another step that warrants caution. As for DNA extraction, microbial forms of life are very diverse and contribute to non-specific consensus. Even if marker genes are chosen theoretically to be targeted similarly in every organisms using the highly conserved regions flanked to the hypervariable regions, these conserved regions are still very different from being clonal from one species to another. The design of degenerated PCR primers (mixture of primers with substitutions for different variants of the same sequence) is then crucial to avoid as much as possible the bias introduced by the primer affinity(Kwok et al., 1994). The more taxonomically broad the PCR primer design is, the harder it will be to reduce the selective priming bias. Recent advances have been made to bias the primers against metazoan species in order to retrieve mainly protist 18S rDNA from environmental DNA (eDNA) and therefore extend the sequencing effort onto real and important diversity for the environment

instead of metazoan contaminant (Bower et al., 2004). Moreover, sequencing techniques, with the notable exception of SMS sequencing technologies, require PCR primers to be relatively close to one another on the DNA strand. This limitation further increases the bias as there are not many possibilities to design primers that are close to each other, relatively conserved but still targeting an hypervariable region for which databases have been compiled with reference sequences. In addition to the PCR primer bias, the errors introduced by the DNA polymerase itself during the PCR amplification process are non-negligible (Shagin et al., 2017); this effect can be countered by pooling PCR replicates (Dopheide et al., 2019; Zinger et al., 2019).

Step 3: Sequencing, filtering and grouping the amplicons The sequencing technology chosen for the analyses has an undeniable effect on the post-sequencing treatment of the amplicons. For example, pyrosequencing technology requires a correction for the error of homopolymers detection. Illumina MiSeq has also been studied for error motifs but these are less systematic and easier to overcome (Schirmer et al., 2015). At this step, there is not much a scientist can do but trust the sequencing companies. The next part however, once the sequencing treatment is carried out and the output files are available, can be controlled.

In the case of the *Illumina* platform, which was employed during the course of this thesis, the output files are multiplexed forward (R1) and reverse (R2) fastq[4] files with amplicons and corresponding Phred scores indicating assessed quality. The pair-end reads R1 and R2 should be pruned from multiplexed identifiers and PCR primers, of which there may be multiple copies, and then merged together into the correct full amplicon length as determined by the initial set of PCR primers chosen (Nguyen et al., 2015). One then needs to discard low-quality amplicons and dereplicate the remaining amplicons (meaning removing redundancy in the pool of amplicons). Then, because of PCR bias, chimeric[5] sequences need to be removed from the pool of high quality full length amplicons. The final amplicons are then pooled into Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs); there are different methods to achieve this

[4]Illumina sequencers standard output format. Text format file storing a DNA sequences as letters and the associated quality for each base as symbols.

[5]Chimeras are build during the PCR process and this happens because a short unfinished sequence can serve as a primer in the next PCR cycle, amplifying another amplicon (Haas et al., 2011)

discussed below.

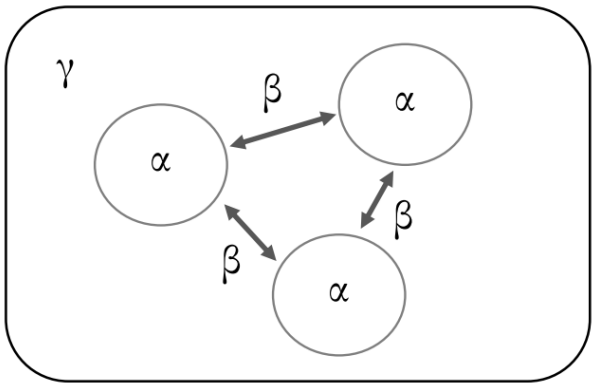
Before 2015, the most widely applied methodology was *de novo* clustering (i.e. using the amplicons only without metadata or reference databases) using pairwise sequence similarities with a 97% threshold for 16S rRNA gene amplicons (as described above in Section 1.2.2.2) or a 98% threshold for 18S rRNA gene amplicons. OTUs are created in an iterative process, where in each step, by grouping the first (typically the longest) amplicon together with other amplicons similar to it (with pairwise sequence similarity surpassing the threshold) into an OTU. In some cases, this methodology has been shown to produce inaccurate results (Nguyen et al., 2016), in part because of the fixed threshold: in a relatively conserved regions, multiple species may be grouped together in the same OTU, while in more variable regions members of the same species may span multiple OTUs. These issues mostly affect deep taxonomic level analysis, with high-level taxonomy used for overall community analyses relatively unaffected. Alternative strategies have been developed to tackle the issue for deep taxonomic level analysis, including Oligotyping (Eren et al., 2013), distribution-based clustering (Preheim et al., 2013), cluster-free filtering (Tikhonov et al., 2015), Swarm (Mahé et al., 2014) and DADA2(Callahan et al., 2016). Oligotyping and DADA2 are methods to tackle very deep variations *i.e. at the species level* in the communities and therefore are not required for overall community analyses. Similarly, distribution-based clustering, which uses a different distance metric, produces results that are not very different from the traditional similarity sequence approach when it comes to higher taxonomic level analyses (Preheim et al., 2013). Cluster-free filtering has been developed for cross-sample analyses and therefore is well suited for time-series analyses or to compare similar communities (at high taxonomical level). However, the process discards low abundance sequences and is unsuitable for population-level alpha or beta diversity analyses (Tikhonov et al., 2015). Finally, Swarm is a *de novo* clustering tool in which similar amplicons are iteratively added together in an OTU until no nearby (up to a distance threshold) amplicons are available to add to the OTU; thus the OTU "grows" until its natural limits are reached. In a way, this is equivalent to using variable similarity thresholds depending on the OTU, with a high threshold for 'tight' OTUs with very similar amplicons and more relaxed thresholds for OTUs with relatively more dissimilar amplicons. While

this is a promising approach that solves the issues raised against the traditional approach, the impact of the improvements on high-level analyses are likely to be minor.

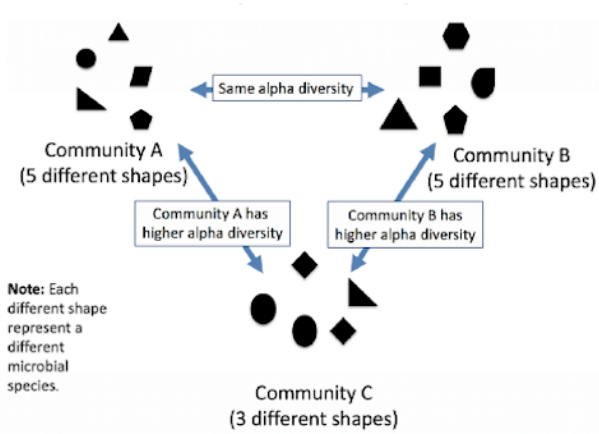
Step4: Metadata and Statistical Analyses The last step in the metabarcoding approach is the bioinformatic analysis of the produced OTU/ASV tables, which typically adopt the double entry table format in which the number of OTU/ASV per sample is available. In addition to these abundances, metadata can be added like predicted taxonomy, function or OTU/ASV statistics. To add these metadata to the OTU/ASVs, their sequences have to be compared with reference databases like Greengenes, RDP, PR2, Silva or others (DeSantis et al., 2006; Larsen et al., 1993; Maidak et al., 1996; Guillou et al., 2013; Quast et al., 2013). Once the metadata are added, statistical analyses can be conducted typically with R (R Core Team, 2017) or Python (`https://www.python.org`) scripts on the OTU/ASV table. These tables are referred as compositional meaning that the element per sample are not independant because of the sequencing process and the library size (Legendre et al., 2005; Gloor et al., 2017; Quinn et al., 2018). Consequently, any comparison or analysis should be preceded by a normalization or a transformation step following a specific strategy depending on the dataset(Weiss et al., 2017). Carrying out statistical analyses of the resulting OTU/ASVs tables can vary considerably according to the scientific hypotheses that the scientist wants to address. Tools like *gusta me* (Buttigieg, Ramette, 2014) are available to help drive the analyses. Nonetheless, some metrics are generally valuable in microbial ecology, such as the alpha- and beta-diversity indices (Figure 1.10; Whittaker (1972)).

The alpha-diversity indicates the species diversity in a single sample. For instance, species richness in a single sample (Figure 1.10), answering the question '*How many species could be detected and identified in this sample?*'. Applied to microbial ecology *via* a metabarcoding approach, this simply means the number of different OTU/ASVs per sample, which would ideally be computed over non-rarefied data (Willis, 2019). Usually, in addition to richness, estimators can be computed. For example, the abundant and rare species ratio. All of these indices are mathematically reviewed in Daly et al. (2018). Additionally, the last alpha-diversity index computed is the evenness. This index is a major component of the understanding of the microbial commu-

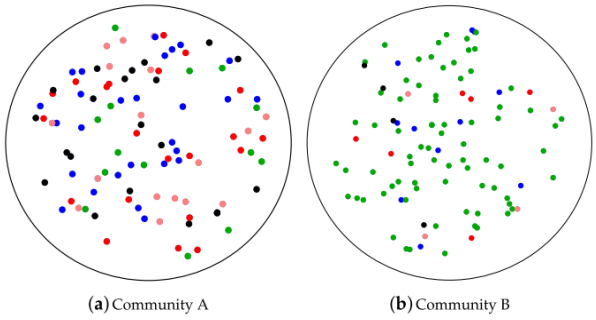
nities as it gives a quick idea of how the community is composed, and in particular if there are major contributors dominating the community or if species abundances are relatively uniformly distributed. The comparison of alpha-diversity indices and metadata variables is important to get a better insight with ecological meaning (Shade, 2017) and can be achieved using ANOVA or Kruskal-Wallis analyses depending on the distribution (normal or not) of the alpha-diversity indices across the samples.



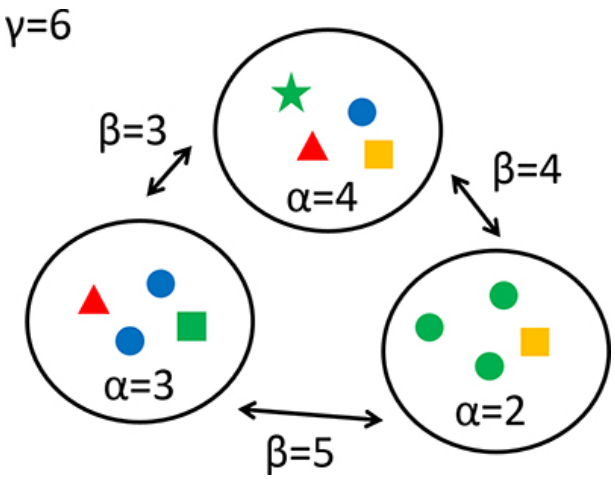
(a) *Whittaker (1972) diversity indices: Gamma, Alpha and Beta-diversity*



(b) *Example of different alpha-diversities in samples*



(c) *Whittaker (1972) diversity indices: Gamma, Alpha and Beta-diversity*



(d) *Example of different alpha-diversities in samples*

Figure 1.10 – Partitioning diversity according to the sampling area scale. Gamma diversity is the total diversity of the sampling area. Alpha diversity is the diversity in a sample in that sampling area. Beta diversity is the diversity between the samples in the sampling area. (sources: (b) https://medium.com/pjtorres-high-gut-alpha-diversity-and-health/high-alpha-diversity-and-health-65e5eca7fa36 ; accessed 17/09/2020) ; (a) and (c) Daly et al. (2018) ; https://oxfordre.com/environmentalscience/view/10.1093/acrefore/9780199389414.001.0001/acrefore-9780199389414-e-33

The beta-diversity is the variation of microbial communities between samples (Figure 1.10), answering the question '*How similar (or how different) are the microbial communities of two samples?*'. In a typical study involving multiple samples, the beta-diversity takes the format of a matrix of pairwise similarity or dissimilarity comparison coefficients. The most common indices are Bray-Curtis (abundance (dis)similarity ; Bray, Curtis (1957)), Jaccard (presence/absence) and UniFrac (phylogenetic and optionally abundance ; Lozupone, Knight (2005)) for microbial communities studies. Then the matrix is generally used by itself or in conjunction with metadata variables within multivariate statistical tests. To statistically verify if there is a particular grouping of samples based on the distance coefficients, one can apply ANOSIM (Clarke, 1993), and to visualise the samples for clustering, one can use the ordination technique NMDS (Non-Metric Multidimensional Scaling ; Kruskal (1964); Clarke (1993)). To statistically test if the distance coefficients are linked with a particular metadata variable(s), PERMANOVA (or NPMANOVA ; Anderson (2001)) are applicable. The equivalent ordination method would be a db-RDA (distance-based ReDundancy Analysis ; Legendre, Andersson (1999)) analysis.

Beside these diversity metrics, graphically, diversity can be displayed directly in the form of barcharts, piecharts or heatmaps of species abundances, gathered into a taxonomic rank allowing data to be readable and understandable. Additionally, analyses can be conducted to look for common or specific OTU/ASVs in a sample or a group of samples.

While metabarcoding presents many opportunities to the field of biology, it has several notable drawbacks (Knight et al., 2018; Pollock et al., 2018; Dopheide et al., 2019; Zinger et al., 2019). Of course, characterizing an entire community is a very challenging task and most of the limitations described below are due to the great diversity in microbial forms of life which makes biology so exciting, difficult, evolving and plural.

The rRna gene, while covering most of the members of the community, is not present in viruses which are therefore not included in metabarcoding analysis. However, metagenocmic approaches allow the recovery of virus particles (see Section 1.2.2.3).

However, the most important issue with metabarcoding may be the reliance on taxonomic

databases. Indeed, with the growing amount of data from the NGS analyses the number of sequences and especially 16S rRNA gene sequences submitted every day to the online database is tremendous. All these datasets need further taxonomical classification and because of the very variable protocols available in metabarcoding sequence analyses, most of the amplicons are deposited as raw data entries as this was already identified as an issue in 1992 (Ward et al., 1992). Therefore there is the need of external laboratories or consortia to use these datasets and create reliable classification of the sequences and remove the chimeras produced during the sequencing process. Indeed, some have risen to the task creating databases and making them available online as is the case with Greengenes, RDP, PR2 or Silva (DeSantis et al., 2006; Larsen et al., 1993; Maidak et al., 1996; Guillou et al., 2013; Quast et al., 2013). However, despite the creation of multiple parallel taxonomy databases, the lack of consensus is an impediment to confidently applying approaches like metabarcoding. This is a persisting issue that would be hard to resolve despite the creation of big consortia lately to improve protists taxonomy (Campo del et al., 2018).

Another issue is the reliance of metabarcoding on PCR amplification and therefore on a single marker gene because of the sequencing limits in term of length of the NGS or Sanger technologies. It has long been known that the PCR primers do not have the same affinity across the species or even kingdoms in the tree of life (Reysenbach et al., 1992). This bias might be reduced as sequencing long fragments by the SMS sequencing technology becomes more widely available, which could allow the design of better set of primers further from each other. These longer reads could also improve the classification of the amplicons by comparing those to full rRNA gene databases (Johnson et al., 2019).

Last but not least, while small rRNAs are not affected by HGT, they have been shown to be present in multiple copies in some taxonomic groups (Kiss et al., 1977; Stoddard et al., 2015) which would bias the metabarcoding analyses in terms of taxonomic profiles and relative abundances of taxa in a given ecosystem.

Overall, although there is room for improvement, metabarcoding is a useful approach for deciphering the microbial diversity in ecosystems and has been largely used to do so over the last twenty years. Metabarcoding can also highlight microorganisms that co-occur in an ecosystem or microbial organisms with physical or chemical measured gradients and thus drive ecological hypotheses. It is possible to analyse more than a marker gene and benefit from the entire DNA composition of a community using metagenomics.

1.2.2.3 Metagenomics: the simple all-in strategy

The metagenomic approach (also called shotgun metagenomics) can be summarized as the genomic approach targeting an entire microbial community as shown on Figure 1.11. The first study to demonstrate the potential of sequencing whole communities without the use of clones was Breitbart et al. (2002), who tackled all viral particles in a (filtered) planktonic surface marine sample. The total amount of viral DNA was very low, they needed to randomly amplify their viral DNA fragments and then performed shotgun metagenomics. Two years later, Venter et al. (2004) were the first to apply metagenomics to investigate the diversity of an environment, but they chose a low complexity sampling site to allow for the fastidious cloning step in their workflow. The same year, Tyson et al. (2004) successfully binned two near-complete genomes from another low complexity sample, an acid-mine drainage, which marked the first Metagenome Assembled-Genomes (MAGs) in a long ever-growing list of studies. In 2006, Poinar et al. (2006) published the first metagenomic study with sequences produced through NGS technology. This and other improvements, notably in bioinformatic tools and wet lab techniques, opened doors to using metagenomics to investigate more and more complex samples without the need of a cloning step, as was anticipated by many (Kowalchuk et al., 2007; Sleator et al., 2008; Simon, Daniel, 2009). Thus, scientists look for unknown microbial groups or divergent key genes (Wu et al., 2011), hoping to reach a breakthrough like Woese, Fox (1977)'s discovery of Archaea through the new technique of 16S rRNA barcoding. In a way, they have succeeded as the field of microbial ecology has rapidly progressed. For example, Wrighton et al. (2012) uncovered the metabolism of anaerobic uncultured bacteria through the use of recovered MAGs, opening

new insights into the microbial ecology of obscure environments. Even beyond microbial ecology, metagnomics through binning has recently led to the discovery of the Lokiarchaeota (Spang et al., 2015) and other Asgard archaea relatives (Zaremba-Niedzwiedzka et al., 2017), findings with implications for the view of eukaryotes' position on the tree of life and hypotheses on the features of the last eukaryotic common ancestor (LECA). In addition to these purposes of binning MAGs and assessing diversity, the metagenomic approach can be applied to metabolic profiling as shown by Edwards et al. (2006), who performed comparative metagenomics analyses between two adjacent sites from the Soudan Mine in Minnesota with differing biogeochemical environments. Moreover, metagenomics holds the potential to provide insights into the key players in a ecosystem (Vieites et al., 2009), which need not always be the predominant species recovered by approaches like metabarcoding.

How does it work? Briefly, all DNA from the community is extracted as for metabarcoding (see Section 1.2.2.2, Step0 and Step1), fragmented and sequenced using NGS (see Section 1.2.2.1) or recently SMS technologies (Nicholls et al., 2019; Hu et al., 2020), resulting in (usually paired-end) DNA fragments which were initially part of theoretically all species' genomes in the targeted microbial community. After quality filtering, these fragments are either assembled into longer fragments termed contigs (assembly approach) or can be used directly (assembly-free approach) (see Figure 1.11b). Taxonomic and functional profiling of the sequenced community is possible with both approaches, but only the assembly approach allows for gene predictions and genome reconstruction (binning). This workflow has been used for over a decade and have barely changed over the years (Riesenfeld et al., 2004; Gilbert, Dupont, 2011; Sharpton, 2014; Jünemann et al., 2017; Pérez-Cobas et al., 2020). Conversely, the dedicated tools have undergone considerable improvements in terms of quality and speed in order to tackle the ever-growing datasets obtained through NGS technologies (Mande et al., 2012; Vollmers et al., 2017; Pérez-Cobas et al., 2020). Accordingly, many tools are currently available at every step to apply metagenomic approaches in microbial ecology (Figure 1.11b). Above all, the critical steps of assembly (Ghurye et al., 2016; Vollmers et al., 2017; Breitwieser et al., 2018), binning (Mande

et al., 2012; Sedlar et al., 2017) and taxonomical affiliations (Mande et al., 2012; Breitwieser et al., 2018) are in constant motion as new tools are released and updated.

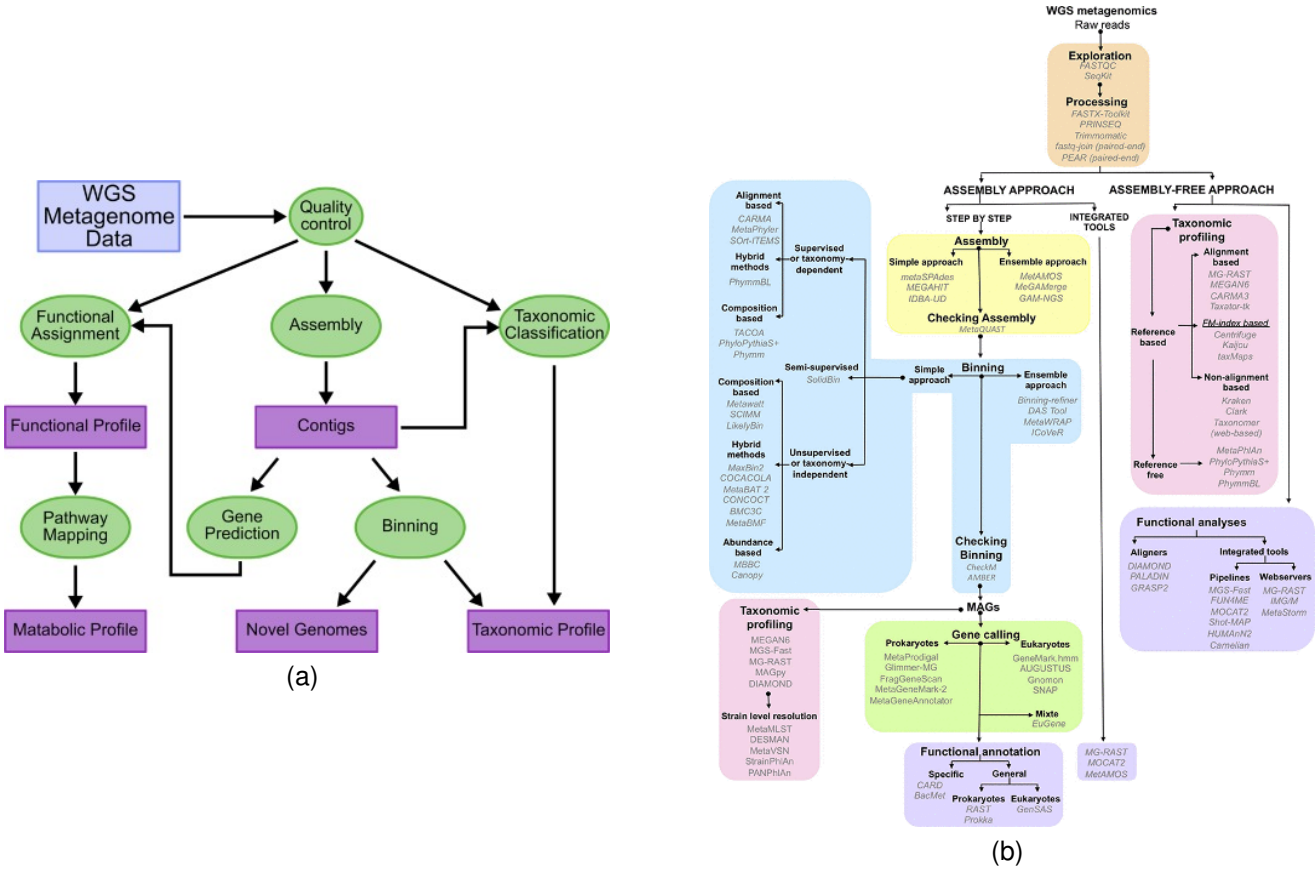


Figure 1.11 – (a) Simple and (b) detailed workflows of standard metagenomics analyses. (a) Purple boxes: data or results; green ovals: processing steps. Note that the direct arrows from 'Quality control' to 'Functional assignment' and 'Taxonimic classification' mark the assembly-free approach, while the paths passing through the step 'Assembly' mark the assembly approach. (b) Bold font: steps; grey italic font: tools used. (source : (a) Jünemann et al. (2017) and (b) Pérez-Cobas et al. (2020))

Phylogenomics In order to explore the phylogenic diversity of a given sample, one can take advantage of the fact that metagenomic approaches make accessible the genomic content. It is therefore a common strategy to look for the sequences of marker genes in the assembled contigs or the raw reads directly, especially for 16S rRNA genes as tools have been designed for this purpose (Lagesen et al., 2007; Miller et al., 2011; Pericard et al., 2018). To learn more about the microbial diversity present in the original sample, one can compute a phylogenetic tree by

comparing these recovered sequences to reference sequences and thus display any close relationships present. The choice of a marker gene to infer the correct phylogenetic tree has been subject to debate and it is possible for two marker genes to lead to different results (Eisen, 1995). This prompted phylogeneticists to start using phylogenomics (multi-marker) instead of phylogenetics (single-marker). Unlike phylogenetic analyses, phylogenomics cannot be directly applied on metagenomic raw reads (as each of those is unlikely to span multiple genes); phylogenomics is therefore only compatible with the assembly approach, after contig reconstruction and binning to form MAGs. Then, the open reading frame (ORF) is predicted and marker genes identified for to phylogenomic analyses, which place the MAG into a reference tree using either the supermatrix method (a phylogenomic tree is inferred using concatenated marker genes) or the supertree method (consensus of several phylogenetic trees, each inferred by its marker gene), with the supermatrix method considered superior for phylogenomic inference (Lang et al., 2013).

Taxonomic classification Taxonomic classification can be obtained using marker genes, mirroring the metabarcoding approach (see Section 1.2.2.2). In addition to providing a set of marker genes which can be used to infer taxonomy through phylogenomic analyses, metagenomic binning can provide new metrics for classification such as average amino acid identity (AAI; Konstantinidis, Tiedje (2005b)) or average nucleotide identity (ANI; Konstantinidis, Tiedje (2005a)). The latter metric has been shown to be useful in determining relatedness between strains. Indeed, Goris et al. (2007) have proposed for ANI to replace DNA-DNA hybridization (DDH), the standard method in delineating microbial species, as the results from the two methods matched very well (with 95% ANI corresponding to the 70% DDH threshold for species delienation). Chan et al. (2012) showed that ANI delivers accurate results in classifying different bacterial strains (from the same genus, Acinetobacter), unlike marker-gene based phylogenomics which were influenced by HGT resulting in some misclassifications. Since then, other studies have confirmed that ANI is a good method to infer taxonomy, preferable to marker genes (Kim et al., 2014; Figueras et al., 2014), and ANI has recently been integrated into a novel taxonomy framework proposal (Parks et al., 2020).

1.2.2.4 Other approaches used in microbial ecology

Metatranscriptomics Metatranscriptomics approaches are similar to metagenomic ones except that they target total messenger RNA (mRNA) instead of total DNA in an ecosystem. Environmental transcriptomic approaches were previously restricted to the use of microarrays (Schena et al., 1995) or transforming mRNA into complementary DNA (cDNA) which needed to be cloned (Poretsky et al., 2005). But with advancing technique on RNA-seq, the fastidious cloning step has been removed and therefore metatranscriptomic analyses could be conducted to gain more insights into microbial ecology especially on species functioning (Leininger et al., 2006; Frias-Lopez et al., 2008). Metatranscriptomic approaches have the major advantage of avoiding contaminant eDNA because sinking or volatile DNA fragments won't be retained nor sequenced. Also, to the extent where it's possible to infer the phylogenetic or phylogenomic trees as well as the metabolic potential through the gene content only, metatranscriptomic can be a good methodological choice. However, metatranscriptomic cannot be assembled neither binned into MAGs because only genes are sequenced. Consequently, all genomic content analyses such as the ones involving gene promoters, RNA derivatives (tRNA, rRNA, siRNA, etc..), genomic islands, codon usage or synteny/operon are incompatible with this technique.

Single-Cell genomics Single-cell oriented analyses have been employed since the beginning of microbial ecology. Initially, single-cells were sorted on plates to grow and create clonal colonies. The growth step was very limiting and only microscopy could confirm growth. Since the 1990's, microbiologists micromanipulate cells and sort them using flow cytometry (Brehm-Stecher, Johnson, 2004; Weibel et al., 2007). Then they usually amplify the low biomass obtained by only one cell using multiple displacement amplification (MDA) (Binga et al., 2008) to ensure enough material is available for sequencing. Single-cell omics (SCO), including single-cell genomics as well as single cell transcriptomics, were elected as the method of the year in 2013 by the Nature publishing group (Nat, 2014) because of its potential to revolutionize microbial ecology (Rinke et al., 2013; Eberwine et al., 2014). Today, SCO are very powerful tools to investigate specific cultured or uncultured taxa from the entire tree of life but also to investigate

interactions between microbes (Yoon et al., 2011; Ku, Sebé-Pedrós, 2019; Santoro et al., 2019).

1.3 Sampling sites

1.3.1 Movile Cave, a million-year-old sealed cave

1.3.1.1 Introduction to Movile Cave uniqueness

Situated in South-East Romania, Movile Cave was discovered in 1986. The cave is partially flooded and fed by thermal sulfidic water and its air is oxygen-depleted, making Movile cave a unique ecosystem Figure 1.12 (Sarbu, Lascu, 1997). Moreover, the cave has been sealed off from the surface for close to 6 million years (Lascu, 1989). The conditions were right for non-photosythetic fauna to thrive and for species to adapt, with ∼30 of the described invertebrate species (∼70%) endemic to the cave (Sarbu et al., 1996; Fišer et al., 2015).

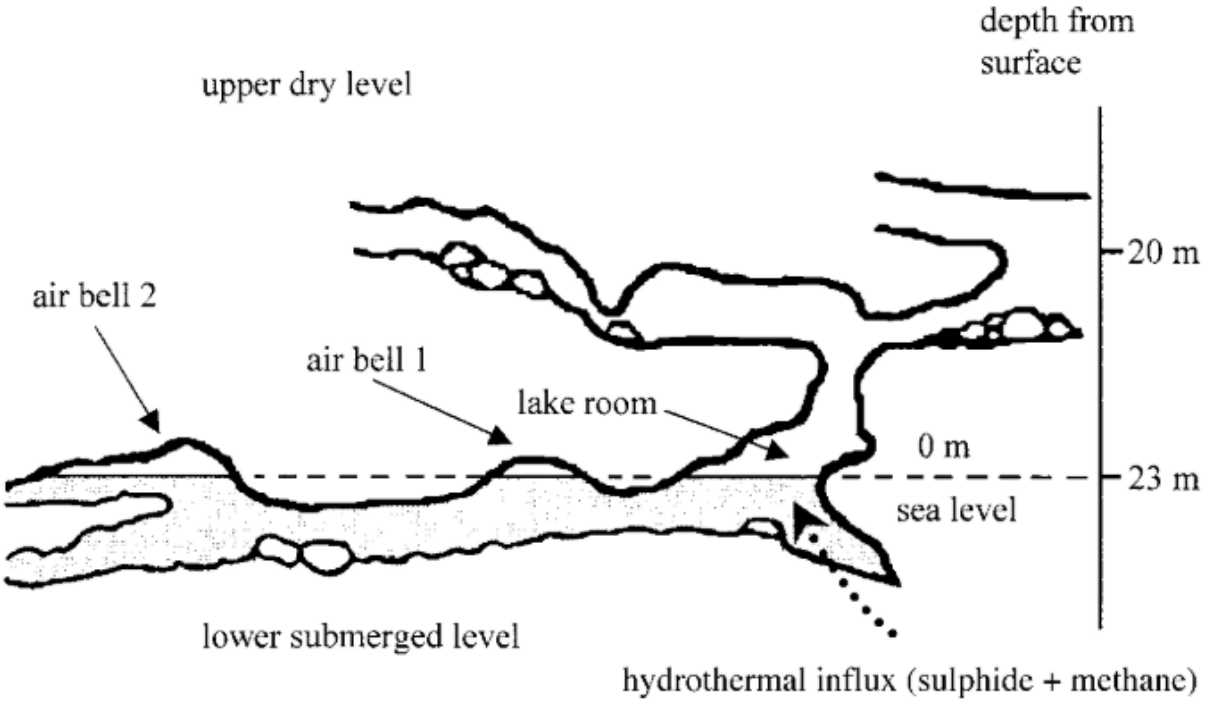


Figure 1.12 – Cross-section of the lake room area of Movile cave. (source : Rohwerder et al. (2003))

1.3.1.2 Past microbial ecology studies on Movile Cave

The role of microorganisms in the cave has been investigated in order to unravel the specific food web system and its key players. Sarbu et al. (1994) identified sulfide-oxydizing chemoautotrophic bacteria as the primary producers in this enclosed ecosystem and later showed the *in situ* autotrophic production sustaining life for many different (micro-)organisms (Sarbu et al., 1996). Rohwerder et al. (2003) discovered that the majority of the metabolic activity takes place on microbial mats floating on the water surface and harboring the elemental sulfur and the primary producers *i.e.* sulfur oxidisers. The authors also highlighted the importance of methylotrophic bacteria, ubiquitous in the cave and later shown to use methylated amines as a carbon source (Wischer et al., 2015). Hutchens et al. (2004) also identified aerobic methanotrophs using the methane present in the *air bells* (see Figure 1.12) as the other key primary producers. Both methylotrophs and methanotrophs have since been isolated and genome sequenced (Ganzert et al., 2014; Kumaresan et al., 2014). In terms of interations, Flot et al. (2014) investigated an association between Amphipods (Niphargus genus) and Thiothrix sulphur-oxidizing ectosymbiotic bacteria. Chen et al. (2009) were the first to conduct broad microbial analyses in Movile Cave. The authors applied a barcoding approach to their samples using multiple marker genes: bacterial and archaeal 16S rRNA, RuBisCO, soxB and amoA. Their results confirmed the chemolithoautotrophic life in the cave and suggested that ammonia- and nitrite-oxidizing may play a bigger role than initially believed. However, microbial eukaryotes have not at all been investigated, with the exception of a study on the diversity of cultivable fungi from the cave (Nováková et al., 2018).

1.3.2 Lake Baikal, a unique water body on Earth

Lake Baikal (Figure 1.13), formed more than 25 million years ago, is the oldest lake on earth. Scientists have been studying lake Baikal since the 18th century, mostly because its originality among water bodies is captivating; Mikhail Kozhov, Brooks (1965) described it as: *a body of water which on the one hand can be considered as a marvellously old and complex lake, and on*

the other as a marvellously simplified ocean.

(a)

(b)

Figure 1.13 – (a) Lake Baikal during ice-cover period and (b) Lake Baikal in July. (source : (a) https://news.algaeworld.org/2016/12/life-thrives-under-ice-covered-lakes/ and (b) Yours truly)

1.3.2.1 Introduction to lake Baikal uniqueness

Lake Baikal is located in Siberia in Russia (Figure 1.13). Attesting to its geological and biological uniqueness, the lake has been a designated UNESCO world heritage site since 1966.

Relative to other lakes (see Figure 1.14), lake Baikal is the deepest (∼1600m) and also the deepest on average (∼750m) freshwater lake on earth followed by the lake Tanganyika in Africa. Reaching ∼1300m below sea level at its deepest, Baikal is also the deepest continental depression on earth just before the Caspian sea (Mikhail Kozhov, Brooks, 1965; Zemskaya et al., 2020). Lake Baikal also contains more than 200 km^3, which corresponds to approximately 20% of the planet's total unfrozen freshwater volume (Sherstyankin et al., 2006). Moreover, lake Baikal ranked second of all freshwater lakes behind Tanganyika in terms of length (∼650km) and sixth in terms of total area (∼32 km^2) behind lakes from North America and Africa.

Geographically, Lake Baikal is divided into three basins of relatively similar size: the Northern, Central and Southern basins. These are delimited by the Academician ridge and the Selenga river (major inflow river) deltas, respectively (Mats, Perepelova, 2011; Touchart, 2012). Because of its high latitude, the lake is frozen from January to May despite recent climate change (Fig-

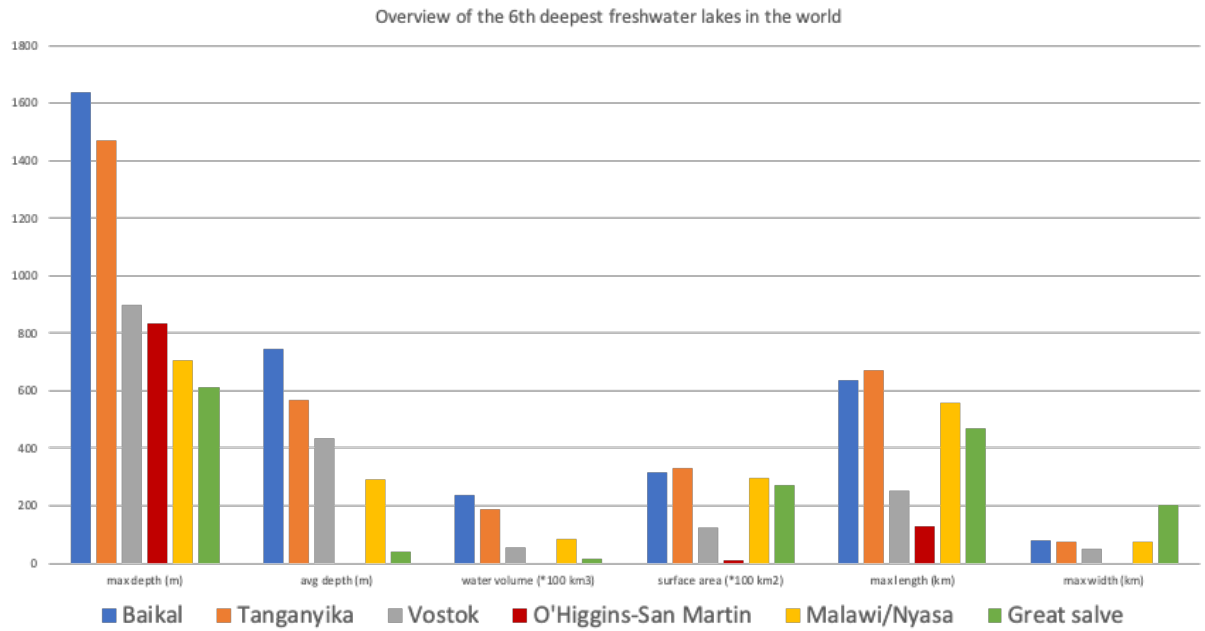


Figure 1.14 – Geographical parameters of the six deepest freshwater lakes on Earth. (source: produced using R with data from `https://web.archive.org/web/20200209063201/https://marine.rutgers.edu/~cfree/ancient-lakes-of-the-world/`)

ure 1.13a; Hampton et al. (2008); Piccolroaz, Toffolon (2018)). This frozen period coupled with strong winds is a very important process for the lake ecosystem. Indeed, it ensures deep-water renewal by coastal downwelling and therefore the stable cold temperature of water in the lake around 4°C as well as the presence of oxygen in the lake water at high depth and the low sedimentation rate (oligotrophic water) (Hohmann et al., 1997; Schmid et al., 2008; Moore et al., 2009; Shimaraev et al., 2011; Troitskaya et al., 2015; Klump et al., 2020). The downwelling process has been shown to represent 50% of the intrusion of oxygenated cold water reaching the bottom of the lake, while the other half can be accounted for by spring events (30%) and under ice events (20%) (Tsimitri et al., 2015). The low water temperature (here due to downwelling) and the high pressure (here due to the depth) are known to facilitate solid phase methane. Consequently, Lake Baikal is so far the only lake harboring methane discharge sites (De Batist et al., 2002; Granin et al., 2019).

Isolated for quite some time, lake Baikal is a great reservoir of endemic species; indeed, 1455 out of 2595 described species (∼60%) are endemic to lake Baikal (Yu Sherbakov, 1999). Many of those have been very well studied especially metazoan and in particular crustaceans. For

example, phylogenetic analysis by Itskovich et al. (2008) confirmed that the family of sponges Lubomirskiidae is endemic to Baikal. Stelbrink et al. (2015) found that a limpet (metazoan) radiation was adapted to the hydrothermal vents and the oil seep in lake Baikal. Such evidence for intra-lacustrine diversification keeps growing with investigations into endemic metazoan Baikal species such as the endemic mollusc from which mitochondrial genomes were recently recovered (Peretolchina et al., 2020). Interest into microbial species and diversity also manifested very early, especially protists were the object of studies from the early 1900s as reviewed by Mikhail Kozhov, Brooks (1965), which due to the available technology at the time, mainly microscopy, were mostly descriptive (Zemskaya et al., 2020).

1.3.2.2 Past microbial ecology studies on lake Baikal

Since the 1990s, many studies started employing the newly developed 16S rRNA gene based approach (see Section 1.2.2.2) to investigate the Baikal microbial communities (Bel'kova et al., 2003). Bel'kova et al. (1996) investigated deep-water prokaryotes and were the first to explore the interesting and very diverse prokaryotic community thriving in the deep waters of lake Baikal, while Semenova, Kuznedelov (1998) focused on planktonic communities. Then, other analyses were carried out on different layers of the lake in order to cover the entire planktonic community and compare the layers (Denisova et al., 1999; Glockner et al., 2000). In addition to depth gradient, geographical comparisons were conducted. Bel'kova et al. (2003) sampled the three basins up to their deepest point (for the Central and North basin) in order to maximize the sampled diversity in terms of cultivable and non-cultivable prokaryotic species. Since then, thanks to the NGS technologies coupled with metabarcoding approaches (see Section 1.2.2.1) many studies have characterized planktonic microbial communities. These include investigations into prosthecate bacteria phylogeny (Lapteva et al., 2007), dinoflagellate (protist) diversity (Annenkova et al., 2009) or dinoflagellates associated with sponges (Annenkova et al., 2011), temporal analyses on phytoplankton differentiating the lake Baikal basins (Mikhailov et al., 2015), temporal and depth dynamics of bacterioplankton (Kurilkina et al., 2016), adaptation and impact of diatom to climate change (Roberts et al., 2018), viral and bacterial communities of coastal water (Butina et al.,

2019), co-occurrence networks using bacterial and eukaryotic OTUs (Mikhailov et al., 2019) or the comparison of free-living, particle-associated bacterial communities (Bashenkhaeva et al., 2020) or planktonic protists origins (Annenkova et al., 2020). Complementary, NGS coupled with metagenomic approaches have been recently been applied to study the planktonic communities in lake Baikal (Cabello-Yeves et al., 2017; Cabello-Yeves, Rodriguez-Valera, 2019; Cabello-Yeves et al., 2019).

Before 2005, only geological studies had been conducted on lake Baikal sediments. Importantly, these studies highlighted that lake Baikal harbors gas hydrates discharge sites (De Batist et al., 2002; Granin et al., 2019). These extremely interesting samples were first analyzed by Shubenkova et al. (2005), who conducted the first phylogeny of microorganisms using 16S rRNA gene clones. The same group later published an updated method to better extract the total DNA from sediments, which would allow one to investigate the ecology of lake Baikal sediments (Chernitsyna et al., 2008). In both reports, they amplified mainly methanotrophic bacteria (Proteobacteria) as well as a few archaeal related sequences. In the second study, they also reported that the deep layers of these gas hydrate discharge sediments harbor divergent sequences, which clustered together in a distinct branch on the phylogenetic tree related to the genus *Pseudomonas* (Gammaproteobacteria). Two years later, the same research group published a study using clones of 16S rRNA genes targeting bacterial and archaeal domains (Zemskaya et al., 2010). From a community thriving close to a gas hydrate bearing sediments, bacterial clones were identified to be from phyla Delta- and Gamma-Proteobacteria, Chloroflexi and OP11, while archaeal clones were from Crenarchaeota and Euryarchaeota (Methanosarcinales order) phyla confirming a thriving methanotrophic lifestyle.

As it became widely available, NGS technology was applied to characterize microbial diversity and ecology in lake Baikal sediments associated with oil seeps or gas discharge. First, Kadnikov et al. (2012, 2013) used a 16S rRNA gene pyrosequencing approach to target the microbial communities of natural methane hydrate and oil seeps sediments. As with previous cloning methods, they retrieved mainly Proteobacteria sequences (*Pseudomonas* - Gammaproteobacteria and uncultured groups – Alpha- and Beta-Proteobacteria) and methanogenic Archaea (*Methanosarci-*

nalles - Euryarchaeota phyla). They also identified new archaeal groups and compared their results to marine environments. They found high similarity in the composition while comparing the bacterial community but the opposite while looking at the archaeal community. Subsequent studies confirmed that aerobic and anaerobic methane oxidation (AOM) take place in both gas-saturated and gas hydrate-bearing sediments of lake Baikal, which decreases the methane flow released into the water column (Pimenov et al., 2014; Lomakina et al., 2014; Zemskaya et al., 2015; Chernitsyna et al., 2016; Bukin et al., 2018; Lomakina et al., 2018). Nonetheless, the major contributors which typically carry out AOM in similar marine environments such as Archaea ANME group 1-3 were not found in lake Baikal. Recently, Lomakina et al. (2020) determined that lake Baikal AOM was instead carried out by NC10 bacteria and ANME2-d Archaea-related species. The same authors investigated the ability of some species from current microbial communities associated with gas discharge to live under thermobaric conditions for several months and hypothesized that these species could have migrated through fault zones together with gas-bearing fluids (Bukin et al., 2016; Pavlova et al., 2019).

The Zemskaya research group never investigated the eukaryotic components of their samples. Only diatoms in lake Baikal sediments have been studied, mainly by classical methods (Kulikovskiy et al., 2011) until the first microbial ecology study related to protists by Yi et al. (2017). They carried out metabarcoding analyses of samples taken in the South basin of lake Baikal ranging from shallow to deep water as well as upper to deeper sediments underneath the water column. This study also revealed for the first time protists from lake Baikal closely related to protists previously only known in marine ecosystems.

Viruses in lake Baikal have also been the object of studies. Butina et al. (2010) were the first, investigating the diversity of T4-like bacteriophages (common planktonic viruses with hosts ranging from Proteobacteria to Cyanobacteria) and their phylogeny in shallow waters using the g23 marker gene (major capsid protein gene) through cloning and Sanger sequencing approach. Recently, the authors updated their g23 marker gene analyses with more sampling sites and different depths, updated protocols and using NGS and metabarcoding approach (Potapov et al., 2018). Reassuringly, they found that $\sim$20% of their sequences matched the clones from the

initial study and ∼60% closely related to phages from other studies of freshwater lakes in France (Le Bourget and Annecy). Meanwhile, metagenomic analyses of Lake Baikal have also shed light on its viral particles. First, analysis of sub-ice waters revealed a phage putatively infecting widespread freshwater bacterium *Polynucleobacter sp.* (Cabello-Yeves et al., 2017). Then, along with a metabarcoding approach targeting the prokaryotic communities of coastal waters, Butina et al. (2019) sequenced the total viral DNA of the same samples. This revealed an unexpected diversity of putative microbial viruses and highlighted the lack of information of viral particles from freshwater ecosystems in the reference databases. Virus exploration of the lake has continued with recent studies of two metaviromes, a holobionts of diseased endemic sponges of lake Baikal (Butina et al., 2020) and pelagic water of lake Baikal (Potapov et al., 2019), and the first discovery of a crenarchaeal phage which was found *adapted* to a freshwater environment (Section 1.3.2.3) (Coutinho et al., 2020).

1.3.2.3 Lake Baikal and Freshwater–Marine transitions

Lake Baikal harbors certain geographical and geological properties, through the lens of which it is more similar to a sea than to a common freshwater lake (see Section 1.3.2) including the only gas hydrates ever found in freshwater lakes while they are common to marine habitats (De Batist et al., 2002). Interestingly, microbial communities associated to these gas hydrates and filling the ecological niche in specific habitats are different from the ones found in similar marine ecosystems (Lomakina et al., 2020). In 2017, metabarcoding (Yi et al., 2017) and metagenomic (Cabello-Yeves et al., 2017) approaches respectively revealed protist and bacterial SAR11 marine-like species in lake Baikal. Thanks to metagenomics, Cabello-Yeves et al. (2017) were able to point out that relative to marine proteomes, the isoelectric point of freshwater proteomes is shifted towards basicity, which they confirmed later in a dedicated manuscript (Cabello-Yeves, Rodriguez-Valera, 2019). Very recently, Coutinho et al. (2020) found the same pattern of isoelectric shifting toward basicity in the first described freshwater Crenarchaeal phages relative to marine ones.

1.4 Thesis objectives

This thesis had three major objectives: implementing and testing an in-house metabarcoding pipeline, applying this pipeline to a more complex study and, finally, using metagenomics to explore the metabolic potential and recovering MAGs from undescribed environments.

First, I developed an in-house metabarcoding pipeline from scratch which I first tested on a case study in order to characterize the protist communities in the suboxic karstic Movile Cave, Romania. Then, I made the pipeline available for all members of the DEEM team and our collaborators. This implementation aims to allow up-to-date treatment of metabarcoding datasets and easy cross-comparisons and replicability to the lab ongoing metabarcoding analyses.

The second aim was to apply this pipeline to a more complex analysis: lake Baikal sediments, Siberia, Russia. Indeed, at the beginning of my PhD we carried out a thorough sampling campaign from which we retrieved deep and shallow surface sediments. Metabarcoding would then be used to describe the microorganisms thriving in these sediments and compare the communities according to depth (possible hydrothermal influence) and across the latitudinal N-S transect of the lake (different river inputs and geology). Also, we wanted to see if there was any trace of typical 'marine' microorganisms to confirm recent findings.

Third, on a selection of the deepest Baikal freshwater sediments, we applied a metagenomic approach to depict the communities' metabolic strategies. Indeed, this approach could allow us to shed light on the key players of these communities and infer their metabolic functions in terms of carbon fixation and energy metabolism. We also aimed at comparing these results with data from other freshwater and marine environments to inspect the freshwater-marine transition hypotheses stated by Cabello-Yeves et al. (2019). In addition, this approach could permit the reconstruction of quality metagenome-assembled genomes from the deepest freshwater surface sediments on earth, allowing us to investigate their potential uniqueness, divergence or adaptation.

CHAPTER

2

METABARCODING PIPELINE AND APPLICATIONS

In this chapter, I will first describe the metabarcoding pipeline I developed in the first year of my PhD. This pipeline is established as the default tool at the DEEM laboratory to analyse amplicon datasets for ecological studies. Consequently, I have been involved in some studies involving both eukaryotic and prokaryotic metabarcoding datasets as described in Section 2.3.

2.1 Implementation of a metabarcoding pipeline

2.1.1 Motivation

At my arrival in the laboratory, there was a pipeline developed in-house compatible with pyrosequencing (*454 Life Sciences*) technology, but new data were sequenced using the *Illumina MiSeq* platform. A simple consequence of this change is switching from single-end relatively long read ($\sim$600 nucleotides) outputs to paired-end relatively short read (300 nucleotides) outputs. Another important consequence was that since the two technologies differ greatly in terms of the introduced bias (see Section 1.2.2.1), different tools were required to treat the produced reads.

With the predecessor pipeline in mind, I implemented a new in-house pipeline for metabarcoding analyses, versatile enough to suit the lab datasets and ongoing projects. I also had the opportunity to supervise Ahmed Ben Brahim[1], a second year Bioinformatics Master student, over a 7-month internship in developing a web interface to provide an easier user access to the data.

The new in-house pipeline was developed in a transition period in terms of open-source bioinformatics pipelines. The most popular suite in processing metabarcoding data was Qiime1 Caporaso et al. (2010). However, in 2017 it was already known that support for Qiime1 would be soon discontinued (which it was in January 2018). Indeed, as the methodological limitations of traditional *de novo* clustering (which Qiime1 was based on) were beginning to be exposed (Nguyen et al., 2016), Qiime2 was released (Bolyen et al., 2019) (see detail on clustering at Section 1.2.2.2, step3). This new version was completely renovated and implemented ASV-based novel clustering methods with tools such as DADA2 (Callahan et al., 2016). However, OTU-based tools such as Swarm (Mahé et al., 2014, 2015) were not and are still not been incorporated.

2.1.2 The pipeline explained

The pipeline is implemented in *bash* in conjunction with a *PostreSQL* (`https://www.postgresql.org`) database in which all the sample metadata is stored (see Section 2.1.3 for details on sample

[1] `https://www.linkedin.com/in/abenbrahim/`

uploading). At every step of the pipeline, the output results are uploaded in the database, which allows one to easily track back from OTU to CMR (Clean Merged Reads, described below). The inputs to the pipeline are the sequences in the format *FASTQ paired-end* and a list of identifiers ('Experiment To Run' or ETR), which are unique indices for the combination of sample, the sequencing run, and MID (multiplexed identifier on the sequencing plate) (see Section 2.1.3 for details on ETR). The pipeline also requires the following user-set parameters: the desired merging size (for step 1) of the overlap between the paired-end reads, taking into account the amplicon length; the trimming length or the minimum CMR length (for step 3), and the choice of tool for computing OTUs (step 5) and its corresponding parameters. The pipeline workflow is as follows:

1. The first step is to merge the paired-end reads. To do so, we incorporated *FLASH* (Magoč, Salzberg, 2011), with the user-defined parameter of the minimum and maximum merging size without penalizing. All merged reads (MR) are passed on to the next step.

2. This second step uses the list of ETR. For every ETR in the metabarcoding analysis, both the MID and reversed MID (rMID) sequences are located and pruned from the merged reads using *cutadapt* (Martin, 2011). Then, PCR primers are also removed from these MID-pruned sequences using *cutadapt* (Martin, 2011). The output sequences are called clean merged reads (CMR).

3. CMRs with length greater than a user-defined threshold are pooled together and dereplicated. Dereplicating here refers to grouping together identical CMR (equal length and nucleotide identity) with the option *–derep_fulllength* of *vsearch* (Rognes et al., 2016). The abundance, *i.e.* the number of sequences in each cluster, is a valuable information for the following step of chimeras removal and therefore is computed and saved (*–sizeout* option).

4. Clusters are then *de novo* checked for chimeras using *vsearch* with the *–uchime_denovo* option (Rognes et al., 2016). *De novo* here means that only the clusters' representative sequences, the clusters' abundances and no external information is used to detect

the chimeric clusters. At the end of this step, every cluster has an attributed value of Y (chimeric), N (not chimeric) or ? (borderline case).

5. Next, non-chimeric clusters (N and ?) are grouped into operational taxonomic unit (OTU). To date, the pipeline is compatible with two clustering alternatives: CD-HIT-EST (95%, 97% or 98% cutoffs; Fu et al. (2012), implementing a traditional *de novo* clustering method (presented in Section 1.2.2.2)) and Swarm v2 (Mahé et al., 2015). Note in addition to these 4 options (3 for CD-HIT-EST and 1 for Swarm), the user can instead opt for any user-set of parameters for CD-HIT-EST or Swarm. The abundance of each OTU is computed as the sum of the abundances of its non-chimeric cluster components.

6. Both strands of the representative sequence for each OTU are compared (global pairwise aligment) to databases of reference sequences using *vsearch* (Rognes et al., 2016) with (*–strand both* and *–usearch_global* options). Only the top hit is retained and the alignment metrics (identity percentage, raw score, query cover percentage, number of identical nucleotides, number of mismatches, length of the alignment) between this top hit and the OTU representative sequence are retained as well.

7. The last step is to produce a double entry table of all retained valuables for the ensuing statistical analyses using *PostgreSQL* (`https://www.postgresql.org`) and *R* (R Core Team, 2017). Each line of the table refers to a OTUs, and the columns represent, respectively, the counts per each ETR (each value representing the number of CMRs per OTU for this ETR) followed by the alignment metrics and the information about the top hit reference sequence (public database of origin, corresponding identifier, taxonomy, full DNA sequence).

2.1.3 The web interface

Because the DEEM team is mainly composed of biologists with limited bioinformatics background, we chose to accompany the metabarcoding pipeline with a browser-run graphical user interface. We chose to use the python web framework *Django* (`https://www.djangoproject.`

com) as we found the user experience intuitive and it was relatively easy to set up and implement on the team local servers.

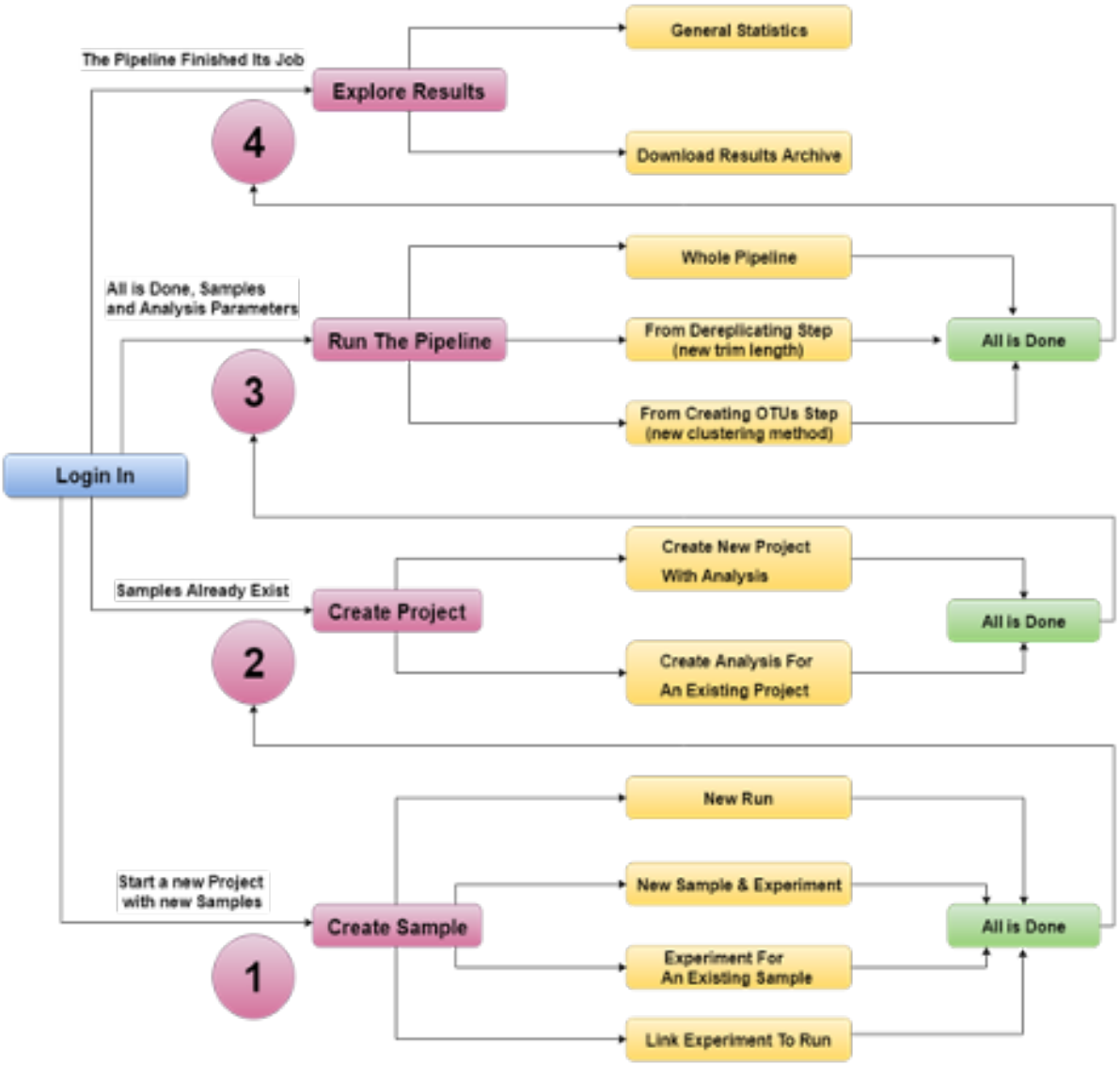


Figure 2.1 – Description of the different step for a user to create and run the metabarcoding pipeline available at the lab. Note that *Sample* in pink block1 should be *ETR*. (source : taken from the intern report and web interface help section

Before being able to do anything, a user needs to log in.

The pipeline first requires a *sample* and its affiliated metadata (same as NCBI SRA submission form) to be uploaded into the database through the corresponding web form. To each sample, one can link one or more *experiments*. In this framework, an *experiment* refers to the molecular biology process of PCR. Each *experiment* is linked to a *sample* and it is possible to

link many *experiment* to the same *sample*. For example, PCRs with primers targeting different marker genes (e.g. one with primers dedicated to a prokaryotic marker gene and another to a eukaryotic marker gene) or different regions of the same marker gene. In addition to these metadata objects, one needs to create a sequencing run (hereafter *run*): an object containing the full path to the sequencing output files as well as information about the sequencing run (if single or paired-end, length of reads), platform and company.

When an *experiment* is uploaded and the corresponding *run* is created, a look-up table is stored in the database, with each line containing a unique identifier, *ETR* ('Experiment To Run'), linking an *experiment* to (its corresponding MIDs in) a *run*.

With the metadata thus organised, a user can create *projects* and *analyses*. A *project* is simply an organisation entity containing the *analyses* for a particular purpose. For example, two analysis under the same project may differ in their sample composition (e.g. on a specific selection of samples) or in the parameters used. The user can create a new *project* or add an *analysis* (a metabarcoding pipeline instance) under an existing *project*. Once an *analysis* is selected the user can launch the metabarcoding pipeline, providing the *ETR*s relevant to the *run*s in the *analysis*.

The web interface allow the user to easily update the database with new *samples* or new *experiments* for existing *samples* or new *runs*. The fact that the linkage resulting into an *ETR* can be done separately allows the user to pre-fill the database with its samples while the sequencing process is still ongoing and the *run* informations are not yet available. Also, it is possible for the user to add new *analyses* to already existing *projects* and to mark the outdated *analyses*. Another major point is that the user can choose to run either the whole pipeline following the steps described above (Section 2.1.2) or, for an existing *analysis*, the user can choose another trimming length to apply on the CMR (see Section 2.1.2, step 3) or another clustering tool (see Section 2.1.2, step 5)

This process and the possible shortcuts or updates are schematized on Figure 2.1.

2.2 Movile Cave case study

In order to test this pipeline in a real scientific context we used samples from the Movile Cave, Romania, (see Section 1.3.1) obtained in 2015 by our collaborator, Alexandra Maria Hillebrand-Voiculescu. This oxygen-depleted environment is interesting because of its unique abiotic and biotic composition. Therefore its planktonic and biofilm samples were ideal to serve as a case study. The published scientific output entitled ***Microbial eukaryotes in the suboxic chemosynthetic ecosystem of Movile Cave, Romania*** (Reboul et al., 2019) can be find below.

environmental microbiology reports

Environmental Microbiology Reports (2019) **00**(00), 00–00 doi:10.1111/1758-2229.12756

Brief Report

Microbial eukaryotes in the suboxic chemosynthetic ecosystem of Movile Cave, Romania

Guillaume Reboul,[1] David Moreira,[1] Paola Bertolino,[1] Alexandra Maria Hillebrand-Voiculescu[2,3] and Purificación López-García[1*]

[1]*Unité d'Ecologie; Systématique et Evolution, CNRS, Université Paris-Sud, Université Paris-Saclay, AgroParisTech, bâtiment 360, 91400 Orsay, France.*

[2]*Department of Biospeleology and Karst Edaphobiology, Emil Racovita Institute of Speleology, Bucharest, Romania.*

[3]*Group for Underwater and Speleological Exploration, Bucharest, Romania.*

Summary

Movile Cave is a small system of partially inundated galleries in limestone settings close to the Black Sea in Southeast Romania. Isolated from the surface for 6 million years, its sulfidic, methane and ammonia-rich waters harbour unique chemosynthetic prokaryotic communities that include sulphur and ammonium-metabolizing chemolithotrophs, methanogens, methanotrophs and methylotrophs. The cave also harbours cave-dwelling invertebrates and fungi, but the diversity of other microbial eukaryotes remained completely unknown. Here, we apply an 18S rRNA gene-based metabarcoding approach to study the composition of protist communities in floating microbial mats and plankton from a well-preserved oxygen-depleted cave chamber. Our results reveal a wide protist diversity with, as dominant groups, ciliates (Alveolata), Stramenopiles, especially bicosoecids, and jakobids (Excavata). Ciliate sequences dominated both, microbial mats and plankton, followed by either Stramenopiles or excavates. Stramenopiles were more prominent in microbial mats, whereas jakobids dominated the plankton fraction of the oxygen-depleted water column. Mats cultured in the laboratory were enriched in Cercozoa. Consistent with local low oxygen levels, Movile Cave protists are most likely anaerobic or microaerophilic. Several newly detected OTU clades were very divergent from cultured species or environmental sequences in databases and represent phylogenetic novelty, notably within jakobids. Movile Cave protists likely cover a variety of ecological roles in this ecosystem including predation, parasitism, saprotrophy and possibly diverse prokaryote-protist syntrophies.

Received 31 December, 2018; revised 26 March, 2019; accepted 8 April, 2019. *For correspondence. E-mail puri.lopez@u-psud.fr; Tel. (+33) 169157608; Fax (+33) 169154697.

Introduction

The Movile Cave harbours a unique underground aquatic ecosystem that has been isolated from the surface for almost 6 million years (Lascu, 1989). Located in a limestone area close to the Black Sea in Southeast Romania, it encompasses several inundated galleries fed by thermal (21°C) sulfidic waters. The first explorations of these galleries showed that some of them contained oxygen-depleted air pockets ('airbells') and floating whitish microbial mats apparently formed of bacteria and fungi (Sarbu *et al.,* 1994; Sarbu *et al.,* 1996). Early stable isotope labelling experiments showed that this subsurface ecosystem is chemosynthetic (Sarbu *et al.,* 1996). Subsequent studies uncovered a wide diversity of prokaryotes and revealed the presence of sulphur- and ammonium-based chemolithotrophy (Chen *et al.,* 2009) but also an important contribution of methanogenesis, methanotrophy and methylotrophy to the carbon cycle in this cave ecosystem (Hutchens *et al.,* 2004; Wischer *et al.,* 2015; Kumaresan *et al.,* 2018). Methanogenic archaea were indeed isolated from floating biofilms (Ganzert *et al.,* 2014) and anoxic sediment (Schirmack *et al.,* 2014).

The chemolithoautotrophic C fixation sustains not only microbial communities but also a variety of obligate cave-dwelling invertebrates, from which more than 30 species are endemic (Sarbu *et al.,* 1996; Fiser *et al.,* 2015). Amphipods are particularly diverse. Species of the prevalent *Niphargus* genus are tightly associated to *Thiothrix* sulphur-oxidizing ectosymbiotic bacteria (Flot *et al.,* 2014). Prokaryote-eukaryote symbioses are widespread in oxygen-depleted ecosystems (Dubilier *et al.,* 2008; Nowack and Melkonian, 2010; Edgcomb, 2016). This

type of symbioses, essential for adaptation to these ecosystems and source of evolutionary innovation, and are particularly widespread in anaerobic microbial eukaryotes (Nowack and Melkonian, 2010; Lopez-Garcia *et al.,* 2017) and might also be prevalent in protists from the suboxic Movile ecosystem. However, the diversity of microbial eukaryotes in this cave is practically unknown. Only a recent, culture-based study provided information about the diversity of culturable fungi in Movile samples (Novakova *et al.,* 2018). This situation mirrors that of other cave ecosystems, which have traditionally attracted interest either on the prokaryotic communities (Northup and Lavoie, 2001) and/or the diversity and specific adaptations of the, very often, endemic animal species (Juan *et al.,* 2010; Casane and Retaux, 2016), while leaving protist diversity largely unexplored.

With the aim to fill this knowledge gap and characterize microbial eukaryotic communities in the chemosynthetic Movile ecosystem, we carried out a study based on high-throughput 18S rRNA gene amplicon sequencing (metabarcoding) of microbial mat and plankton samples from an oxygen-depleted 'airbell' compartment. Our results revealed a considerable diversity of likely anaerobic and/or microaerophilic protists, several of which represent divergent groups from known taxa.

Results and discussion

Movile Cave is a small cave (~ 250 m length) developed in Sarmatian limestone partially flooded with mesothermal (22°–23°C) sulfidic (H_2S, 0.3 mM) water enriched in CH_4 (0.2 mM) and NH_4^+ (0.3 mM). The dissolved oxygen ranges between 9 and 16 μM at the water surface and less than 1 μM below the upper 3–4 cm of the water column, which becomes anoxic towards the bottom (Sarbu *et al.,* 1994; Chen *et al.,* 2009). We collected water and microbial mat samples from the second Movile Cave chamber, more remote from the entry, containing an airbell ('AirBell2'). The atmosphere of this chamber was oxygen-poor (8%–10% O_2) and contained high relative concentrations of CO_2 (2.5%) and CH_4 (2%). As previously described, whitish microbial mats were observed floating on the water surface, sometimes retaining bubbles of reduced gases coming from below (Fig. 1). We collected a fraction of this mat (Mov6, surface of ca. 20 cm^2), which was fixed in ethanol for subsequent DNA purification and 18S rRNA gene metabarcoding analysis (see Supporting Information). A water sample of 0.8 l collected below the surface was prefiltered through a 200 μm mesh to eliminate large particles and the planktonic biomass was retained in 0.2 μm pore size filters (Mov4). We also included in our study a sample of a microbial mat collected in AirBell2 two months earlier and maintained in culture in a sealed cave water-containing bottle in the laboratory (Mov2).

After DNA purification, we amplified 18S rRNA gene fragments of approximately 550 bp length comprising the hypervariable V4 region using primers EK-565F (5′-GCAGTTAAAAAGCTCGTAGT-3′) and 18S-EUK-1134-R-UNonMet (5′-TTTAAGTTTCAGCCTTGCG-3′) tagged with different molecular identifiers for each sample. After mixing the products of several independent PCR reactions to minimize amplification biases, we purified, pooled, and sequenced amplicons using MiSeq paired-end (2 x 300 bp, chemistry v3) Illumina technology. We merged and treated paired-end sequence reads using an in-house bioinformatic pipeline to check quality

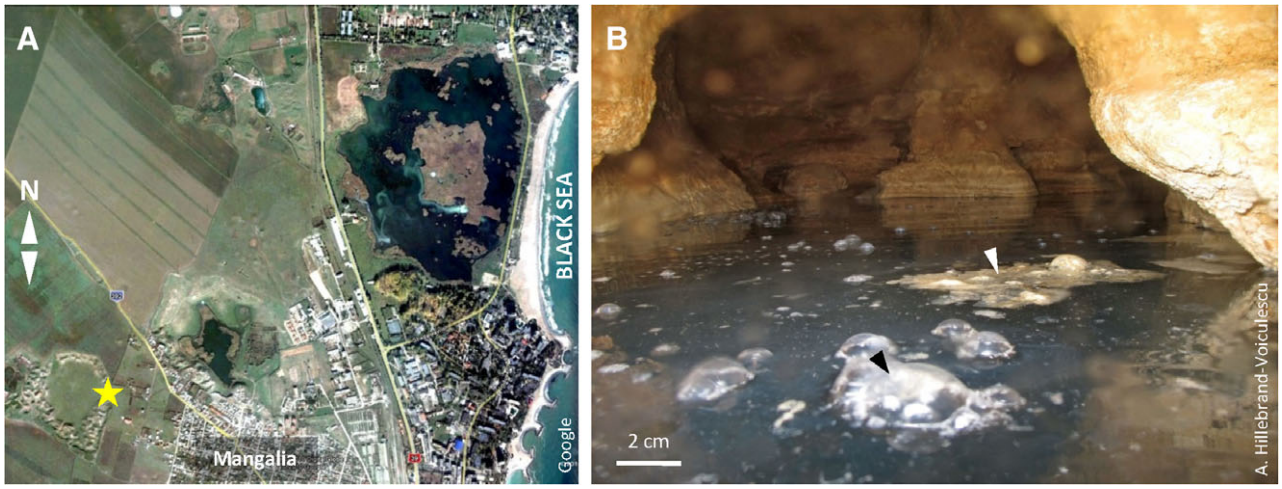


Fig. 1. Sampling at Movile Cave.
A. Location of Movile Cave in the vicinity of Mangalia village and the Black Sea. The entrance of Movile Cave is indicated by a yellow star.
B. sampling site at 'airbell 2'. The sampled floating biofilms are indicated by a white arrowhead. The black arrowhead points at methane bubbles accumulating at the surface of the cave water.

Table 1. Features and sequence statistics of Movile Cave samples analysed in this study. OTU, operational taxonomic unit. CMR, clean merged reads. Clusters refer to groups of 100% identical CMRs.

Sample name	Mov2	Mov4	Mov6
Sample type	microbial mat	plankton	microbial mat
Description	floating microbial mat maintained in lab culture for 2 months	0.2–200 μm fraction (0.8 l)	floating microbial mat fixed after collection
Collection date	06/09/2015	10/11/2015	10/11/2015
Number of raw reads successfully merged	8,281	178,605	137,203
Number of retained reads (CMRs)	8,227	36,133	114,508
Number of Dereplicated CMRs (clusters)	6,945	20,535	46,271
Number of CMRs corresponding to non-chimeric clusters	7,813	34,931	111,419
Number of non-chimeric clusters	6,540	19,443	43,440
Number of 0TUs	633	2,464	4,430
Number of CMRs corresponding to non-singleton OTUs	7,236	32,764	107,361
Number of non-singleton 0TUs	56	297	372

and eliminate primers and molecular identifiers. We also eliminated potentially chimeric sequences (see Supporting Information). We then dereplicated the resulting clean merged reads (CMRs) and used them to define operational taxonomic units (OTUs) at 97% identity cut-off (see Supporting Information). We chose this cut-off value as a good compromise offering a reasonable operational approximation to the genus-species level diversity while producing a manageable number of OTUs to be included in specific phylogenetic trees (see below). Collectively, this yielded a total of 7,454 OTUs, including shared OTUs among samples, but most of them were singletons and were discarded for the rest of the analysis. In total, we retained 652 OTUs (some of them shared between samples) (Table 1). We assigned these OTUs to known taxonomic groups based on their similarity with sequences of a local database that included sequences from cultured/described organisms and environmental surveys retrieved from SILVAv128 (Quast *et al.,* 2013) and PR2v4.5 (Guillou *et al.,* 2013). We further refined the phylogenetic assignation by the phylogenetic placement of our OTU sequences in a reference phylogenetic tree (Supporting Information).

We retrieved OTUs belonging to all major super-groups of microbial eukaryotes including Amoebozoa, Opisthokonta

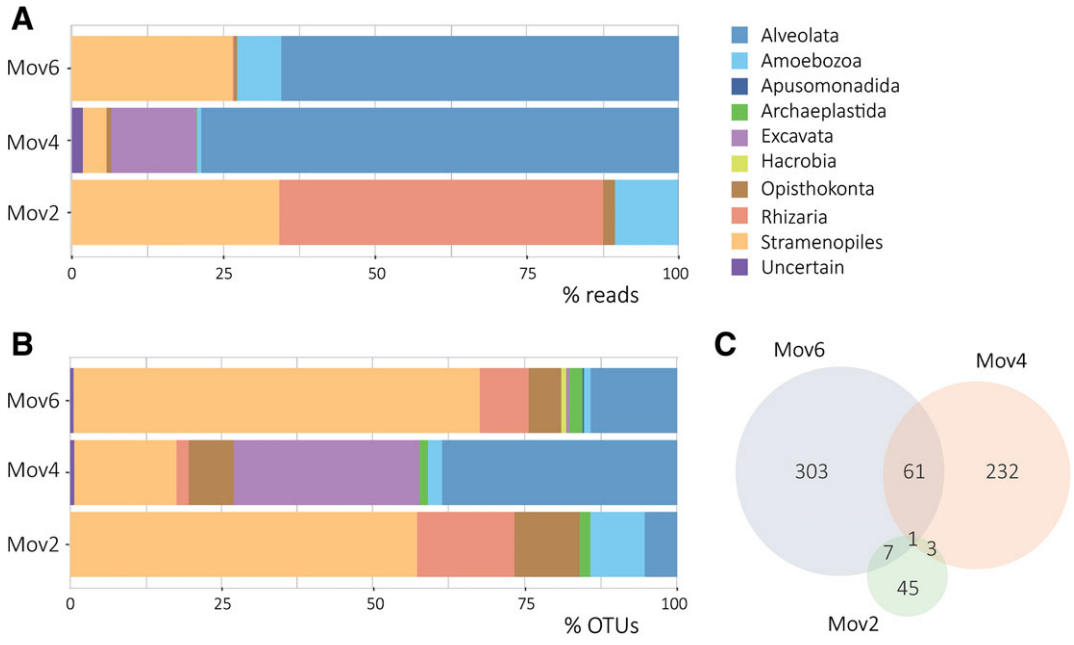


Fig. 2. Relative abundance of microbial eukaryotes in Movile Cave samples.
A. Relative abundance of 18S rRNA gene amplicon reads.
B. Relative abundance of operational taxonomic units (OTUs).
C. Venn diagrams showing specific and shared OTUs among samples.

(including apusomonads), Excavata, Archaeplastida, and the SAR clade (Stramenopiles, Alveolata, Rhizaria), as well as sequences of uncertain classification or belonging to groups of unresolved phylogenetic placement such as haptophytes, katablepharids and telonemids, sometimes referred to as Hacrobia (Okamoto *et al.,* 2009) (Fig. 2). 'Fresh' samples harboured most of the diversity with 372 and 297 OTUs for, respectively, the biofilm Mov6 and the plankton sample Mov4 (Table 1). In both samples, alveolate (ciliate, in particular) sequences dominated (ca. 70%–80%; Fig. 2A) although, in general, OTUs from other groups collectively accounted for a larger diversity (Fig. 2B). However, whereas Stramenopiles, followed to some extent by Amoebozoa, were the subsequent most prevalent groups in the microbial mat Mov6, Excavata were the more relatively abundant in the planktonic Mov4 sample. A small fraction of OTUs was shared by the plankton and the microbial mat samples, highlighting their different community composition (Fig. 2C). The most abundant shared OTUs belonged to Stramenopiles and Amoebozoa (Supporting Information Table S1). As expected, Mov2, the biofilm sample that was maintained in culture for two months in the laboratory, was less diverse and had a different community composition as compared to the 'fresh' sample Mov6. Interestingly, although Mov2 had similar proportions of OTUs across taxa (Fig. 2B), the relative abundance of reads was very different from Mov6 (Fig. 2A). This implies that, although the phylogenetic diversity of OTUs was maintained in mat cultures over time (Fig. 2B), the relative proportion of the different taxa considerably shifted (Fig. 2A). In particular, Rhizaria, and more specifically members of the Cercomonadida, opportunistically proliferated under laboratory conditions. Mov2 and the other two samples shared very few OTUs (Fig. 2C).

In general, OTU sequences retrieved from Movile samples resembled more sequences retrieved from environmental surveys than sequences from cultured/described species, as shown in divergence plots (Fig. 3). These plots also show that, on average, Excavata and Amoebozoa included the most divergent 18S rRNA gene sequences as compared to those existing in databases. Although some ciliate sequences were also divergent, most of them had closer relatives in databases. In order to explore better the phylogenetic diversity within the dominant and most diverse protist groups identified in Movile Cave, we reconstructed phylogenetic trees for Alveolata, Stramenopiles and Excavata. Because our amplicon sequences were relatively short and contained limited phylogenetic information, we first built an alignment of taxon-specific near full-length reference 18S rRNA gene sequences including the closest blast hit sequences to our OTUs with Mafft-linsi v7.38 (Katoh and Standley, 2013) and trimmed gaps and ambiguously aligned positions (Capella-Gutierrez *et al.,* 2009) before building reference trees. Subsequently, we included our OTU

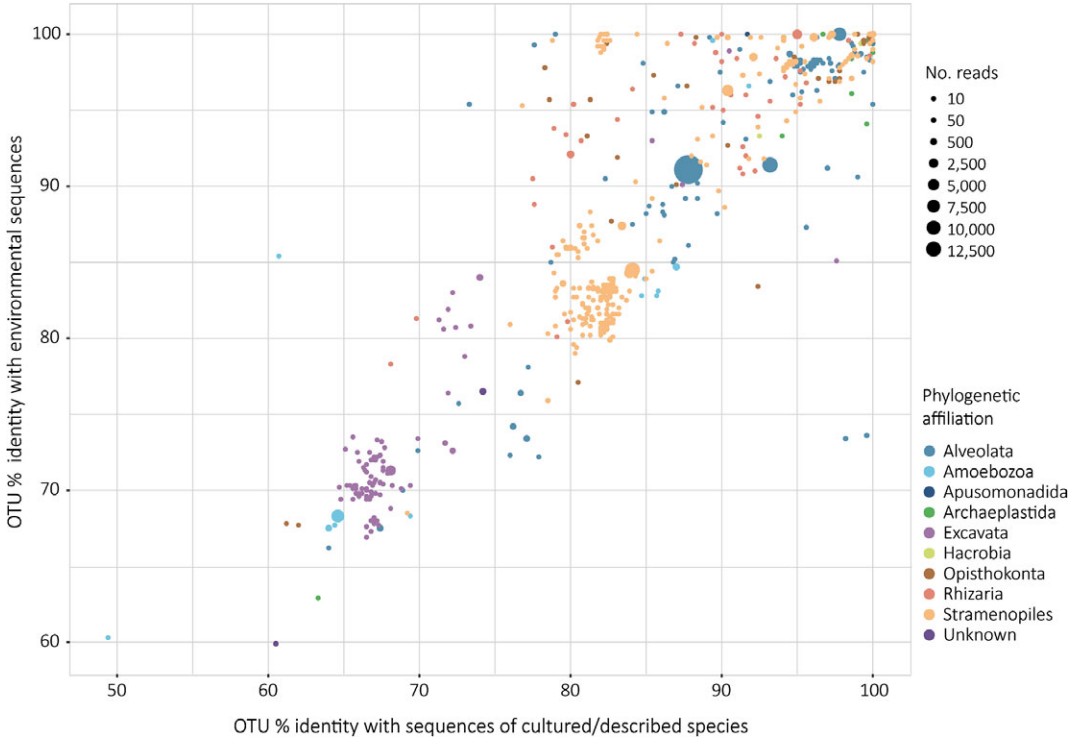


Fig. 3. Divergence plots of eukaryotic OTUs from the Movile Cave with respect to 18S rRNA gene sequences of cultured/described protists and environmental surveys. The size of the dots is proportional to the number of reads. Their colour indicates their phylogenetic affiliation.

sequences to the corresponding alignments using the Mafft-linsi 'addfragments' option. We then reconstructed maximum likelihood phylogenetic trees using IQ-TREE v1.6.5 (Nguyen *et al.,* 2015) applying a GTR model of sequence evolution with a Gamma law and taking into account invariant positions (see Supporting Information).

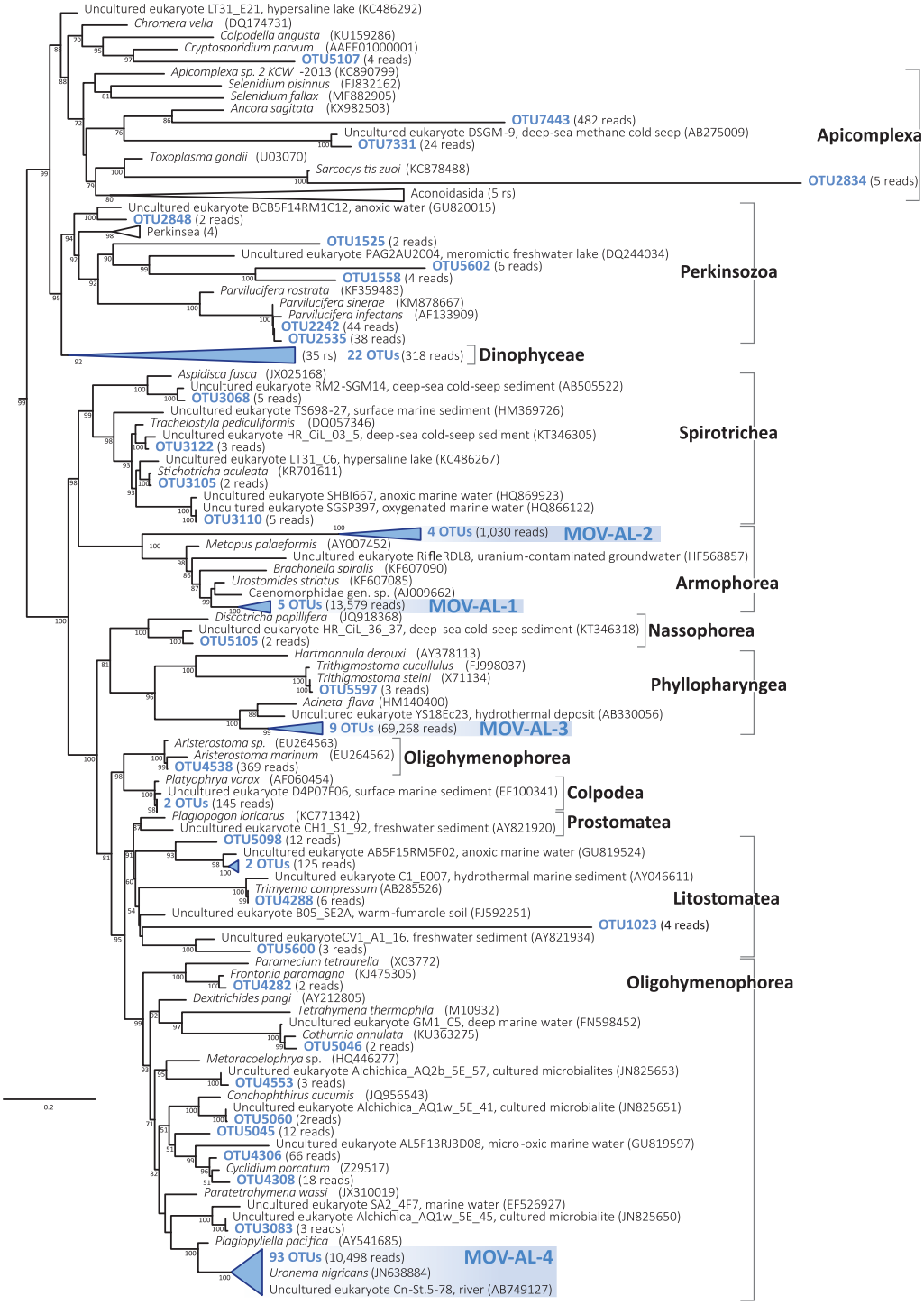


Fig. 4. Maximum likelihood (ML) phylogenetic tree of partial 18S rRNA gene sequences showing the position of OTUs affiliating to Alveolata. The number of reads per OTU or group of OTUs as well as the number of reference sequences (rs) in the case of nodes that have been collapsed (triangles) is indicated. A total of 1,603 unambiguously aligned positions and 284 sequences were used to reconstruct the tree. Bootstrap values higher than 50% are given at nodes. The scale bar represents the number of estimated substitutions per position for a unit branch length. The detailed tree is provided as Supporting Information Fig. S2.

The vast majority of alveolate OTUs corresponded to ciliates, but three OTUs clustered within the Apicomplexa, corresponding most likely to parasites of protists or animals (Fig. 4). The most relatively abundant of them (OTU7443) was distantly related to gregarines (e.g., *Ancora* spp.). We also detected a few OTUs related to the parasitic perkinsids, as well as several dinoflagellate OTUs (22), all of them in very low abundances (Fig. 4 and Supporting Information Fig. S2). Dinoflagellates are typically photosynthetic, although many have lost photosynthesis and become bacterivorous (Boenigk and Arndt, 2002). Many of our OTUs were very similar to environmental sequences from oxygen-deprived settings or deep marine sediments, suggesting that they may be actually heterotrophic (Supporting Information Fig. S2). Other OTUs were more closely related to typical photosynthetic species, and we cannot discard the possibility that they infiltrated from marine waters, given the proximity of the Black Sea, or are low-frequency contaminants introduced during diving (through diving equipment). At any rate, most alveolate sequences were scattered in various ciliate classes (Fig. 4). Three of them contained clades of Movile OTUs that were particularly abundant. The first of them was the class Armophorea, which includes anaerobic and microaerophilic ciliates from diverse environments (Vdacny *et al.,* 2018), often containing prokaryotic endosymbionts (Nowack and Melkonian, 2010). Armophorea encompassed two clades of relatively abundant OTUs that seem related to metopids, a family of anaerobic ciliates, MOV-AL-1 and MOV-AL-2. MOV-AL-2 appeared also forming a clade with metopids but branched at the base of the group and had a longer branch, suggestive of a potential parasitic lifestyle (Fig. 4). The class Phyllopharyngea comprised a clade of nine related OTUs, MOV-AL-3, which was by far the most represented in Movile Cave. MOV-AL-3 likely represents a new ciliate clade, being divergent with respect to their closest relative, a sequence from a hydrothermal deposit in the Mariana Trough. Finally, the class Oligohymenophorea encompassed the largest diversity of OTUs. Many of them were scattered in the class, having as closest relatives sequences retrieved from anoxic or suboxic settings, such as the Cariaco Basin (Edgcomb *et al.,* 2011), the Guaymas hydrothermal sediment (Edgcomb *et al.,* 2002) or the Framvaren fjord (Behnke *et al.,* 2006), and microbialites from alkaline lakes (Couradeau *et al.,* 2011), displaying similar physico-chemical conditions to those of karstic systems. The most diverse clade, MOV-AL-4, comprised 93 OTUs together with one environmental sequence and *Uronema nigricans*, an opportunistic marine parasite of animals (Crosbie and Munday, 1999).

The stramenopiles were also diverse, but most of the OTUs clustered in three major groups, which were also relatively abundant, MOV-ST-1 (bicosoecids, 156 OTUs), MOV-ST-2 (labyrinthulids, 25 OTUs) and MOV-ST-3 (chrysophytes, 20 OTUs) (Fig. 5 and Supporting Information Fig. S3). Although many ochrophytes are photosynthetic (e.g. diatoms or chrysophytes), reversion to heterotrophy has occurred several times independently within this group. Although some diatom and chrysophyte sequences in low frequency might be photosynthetic contaminants introduced during diving, some clades such as MOV-ST-3, and the diatom clades MOV-ST-4 and MOV-ST-5, relatively abundant and related to sequences retrieved from deep-sea or freshwater sediments (Fig. 5), likely correspond to heterotrophic lineages that dwell in the cave ecosystem. However, many other OTUs belong to clear heterotrophic clades, such as the MAST-12 and MAST-3 clades, the saprophytic labyrinthulids or bicosoecids. By far, the most diverse abundant clade, which also included one environmental sequence retrieved from a shallow subtropical lake, was MOV-ST-1. It comprised three major subclusters of OTUs amounting a total of 156 OTUs (Fig. 5 and Supporting Information Fig. S3). In agreement with the local physico-chemical conditions of the cave, and as in the case of alveolates, most of the closest environmental sequences to the Movile OTUs were retrieved from oxygen-depleted habitats or correspond to microaerophilic or anaerobic species.

Excavates comprised very divergent OTUs, with average 18S rRNA gene similarities of approximately 70%–75% (Fig. 3). Many OTUs, in particular the clades MOV-EX-1 and MOV-EX-2, are associated to the family Stygiellidae, which encompasses the genera *Stygiella* and *Velundella.* This is a highly diverse jakobid family whose members typically inhabit anoxic, sulfide- and ammonium-rich marine habitats worldwide (Panek *et al.,* 2015). *Stygiella incarcerata* contains hydrogenosomes, mitochondria-related organelles typical of many anaerobic protists (Leger *et al.,* 2016). However, the most diverse and relatively abundant clade, MOV-EX-3, comprised 76 OTUs and formed an independent lineage with some affinity to jakobids (Fig. 6). This group likely represents either a new jakobid family or a novel euglenozoan lineage.

In addition to alveolates, stramenopiles and excavates, several other taxa were represented in our samples. The most divergent of them corresponded to Amoebozoa (Fig. 3) and were member of the Lobosa or were unassigned (Supporting Information Table S2). This is not surprising given that amoeba have often fast-evolving 18S rRNA sequences and contain insertions. Anaerobic amoeba are relatively poorly known and some of them are so divergent that are usually classified as *incertae sedis* (Taborsky *et al.,* 2017). Among Opisthokonta, we detected OTUs affiliating to apusomonads, metazoans (calcareous sponge), Ichthyosporea, choanoflagellates

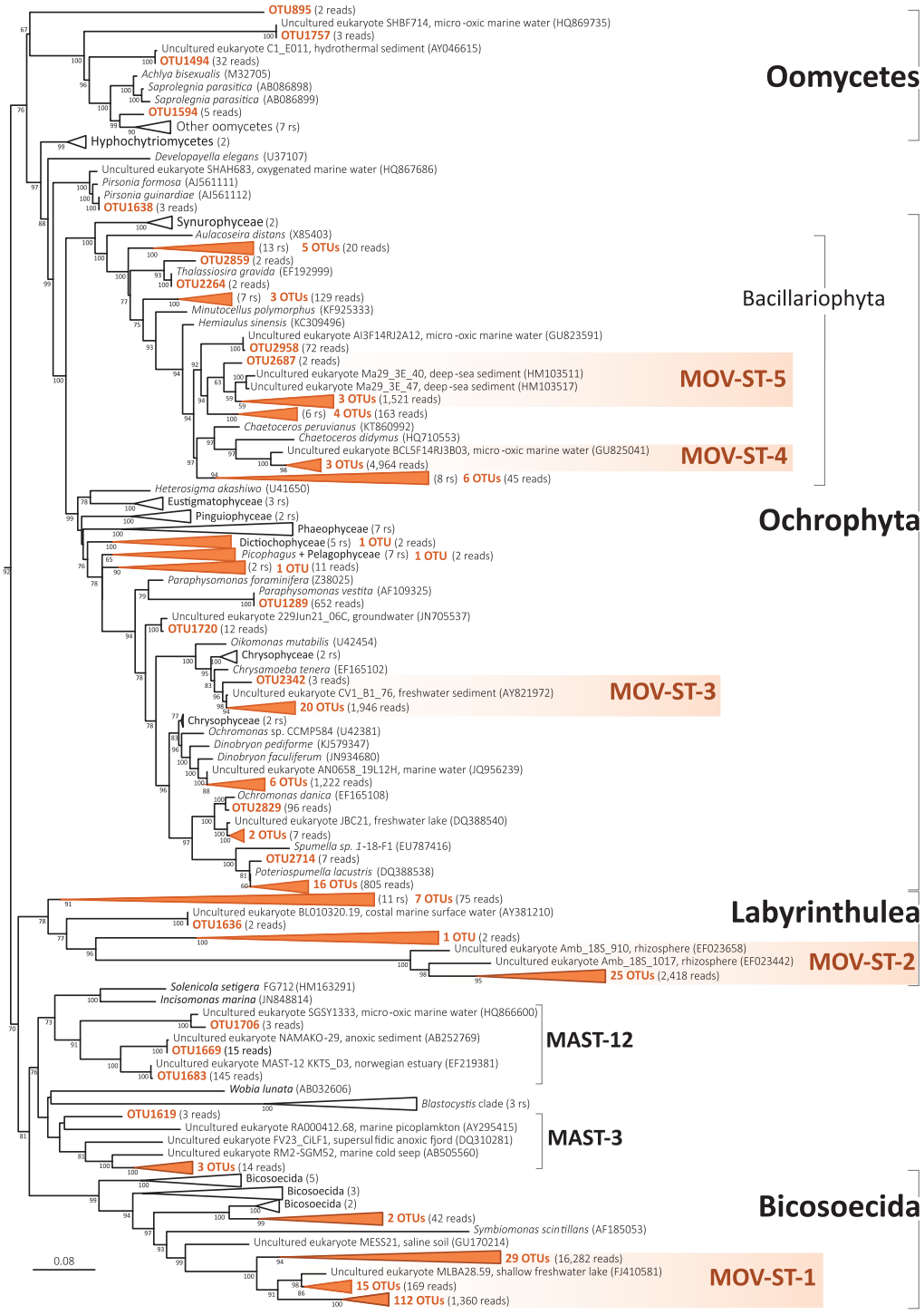


Fig. 5. ML phylogenetic tree of partial 18S rRNA gene sequences showing the position of OTUs affiliating to stramenopiles. The number of reads per OTU or group of OTUs as well as the number of reference sequences (rs) in the case of nodes that have been collapsed (triangles) is indicated. A total of 1,311 unambiguously aligned positions and 446 sequences were used to reconstruct the tree. Bootstrap values higher than 50% are given at nodes. The scale bar represents the number of estimated substitutions per position for a unit branch length. The detailed tree is provided as Supporting Information Fig. S3.

and various fungal and fungi-related taxa (Supporting Information Table S2). Within Rhizaria, we detected one acantharian member and several cercozoan OTUs, notably in the Mov2 sample. This suggests that cercozoa are opportunistic predators that developed better in the laboratory conditions. Finally, a few OTUs represented

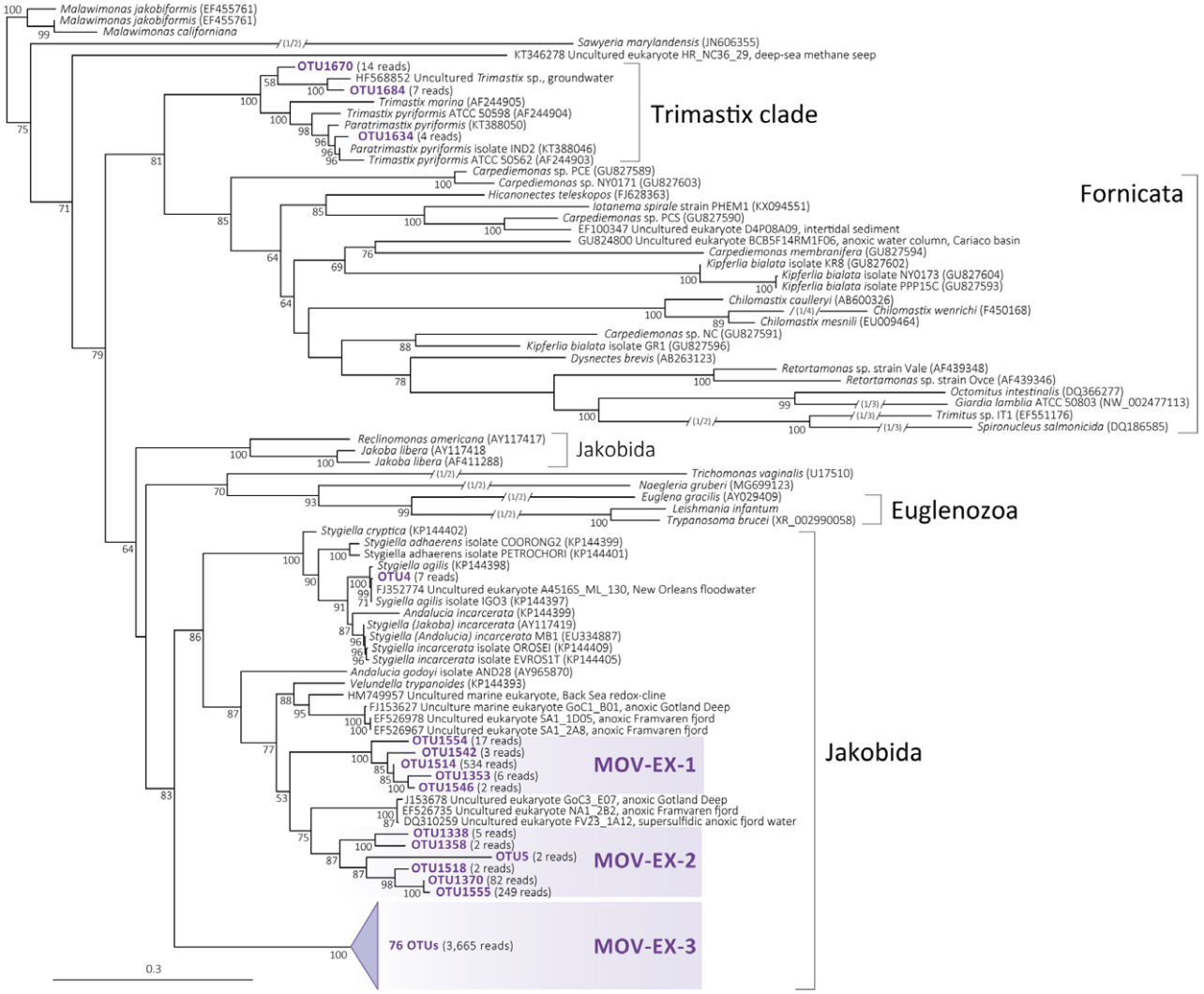


Fig. 6. Approximate ML phylogenetic tree of partial 18S rRNA gene sequences showing the position of OTUs affiliating to excavata. The number of reads per OTU or group of OTUs (triangles) is indicated. A total of 521 unambiguously aligned positions and 153 sequences were used to reconstruct the tree. The scale bar represents the number of estimated substitutions per position for a unit branch length.

by few sequences belonged to Archaeplastida, breviates, prymnesiophytes, telonemids and katablepharids (Supporting Information Table S2). Some of these might be local inhabitants of the cave belonging to the rare biosphere; others, notably those potentially photosynthetic, might be dispersal forms infiltrated from oceanic waters or human-introduced contaminants (e.g. through diving suits).

Our results show that Movile Cave harbours a wide diversity of protists belonging to most major eukaryotic super groups, with ciliates (alveolates), stramenopiles and jakobids (excavates) being the dominant and most varied groups. However, while stramenopiles are more abundant in the floating microbial mats, jakobids seem clearly planktonic, thriving in the oxygen-deprived water column. By contrast, mats cultured in the laboratory for several weeks show protist community shifts, with cercozoans becoming dominant community members. Most of the diversity observed correspond to lineages that have as closest relatives anaerobic or microaerophilic protists or, else, environmental sequences coming from oxygen-deprived habitats. This strongly suggests that Movile Cave protists are mostly anaerobic or microaerophilic. It also seems that protists in the Movile Cave might have both, freshwater and marine, origins. Indeed, the diversity found in this chemosynthetic ecosystem bears resemblance with that of protists found in sulfurous lakes and lagoons, including karstic sites (Triado-Margarit and Casamayor, 2015). At the same time, many of the closest relatives of the Movile OTUs have been identified in anoxic seawater columns (Edgcomb *et al.,* 2011) or sediments (Edgcomb *et al.,*

2002). Given that many of these protists seem anaerobic, it is likely that prokaryote-protist symbioses are prevalent in this chemosynthetic ecosystem. Like in other oxygen-depleted ecosystems (Edgcomb, 2016), Movile Cave protists are thus likely important members of this chemosynthetic microbial ecosystem, covering a range of ecological functions from predation, saprotrophy and parasitism to more subtle hubs of metabolic exchange through syntrophy.

Acknowledgements

This research was funded by the European Research Council advanced grant ProtistWorld (no. 322669) to P.L.G. under the European Union's Seventh Framework Program. A.M. H-V was supported by a grant of the Romanian Ministry of Research and Innovation (CNCS-UEFISCDI; project number PN-III-P4-ID-PCCF-2016-0016, PCCF16/2018). We also thank the COST Action TD1308 'Origins' for facilitating interdisciplinary networking and supporting a short stay of A.M. H-V. in the ESE laboratory at Orsay.

Data accessibility

Sequence data have been deposited in the GenBank Short Read Archive (SAR) under the BioProject ID number PRJNA528591.

References

Behnke, A., Bunge, J., Barger, K., Breiner, H.W., Alla, V., and Stoeck, T. (2006) Microeukaryote community patterns along an O2/H2S gradient in a supersulfidic anoxic fjord (Framvaren, Norway). *Appl Environ Microbiol* **72**: 3626–3636.

Boenigk, J., and Arndt, H. (2002) Bacterivory by heterotrophic flagellates: community structure and feeding strategies. *Antonie Van Leeuwenhoek* **81**: 465–480.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Casane, D., and Retaux, S. (2016) Evolutionary genetics of the cavefish *Astyanax mexicanus*. *Adv Genet* **95**: 117–159.

Chen, Y., Wu, L., Boden, R., Hillebrand, A., Kumaresan, D., Moussard, H., *et al*. (2009) Life without light: microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile cave. *ISME J* **3**: 1093–1104.

Couradeau, E., Benzerara, K., Moreira, D., Gerard, E., Kazmierczak, J., Tavera, R., and Lopez-Garcia, P. (2011) Prokaryotic and eukaryotic community structure in field and cultured microbialites from the alkaline Lake Alchichica (Mexico). *PLoS One* **6**: e28767.

Crosbie, P.B., and Munday, B.L. (1999) Environmental factors and chemical agents affecting the growth of the pathogenic marine ciliate *Uronema nigricans*. *Dis Aquat Organ* **36**: 213–219.

Dubilier, N., Bergin, C., and Lott, C. (2008) Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat Rev Microbiol* **6**: 725–740.

Edgcomb, V.P. (2016) Marine protist associations and environmental impacts across trophic levels in the twilight zone and below. *Curr Opin Microbiol* **31**: 169–175.

Edgcomb, V.P., Kysela, D.T., Teske, A., De Vera Gomez, A., and Sogin, M.L. (2002) Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc Natl Acad Sci U S A* **99**: 7658–7662.

Edgcomb, V., Orsi, W., Taylor, G.T., Vdacny, P., Taylor, C., Suarez, P., and Epstein, S. (2011) Accessing marine protists from the anoxic Cariaco Basin. *ISME J* **5**: 1237–1241.

Fiser, C., Lustrik, R., Sarbu, S., Flot, J.F., and Trontelj, P. (2015) Morphological evolution of coexisting amphipod species pairs from sulfidic caves suggests competitive interactions and character displacement, but no environmental filtering and convergence. *PLoS One* **10**: e0123535.

Flot, J.F., Bauermeister, J., Brad, T., Hillebrand-Voiculescu, A., Sarbu, S.M., and Dattagupta, S. (2014) Niphargus-Thiothrix associations may be widespread in sulphidic groundwater ecosystems: evidence from southeastern Romania. *Mol Ecol* **23**: 1405–1417.

Ganzert, L., Schirmack, J., Alawi, M., Mangelsdorf, K., Sand, W., Hillebrand-Voiculescu, A., and Wagner, D. (2014) Methanosarcina spelaei sp. nov., a methanogenic archaeon isolated from a floating biofilm of a subsurface sulphurous lake. *Int J Syst Evol Microbiol* **64**: 3478–3484.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., *et al*. (2013) The Protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: D597–D604.

Hutchens, E., Radajewski, S., Dumont, M.G., McDonald, I. R., and Murrell, J.C. (2004) Analysis of methanotrophic bacteria in Movile Cave by stable isotope probing. *Environ Microbiol* **6**: 111–120.

Juan, C., Guzik, M.T., Jaume, D., and Cooper, S.J. (2010) Evolution in caves: Darwin's 'wrecks of ancient life' in the molecular era. *Mol Ecol* **19**: 3865–3880.

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

Kumaresan, D., Stephenson, J., Doxey, A.C., Bandukwala, H., Brooks, E., Hillebrand-Voiculescu, A., *et al*. (2018) Aerobic proteobacterial methylotrophs in Movile cave: genomic and metagenomic analyses. *Microbiome* **6**: 1.

Lascu, C. (1989) Paleogeographical and hydrogeological hypothesis regarding the origin of a peculiar cave fauna. *Mics speol Rom* **1**: 13–18.

Leger, M.M., Eme, L., Hug, L.A., and Roger, A.J. (2016) Novel hydrogenosomes in the microaerophilic jakobid *Stygiella incarcerata*. *Mol Biol Evol* **33**: 2318–2336.

Lopez-Garcia, P., Eme, L., and Moreira, D. (2017) Symbiosis in eukaryotic evolution. *J Theor Biol* **434**: 20–33.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B. Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268–274.

Northup, D.E., and Lavoie, K.H. (2001) Geomicrobiology of caves: a review. *Geomicrobiol J* **18**: 199–222.

Novakova, A., Hubka, V., Valinova, S., Kolarik, M., and Hillebrand-Voiculescu, A.M. (2018) Cultivable microscopic fungi from an underground chemosynthesis-based ecosystem: a preliminary study. *Folia Microbiol (Praha)* **63**: 43–55.

Nowack, E.C., and Melkonian, M. (2010) Endosymbiotic associations within protists. *Philos Trans R Soc Lond B Biol Sci* **365**: 699–712.

Okamoto, N., Chantangsi, C., Horak, A., Leander, B.S., and Keeling, P.J. (2009) Molecular phylogeny and description of the novel katablepharid Roombia truncata gen. et sp. nov., and establishment of the Hacrobia taxon nov. *PLoS One* **4**: e7080.

Panek, T., Taborsky, P., Pachiadaki, M.G., Hroudova, M., Vlcek, C., Edgcomb, V.P., and Cepicka, I. (2015) Combined culture-based and culture-independent approaches provide insights into diversity of jakobids, an extremely plesiomorphic eukaryotic lineage. *Front Microbiol* **6**: 1288.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al*. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Sarbu, S.M., Kinkle, B.K., Vlasceanu, L., Kane, T.C., and Popa, R. (1994) Microbiological characterization of a sulfide-rich groundwater ecosystem. *Geomicrobiol J* **12**: 175–182.

Sarbu, S.M., Kane, T.C., and Kinkle, B.K. (1996) A chemoautotrophically based cave ecosystem. *Science* **272**: 1953–1955.

Schirmack, J., Mangelsdorf, K., Ganzert, L., Sand, W., Hillebrand-Voiculescu, A., and Wagner, D. (2014) Methanobacterium movilense sp. nov., a hydrogenotrophic, secondary-alcohol-utilizing methanogen from the anoxic sediment of a subsurface lake. *Int J Syst Evol Microbiol* **64**: 522–527.

Taborsky, P., Panek, T., and Cepicka, I. (2017) Anaeramoebidae fam. nov., a novel lineage of anaerobic amoebae and amoeboflagellates of uncertain phylogenetic position. *Protist* **168**: 495–526.

Triado-Margarit, X., and Casamayor, E.O. (2015) High protists diversity in the plankton of sulfurous lakes and lagoons examined by 18s rRNA gene sequence analyses. *Environ Microbiol Rep* **7**: 908–917.

Vdacny, P., Rajter, L., Stoeck, T., and Foissner, W. (2018) A proposed timescale for the evolution of Armophorean ciliates: Clevelandellids diversify more rapidly than Metopids. *J Eukaryot Microbiol* **66**: 167–181.

Wischer, D., Kumaresan, D., Johnston, A., El Khawand, M., Stephenson, J., Hillebrand-Voiculescu, A.M., *et al*. (2015) Bacterial metabolism of methylated amines and identification of novel methylotrophs in Movile cave. *ISME J* **9**: 195–206.

Supporting Information

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

Fig. S1. Optical microscopy images of Movile Cave microorganisms. A, unidentified biofilm microbes; B, C, *Beggiatoa*-like filaments among smaller prokaryotic cells; D-F, amoeboflagellated eukaryotes. Size bar, 10 μm.

Table S1. Shared protist operational taxonomic units (OTUs) in Movile Cave samples. CMR, clean merged reads.

Table S2. Diversity, abundance and affiliation of OTUs ascribing to eukaryotic supergroups not shown in phylogenetic trees (Figs.).

Fig. S2. Detailed Maximum Likelihood (ML) phylogenetic tree of partial 18S rRNA gene sequences showing the position of OTUs affiliating to Alveolata. The number of reads per OTU is indicated. A total of 1,603 unambiguously aligned positions and 284 sequences were used to reconstruct the tree. The scale bar represents the number of estimated substitutions per position for a unit branch length.

Fig. S3. Detailed ML phylogenetic tree of partial 18S rRNA gene sequences showing the position of OTUs affiliating to Stramenopiles. The number of reads per OTU is indicated. A total of 1,311 unambiguously aligned positions and 446 sequences were used to reconstruct the tree. The scale bar represents the number of estimated substitutions per position for a unit branch length.

Microbial eukaryotes in the suboxic chemosynthetic ecosystem of Movile Cave, Romania

Guillaume Reboul, David Moreira, Paola Bertolino, Alexandra Maria Hillebrand-Voiculescu and Purificación López-García

---

Supplementary Methods

Sampling

Samples were collected from AirBell 2 in the Movile Cave (43°49'32"N, 28°33'38"E), Romania, in 2015. One sample (Mov2, ca. 20 cm^2 surface) of a floating mat was collected in September and maintained in a sealed bottle with water from the cave in the laboratory for two months. The other two cm-scale samples were collected in November and processed after collection. Sample Mov4 corresponded to the planktonic cell-size fraction range of 0.2-200 µm from 0.8 l of Cave water. Sample Mov6 corresponded to a floating mat that was fixed in absolute ethanol (final concentration >80%) after collection (Table 1 and Fig.1).

DNA extraction, 18S rRNA gene amplification and massive sequencing (metabarcoding)

DNA was purified from mat fragments (ca. 300 µl) using the PowerBiofilm™ DNA purification kit and from biomass retained in 0.2 µm size-pore filters using the PowerSoil™ DNA purification kit (MoBio, Carlsbad, CA), following the manufacturer's instruction. 18S rRNA gene fragments of approximately 550 bp length comprising the hypervariable V4 region were amplified using the forward primer EK-565F (5'-GCAGTTAAAAAGCTCGTAGT-3') and the reverse primer 18S-EUK-1134-R-UNonMet (5'-TTTAAGTTTCAGCCTTGCG-3') biased against Metazoa (Bower et al., 2004). Both forward and reverse primers were tagged with 3 different 10 bp molecular identifiers (MIDs) to allow pooling and later differentiation of PCR products from the 3 distinct samples. Amplicons from 5 independent PCR products for each sample were pooled together and then purified using the QIAquick PCR purification kit (Qiagen), according to the manufacturer's instructions. The same amounts (around 200 ng) of purified amplicons from the 3 samples were pooled and sent for sequencing. Amplicons were sequenced using the MiSeq paired-end (2x300 bp) technology from Illumina by Eurofins Genomics (Ebersberg, Germany).

Bioinformatic pipeline for sequence analysis

The paired-end reads obtained after sequencing were treated using an in-house pipeline following a standard protocol. First, the paired-end reads were merged using FLASH v1.2.11 (Magoc and Salzberg, 2011) with --min-overlap and --max-overlap parameters set to 5 and 100, respectively. Then, the full amplicons were check for quality trimming in four steps: i) for each MID, full amplicons were kept only when the corresponding 10-bp MID sequence was found in both paired-end sequences, ii) MIDs and PCR primer sequences were removed from forward and reverse sequences using the multiple command line of cutadapt v1.14 (Martin, 2011), allowing an error rate of 10% and the possibility of cutting up to 5 times the PCR primers (in case of multiple primer joining during PCR), iii) we merged the two paired-end sequences and produced the reverse complement sequence of amplicons for which the reverse PCR primer was found in 5' and forward primer in 3' and iv) we discarded merged amplicons which contained MID sequences within the full-length sequence to avoid possible chimeras. At the end

of this four-step treatment, full amplicons cleaned from MID and PCR primer sequences (named CMR, for clean merged reads) were included in a local database. Only CMRs above 400 nucleotides in length were selected to be dereplicated by clustering CMRs sharing the exact same sequence and length using vsearch v2.3.4 (Rognes et al., 2016), options --derep_fulllength and –sizeout. We retained only one sequence for each cluster while keeping the information about the corresponding sequence abundance. We used vsearch v2.3.4 in de novo mode. Potential chimeras were detected using --uchime_denovo with default parameters for --minh and --xn options. We tagged all the dereplicated CMRs in our database as either non-chimeric (N), chimeric (Y) or potentially chimeric (?). CMRs from the two latter categories were retrieved from the database and clustered using CD-HIT-EST v4.6 (Li and Godzik, 2006; Fu et al., 2012) into operational taxonomic units (OTUs) at 97% sequence identity (options -c 0.97), -n 10 for word length and -g 1 (attribution to the most similar cluster if multiple choices are available for a sequence).

Taxonomic assignation and phylogenetic placement of OTUs

All OTUs were blasted against a home-made database including sequences from cultured/described organisms and from environmental surveys based on SILVAv128 (Quast et al., 2013) and PR2v4.5 (Guillou et al., 2013) using the vsearch v2.3.4 (Rognes et al., 2016) pairwise alignment tool. Only the best hit was retrieved and used as a taxonomic proxy. We refined the assignations of all OTUs that had as best hit either an environmental sequence with less than 80% identity or a sequence from a cultured species with less than 70% identity. To do so, we first built a multiple alignment (mafft-linsi v7.388 (Katoh and Standley, 2013) that contained 157 reference sequences included in a recent update of the tree-of-life phylogeny (Hug et al., 2016) as well as 595 sequences of cultured organisms and 723 environmental sequences identified as best hits to our sequences in the non-redundant GenBank database. Then we trimmed the multiple alignment with trimAl v1.4 (Capella-Gutierrez et al., 2009) option -gt 0.30, allowing to retain at least 70% of non-gap sites among the total number of sequences aligned. Using this trimmed multiple alignment, we inferred a backbone phylogenetic tree using IQ-TREE v1.6.5 (Nguyen et al., 2015) with -m GTR+G+I (GTR + discrete gamma + invariable sites model). Finally, we phylogenetically placed our OTUs within this reference tree using EPA-ng (Barbera et al., 2018) and display the results with FigTree (http://tree.bio.ed.ac.uk/software/figtree/) or Dendroscope (Huson et al., 2007) softwares.

Phylogenetic analysis

Individual phylogenetic trees were done for the most abundant and diverse eukaryotic supergroups in plankton and mats, namely Alveolata, Stramenopiles and Excavata. To do so, we built an alignment of near full-length 18S rRNA gene sequences including selected reference sequences for the respective groups and the closest blast hit sequences to our OTUs with Mafft-linsi v7.38 (Katoh and Standley, 2013). We trimmed the alignment using trimAL (option -gt 0.3; Capella-Gutierrez et al., 2009). We reconstructed maximum likelihood phylogenetic trees using IQ-TREE v1.6.5 (Nguyen et al., 2015) applying the options -m GTR+G+I (GTR model of sequence evolution taking into account a discrete gamma law and including invariable sites). We then incorporated our OTU representative sequences to the corresponding reference alignment using mafft -linsi v7.38 and the ‘addfragments’ option. We then applied trimAL with the same option than before to eliminate gaps and ambiguously aligned positions. The final phylogenetic tree was inferred using IQ-TREE v1.6.5 and the same evolutionary model of sequence evolution as before. The number of positions and sequences retained for the phylogenetic

analysis were as follows: Alveolata, 1,603 sites, 284 species; Stramenopiles, 1311 sites, 446 sequences; Excavata, 521 sites, 153 species.

References

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2018) EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst Biol,* doi: 10.1093/sysbio/syy054. [Epub ahead of print].

Bower, S.M., Carnegie, R.B., Goh, B., Jones, S.R., Lowe, G.J., and Mak, M.W. (2004) Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. *J Eukaryot Microbiol* **51**: 325-332.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150-3152.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L. et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: D597-D604.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J. et al. (2016) A new view of the tree of life. *Nat Microbiol* **1**: 16048.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 460.

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

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

Magoc, T., and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957-2963.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJournal* **17**: 10-12.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-596.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahe, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Table S1. Shared protist operational taxonomic units (OTUs) in Movile Cave samples. CMR, clean merged reads.

Shared in samples	Number of shared OTUs	Number of CMRs	Taxonomic affiliation		Percentage of reads
			Class/Phylum	Super-group	
Mov 6 - Mov4	3	406	Ciliophora	Alveolata	1.09
	3	7,966	n.d.	Amoebozoa	21.46
	1	59	Chlorophyta	Archaeplastida	0.16
	2	21	Metamonada	Excavata	0.06
	7	413	Fungi	Opisthokonta	1.15
	1	13	Mesomycetozoa		
	2	49	Cercozoa	Rhizaria	0.13
	15	16,654	Bicoecea	Stramenopiles	74.34
	18	10,706	Ochrophyta		
	6	218	Other stramenopiles		
	1	11	Labyrinthulea		
	2	598	Unknown	Unknown	1.61
Mov6 - Mov2	1	2	Apicomplexa	Alveolata	99.94
	1	69,251	Ciliophora		
	1	22	Fungi	Opisthokonta	0.03
	4	17	Ochrophyta	Stramenopiles	0.02
Mov4 - Mov2	1	123	Fungi	Opisthokonta	71.1
	1	2	Cercozoa	Rhizaria	1.16
	1	48	Labyrinthulea	Stramenopiles	27.75
Mov6 - Mov4 - Mov2	1	5	Streptophyta	Archaeplastida	100

Table S2. Diversity, abundance and affiliation of OTUs ascribing to eukaryotic supergroups not shown in phylogenetic trees (Figs. 4-6).

Supergroup	Phylum / high-rank taxonomic categories		No. of OTUs	No. of reads	Comments
Opisthokonta	Apusomonadida		1	4	100% identical to clone SHAO486, suboxic sea zone (HQ867866
	Metazoa		1	4	calcareous sponge
	Ichtyosporea		1	13	99% identical to a fish parasite (AB191433)
	Choanoflagellida	Craspedida	3	3-4	
	Fungi and relatives	Ascomycota	12	2-260	
		Basidiomycota	9	2-90	
		Zoopagomycota	1	8	
		Chytridiomycota	2	2-2	
		Cryptomycota	6	2-22	
		Unassigned	3	2	
Amoebozoa	Conosa		1	16	
	Lobosa		6	up to 753	some very divergent OTUs
	Unassigned		5	10-7545	
Rhizaria	Radiolaria	Acantharea	1	15	100% identity with marine Acantharea Ma121_1A29 (FJ032649)
	Cercozoa		41	2-2761	2 dominant OTUs affiliate to Vampyrellida and Cercomonadida
Archaeplastida	Chlorophyta		6	2-59	
	Streptophyta		3	5	
	Unassigned		1	2	
Incertae sedis	Breviatea		1	3	
	Hacrobia	Prymnesiophyta	1	8	99% identitcal to uncultured haptophyte WS071.030 (KP404661
		Telonemia	1	8	99% identitcal to marine clone RA001219.10 (AJ564769)
		Katablepharida	1	5	

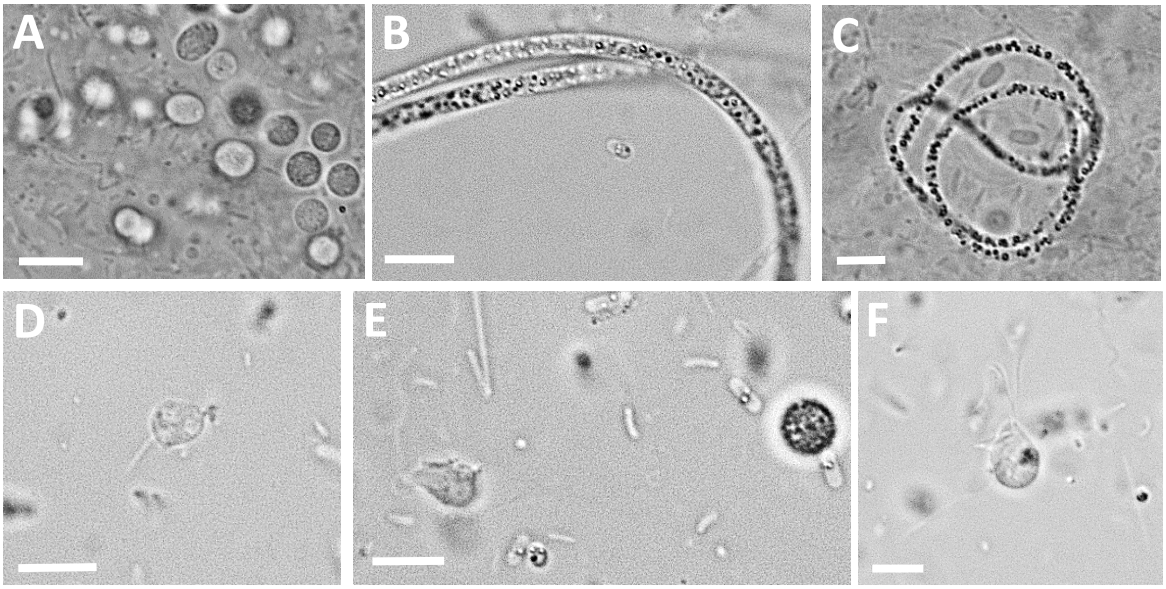


Fig. S1. Optical microscopy images of Movile Cave microorganisms. A, unidentified biofilm microbes; B, C, *Beggiatoa*-like filaments among smaller prokaryotic cells; D-F, amoeboflagellated eukaryotes. Size bar, 10 µm.

2.3 Other applications and scientific contributions

In the development and the support of this pipeline, I was able to collaborate with researchers from the DEEM team itself but also from other research units and laboratories.

2.3.1 Dallol extreme environment diversity (with Jodie Belilla)

This project aimed at exploring the microbial life thriving at Dallol–Danakil, Ethiopia. This area harbors many intriguing sampling sites including some poly-extremes, in other words, combining extreme factors such as very low pH, high salt and high temperature. The goal of the project was to identify whether the physico-chemical conditions determined in a sampling site would harboring life forms and which ones and to therefore shed light on the limits for life and extend this to theories on the origin of life. Environments with high chaotropicity and low water activity were incompatible with the thriving of life. On the other hand, we detected diverse and abundant archaea in hypersaline conditions, which was suprising as it would suggest independent adaptations to hypersalinity. It is possible that some of the thermophilic adaptations described in the deepest archael branches, could be co-opted as adaptations to hypersaline environments. These adaptations, however, are incompatible with extreme acidity, as we found that the combination of high salt ($>35\%$) and low pH ($\sim$0) in the Dallol dome ponds made it inhabitable, possibly due to membrane problems under these conditions.

My contribution to this project was in the metabarcoding data treatment through the pipeline I developed and in bioinformatic assistance, especially for the phylogenetic placement tree on Figure 3.c or on Extended Data Fig. 7.

This work has been published in *Nature Ecology and Evolution* under the title ***Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area***. The article is enclosed in Appendix A (Belilla et al., 2019) of this manuscript.

2.3.2 Dinoflagellates in sand beaches (with Albert Reñé)

For this collaboration, I performed metabarcoding support with the pipeline in the context of two scientific studies.

The first was a methodological study about how metabarcoding analyses of rich complex environments like soils and sediments can be affected by different DNA extraction methods. Both kinds of sample treatment protocols were applied on coastal sediment samples from the Mediterranean Sea and we compared the resulted metabarcoding datasets. The extracting method had a significant impact on the relative abundances of the detected taxa, with the melting seawater-ice elution method resulting in a much higher protist richness estimation than direct lysis. This work was published by the *Environmental Microbiology Reports* under the title ***Performance of the melting seawater-ice elution method on the metabarcoding characterization of benthic protist communities***. The article is enclosed in Appendix B (Reñé et al., 2020) of this manuscript.

The second study describes the diversity, structure and spring-summer temporal dynamics of the benthic dinoflagellate communities in those same sediments. We found characteristic benthic sand-dwelling dinoflagellate taxa clearly distinct from planktonic ones, with many components of the communities corresponding to unknown species. There was also a temporal gradient to dinoflagellate diversity, with higher diversity in spring samples and higher similarity in summer samples. The results are being considered for publication, entitled ***Composition and temporal succession of sand-dwelling dinoflagellate communities from three Mediterranean beaches***.

2.3.3 Mexican lakes microbial mats diversity (with Miguel Iniesto)

Microbialites are rocks formed by microbial communities and are considered as the oldest traces of life on earth to date. However, their formation is still not completely understood, especially the sequestration of CO_2 as biomass and carbonates. In this study, we identified an abundant core microbiome thriving in all microbialites taken from environments with different physico-chemical

parameters.

My contribution to this work was initially the metabarcoding analyses using the pipeline but I also provided help and discussion about the statistical analyses and the use of *R* (R Core Team, 2017).

This paper has been accepted in *Environmental Microbiology* very recently and is entitled ***Core microbial communities of lacrustine microbialites sampled along an alkanility gradient***, enclosed in Appendix C of this manuscript.

2.3.4 Planktonic protists in lake Baikal (with Gwendoline David)

In same sampling cruise in 2017 which started the project on the lake Baikal sediments (Chapter 3), we also sampled water columns across the lake. In this study, we aimed at investigating the abiotic and biotic effects on the protist structure in lake Baikal water columns. We showed that depth has a strong effect on community stratification in contrast to the latitudinal gradient or coastal/open water abiotic parameter, which had little to no effect.

The manuscript, entitled ***Environmental drivers of plankton protist communities along latitudinal and vertical gradients in the oldest and deepest freshwater lake***, has been recently submitted (also available on bioRxiv David et al. (2020)); it is also enclosed in Appendix D of this manuscript.

CHAPTER

3

METABARCODING ANALYSIS OF BAIKAL LAKE SEDIMENTS

3.1 Context and objectives

In this chapter, I will present my study of lake Baikal upper layer sediment using a metabarcoding approach, one I detailed in the previous chapter (Section 2.1.2). The goal of this study was to investigate the effect of environmental parameters on the sediment communities by sequencing both 16S and 18S rRNA genes. The following manuscript has recently been submitted to ISME journal.

3.2 Final version of the manuscript draft

Marine signature taxa and microbial community stability along latitudinal and vertical gradients in sediments of the deepest freshwater lake

Guillaume Reboul[1], David Moreira[1], Nataliia V. Annenkova[2], Paola Bertolino[1], Konstantin E. Vershinin[2] and Purificación López-García[1]

[1] Ecologie Systématique Evolution, Centre National de la Recherche Scientifique - CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France

[2] Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

For correspondence: puri.lopez@u-psud.fr

Running title: Benthic microbial communities of Lake Baikal

Lake Baikal is the deepest (~1.6 km) and most voluminous freshwater reservoir on Earth. Whereas its planktonic communities have been studied to some detail, benthic microbial communities remain poorly explored, as in most freshwater systems. Here, we analyzed the structure of microbial communities associated to sediment upper layers (0-1 cm) across a North-South latitudinal transect covering the three basins of the lake and from littoral to bathybenthic depths (0.5 to 1450 m). Metabarcoding of 16S and 18S rRNA genes revealed rich (74419 prokaryotic and 10563 eukaryotic operational taxonomic units; OTUs) and even communities dominated by rare OTUs. Archaea represented up to 25% or prokaryotic sequences; Thaumarchaeota, Woesearchaeota, Pacearchaeota and Thermoplasmata being more relatively abundant. Among bacteria, members of the PVC (Verrucomicrobia, Planctomycetes, Omnitrophica) and Acidobacteria were relatively abundant, followed by FCB members (Bacteroidetes, Latescibacteria, Ignavibacteria, Gemmatimonadetes), Proteobacteria, Chloroflexi and Nitrospinae. Stramenopiles, Alveolata, Rhizaria, Fungi and sometimes Archaeplastida dominated eukaryotic communities. Baikal sediments harbored typically marine low-frequency eukaryotic and prokaryotic OTUs recently identified in some lakes (diplonemids, Bolidophyceae, SAR202, marine-like *Synechococcus*, Pelagibacterales) but also SAR324, Syndiniales and, surprisingly, Radiolaria, never reported in freshwater ecosystems. These OTUs likely sediment from the water column, contributing to the rare OTU pool. Baikal benthic communities displayed remarkable stability across sites and seemed not determined by depth or latitude. Comparative analyses with other freshwater, brackish and marine sediment prokaryotic communities confirmed the distinctness of Baikal benthic communities, which show some similarity to marine and hydrothermally-influenced systems likely owing to its high oligotrophy, depth and fault-associated seepage.

Keywords: Lake Baikal; benthos; 16S/18S rRNA metabarcoding; archaea; bacteria; protist; marine-freshwater transition

Introduction

Lake Baikal is the oldest, deepest, and largest (by volume) freshwater lake on Earth. As such, it represents a unique ecosystem akin, in several respects, to sea environments. Associated to the Baikal Rifting Zone in Southern Siberia, the lake formed ca. 30 Myr ago [1]. It is located at an elevation of 455 m above sea level, attains a depth of ca. 1 650 m (average depth, ca. 750 m) and has a volume of ca. ~23 000 km^3, accounting for 20% of the Earth's unfrozen freshwater. Its basin's catchment area occupies territories of Russia and Mongolia, with 53 % of the inflowing river water coming from Buryatia [2, 3]. Its strong wind regime and the fact that, despite recent climate change, its surface freezes during several months in winter lead to coastal downwelling and deep-water ventilation [4, 5]. Consequently, its water body remains cold (~4°C), oxygen-rich (dissolved oxygen levels often exceeding 10 $mg.L^{-1}$) and ultra-oligotrophic, especially at the bottom of the lake [4-7]. High pressure and low temperatures in bathyal areas facilitate the formation of solid phase methane, such that Lake Baikal is the only lake on Earth known to host methane hydrates [8, 9]. Geographically, the Academician Ridge and the Selenga river delta (major inflow river) divide the lake into three basins: Northern, Central, and Southern [10]. Lake Baikal is listed as UNESCO World Heritage Site for its unique geomorphology, biology and ecology (including as socio-ecosystem) [3].

Being an ancient ecosystem, Lake Baikal hosts a broad biodiversity with many endemic metazoan species (e.g. 1455 animal endemic species from 2595 described) [4, 11], some of which underwent adaptive radiations [12]. The endemic fauna and flora of the lake were thoroughly studied during the past century, as well as the diversity of microbial life, mainly from plankton, by classical observation and cultural approaches [13-15]. Molecular approaches to characterize planktonic microbial communities in the lake started to be applied at the onset of the 21st century [16] and have largely expanded in recent years with high-throughput sequencing. A variety of studies has focused on the diversity of pelagic bacteria [17], microbial eukaryotes through the water column [18] or across the lake surface waters [19], spring bloom-associated bacteria and eukaryotes [20-22], sub-ice bacteria and algal communities [23] or bacteria in deep waters influenced by oil-methane seeps [24]. Other studies concerned specific groups, such as diatoms [25] or dinoflagellates [26] or bacteria [27]. More recently, metagenomic analyses have targeted planktonic communities from sub-ice [28] and deep waters [29], virus-bacteria assemblages in coastal waters [30] or viruses from the pelagic zone [31].

Comparatively, benthic microbial communities remain surprisingly poorly known. Punctual studies have explored biofilms in littoral zones of Lake Baikal [32], specific bacterial lineages in intertidal zones [33] or archaea and bacteria in bottom sediments [34, 35], notably influenced by methane seeps and oil-bearing fluids [35, 36]. Sediment-associated eukaryotes have only been sporadically studied [18]. Sediment ecosystems remain under-explored not only in Baikal but,

generally, in lacustrine environments. Yet, benthic communities are usually more complex and diverse than plankton [37], being crucial for organic matter remineralization and the completion of the carbon cycle [38-40]. Although they account for only a small portion of the total Earth's living biomass, sediment microorganisms might be, in terms of numbers, as abundant as in soil or plankton [41]. Being little studied, sediment-associated communities are an invaluable source of phylogenetic novelty, with several highly divergent archaeal and bacterial lineages discovered in recent years by molecular, including metagenomic, approaches, essentially in the oceanic realm [40, 42-49]. In addition, comparative studies of sediment ecosystems in oceans and continental systems are virtually lacking; attempts to compare sediment microbial communities along a salinity gradient are extremely rare [50].

In this work, we carry out a comparative study of benthic prokaryotic and eukaryotic microbial communities in Lake Baikal across a ~600 km latitudinal N-S gradient traversing the three lake basins and from surface (littoral sediment) to the greatest depths (>1 400 m), using a 16S/18S rRNA gene metabarcoding approach. Our results show complex and diverse microbial communities that, surprisingly for an extremely low-salt water body, include several typical marine prokaryotic and eukaryotic lineages. The comparison of communities associated to upper-layer sediment in Lake Baikal and other benthic ecosystems across different depth and salinity ranges set it apart from other freshwater and marine systems.

Materials & methods

Sample collection and DNA extraction

Samples were collected during a joint French-Russian research cruise carried out between June 28th and July 7th, 2017. Thirteen sediment push cores were retrieved along a North-South transect from depths ranging from 0.5 to 1 450 m. Four sediment samples were taken from the northern basin, five from the central basin and four from the southern basin (Fig.1; Supplementary Table 1). In each basin, we collected samples from its highest depths, littoral zones and close to inter-basin transition zones. The physicochemical parameters of lake waters close to the bottom were measured in situ with a CTD probe. For this study, we collected the sediment (ca. 0–1 cm) of the core surface, including the water interface. In BK04S (coarse sand deposits), we extracted interstitial water with a syringe prior to biomass and particle concentration. Despite its high transparency, since Lake Baikal locates in Southern Siberia, at relatively high latitude, light penetrates less than at the Equator. Hence, we defined three categories of samples according to depth: shallow (0-100 m, including the upper epibenthic zone), medium (100-800 m, including the lower epipelagic and most of the mesobenthic zone) and deep (> 800 m, including the deep mesobenthos and the bathyal zone) (Supplementary Table 1). Sediment samples from the chosen horizon were fixed in ethanol (>80% v/v) and stored at -20°C until processed.

DNA purification, 16S/18S rRNA gene amplification and sequencing

After ethanol elimination, ~2 g sediment samples were let rehydrate for 2-4 h at 4°C and DNA was extracted using the Power Soil™ DNA purification kit (Qiagen, Hilden, Germany). For each sample, 16S rRNA gene fragments (~290 bp) encompassing the V4-V5 region and 18S rRNA gene fragments (~530 bp) encompassing the V4 region were PCR-amplified using, respectively, primers U515F (5'-GTGCCAGCMGCCGCGGTAA-3')-U806R (5'-GGACTAVSGGGTATCTAAT-3') and EK-565F-NGS (5'-GCAGTTAAAAAGCTCGTAGT-3')-UNonMet (5'-TTTAAGTTTCAGCCTTGCG-3'), the latter biased against metazoans [51]. Primers were tagged with specific 10-bp molecular identifiers (MIDs) for multiplexed sequencing. 25-µl amplification reaction mixtures contained 0.5-3 µl of eluted DNA, 1.5 mM $MgCl_2$, 0.2 mM deoxynucleotide (dNTP) mix, 0.3 µM of each primer and 0.5 U Platinum Taq DNA Polymerase (Invitrogen, Carlsbad, CA). Five PCR reactions were carried out in parallel for each sample, and then pooled, to minimize PCR-associated biases. PCR reactions comprised 35 cycles (94°C for 30 s, 55-58°C for 30-45 s, 72°C for 90 s) preceded by 2 min denaturation at 94°C and followed by 5 min extension at 72°C. Pooled amplicons were purified using QIAquick PCR purification kit (Qiagen, Hilden,Germany). Amplicons were sequenced using paired-end (2x300 bp) Illumina MiSeq by Eurofins Genomics

(Ebersberg, Germany). Sequences have been deposited in GenBank under the BioProject number: XXXXXXXX.

Sequence and phylogenetic analyses

Raw sequences were treated using an in-house bioinformatic pipeline. Briefly, we merged paired-end reads according to strict criteria using flash [52] and attributed them to specific MID-identified samples. For each sample, we pruned MID and primer sequences using cutadapt [53], generating cleaned merged reads (CMRs). CMRs were next dereplicated to unique sequences ('clusters') using the vsearch tool [54]. Chimeric clusters were detected de novo using the vsearch tool and excluded from further analyses. All non-chimeric clusters were pooled together to define operational taxonomic units (OTUs) at 97% identity for 16S rRNA genes and 98% identity for 18S rRNA genes using cd-hit-est [55]. We phylogenetically assigned OTUs to taxa using vsearch pairwise comparisons with local rRNA databases build from SILVAv128 [56] and PR2v4.5 [57]. OTUs affiliating to chloroplast and mitochondria were removed. OTUs sharing <80% identity against their best environmental hit were blasted against the GenBank *nr* database (NCBI; https://ncbi.nlm.nih.gov/) and assigned manually by phylogenetic placement analyses. To this end, the closest hits to our sequences in SILVA and PR2 were aligned with full 16S/18S rDNA reference sequences covering the tree-of-life diversity [58] using mafft [59] (options *L-INS-I, --reorder, --adjustdirection* and *–preservecase).* Uninformative sites were removed from the alignment using trimAl [60] (*-automated1* option). The reference phylogenetic tree was built with IQtree [61] (options *-bb 1000*, *-nt AUTO*, *-m GTR-+G+I*). We then used mafft again to align our OTU sequences to the reference alignment (options *L-INS-I, --addfragments, --keeplength, --adjustdirection, --reorder, --preservecase*). OTU sequences were then placed into the reference phylogenetic tree using alignment files and the reference tree using EPA-ng [62] (with *--model GTR+FU+G4+IU*). Genesis [63] was used to transform the output JPLACE in NEWICK format. OTUs with no reliable affiliation were maintained as 'Unclassified'. We followed a validated taxonomy scheme when possible and the following large supergroups: FCB (Fibrobacteres–Chlorobi–Bacteroidetes), PVC (Planctomycetes–Verrucomicrobia–Chlamydia), DPANN (Diapherotrites–Parvarchaeota–Aenigmarchaeota–Nanoarchaeota–Nanohaloarchaeota, including also Micrarchaeota, Woesearchaeota and Pacearchaeota), TACK (Thaumarchaeota–Aigarchaeota–Crenarchaeota–Korarchaeota) and CPR (Candidate Phyla Radiation [64]). The assignment of putative marine taxa (31 OTUs) was validated by phylogenetic analyses as described.

Diversity indices and statistical analyses

We generated tables of prokaryotic and eukaryotic OTU abundance at different levels of resolution (Supplementary Tables 2-7) for diversity and statistical analyses. OTUs were considered rare when they represented <0.1% total CMRs. Rare OTUs were grouped as 'Other Bacteria' for diversity index calculations. Diversity plots and statistical analyses were carried out using R [65]. Stacked barplots were produced using ggplot2 [66] based on the raw data matrix with 'reshape2' [67]. Non-metric Multidimensional Scaling (NMDS) and UpSetR [68] analyses were conducted with OTU count tables normalized using the rarefaction principle. OTU rarefaction curves were plotted using *rarecurve* and raw data then rarefied using *rrarefy* (vegan R package [69]) (Supplementary Tables 2, 5 and 6). To avoid subsampling bias, the rarefaction process was done 500 times and the mean matrix was used for further analyses. This process led to a total 614 693 CMRs and 58 433 prokaryotic OTUs, and 50 603 CMRs and 2 887 eukaryotic OTUs. To visualize shared OTUs among samples, we generated advanced Venn diagrams with the *upset* function [68]. To include measured environmental factors in our statistical analyses, their values were centered and scaled. Alpha diversity and richness indices were computed using respectively *diversity* and *estimateR* functions from the vegan R package [69]. Evenness was computed as the Shannon diversity indices divided by the log of the number of OTUs per site (*specnumber* in 'vegan'). Normal distributions were tested by the Shapiro-Wilk tests through the *shapiro.test* function (stats R package). ANOVA and linear regression model analyses were performed using *aov* and *glm* functions (stats R package). For Beta-diversity analyses, NMDS analyses were conducted with the *metaMDS* function ('vegan') using Bray-Curtis dissimilarities of matrices from either i) a Wisconsin double standardization of the rarefied CMRs counts (*wisconsin* function in 'vegan') or ii) a phyla-based sum of the rarefied CMRs counts per sample without standardization process. PERMANOVA and ANOSIM tests were applied to the same matrices using the *adonis* and *anosim* functions (vegan R package [69]), respectively, to test the influence of metadata variables and their interactions on the Bray-Curtis dissimilarity distances between samples (Supplementary Table 8). For comparative analyses with other studies, we recovered the metadata (salinity, temperature, depth, environment, sequencing methodology, primers) and relative read abundance per taxon (Supplementary Tables 9-10) [70-76]. Taxonomy was adapted to match equivalent phylum-like level. Phyla with missing data in some studies were gathered in the category 'Other Bacterial Phyla' to reduce noise due to changes in nomenclature and sequencing methodology between the different studies. NMDS analyses were conducted on phylum-like taxon frequencies. The dendrogram was computed using Bray-Curtis dissimilarities (*vegdist,* vegan R package [69]) clustered using UPGMA.

Results and discussion

Overall structure of sediment microbial communities

To study microbial communities associated to Lake Baikal sediments, we collected samples from thirteen locations following a North-South latitudinal gradient extending over ~600 km and different depths, from littoral to the deepest, bathyal zones in the three Baikal basins (Fig.1; Supplementary Table 1). The Northern basin is delimited to the South by the Academician Ridge [10], while the Central and Southern basins are separated by the Selenga river delta, which represents an input of organic matter but also pollutants, notably polycyclic aromatic compounds [77] and mercury [78]. We purified DNA from the upper sediment layer (ca. 0-1 cm) and massively sequenced amplicons of 16S (V4-V5 region) and 18S (V4 region) rRNA genes. After excluding low-quality sequences, we obtained, respectively, 1 774 112 and 1 628 588 clean merged reads (CMRs) that clustered in 25 229 and 8 139 operational taxonomic units (OTUs; see Methods) for prokaryotes and eukaryotes. Rarefaction curves suggested that benthic prokaryotic and eukaryotic diversity was relatively well captured by these OTUs, except for prokaryotic shallow datasets (BK01S.1, BK01S.2, BK04S and BK11S) and eukaryotic BK03S and BK01S.1 datasets, which did not reach a plateau owing to the lower number of retrieved CMRs. Collectively, accumulation curves did not reach clear saturation, highlighting the high diversity of sediment-associated microbes (Supplementary Fig.1). To avoid sequencing depth biases, we rarefied CMR matrices (Supplementary Tables 2, 5 and 6) for comparative analyses and the computation of richness, diversity and evenness indices (Supplementary Table 1). Eukaryotic and, most especially, prokaryotic communities exhibited high richness and diversity scores. For instance, Shannon index values ranged from 5.5 to 7.1 (prokaryotes) and 2.4 to 5.0 (eukaryotes). In contrast with planktonic communities, which tend to display marked rank:abundance curves in both, marine [79] and freshwater systems [80], Baikal sediment communities displayed high evenness values, ranging from 0.7 to 0.85 (prokaryotes) and 0.6 to 0.85 (eukaryotes). Therefore, Baikal sediments harbor complex communities with no clear dominant species.

Prokaryotic communities

At the domain level, although Bacteria dominated over Archaea in terms of relative abundance (CMR counts) and diversity (OTU counts), archaea reached up to 25% abundance in some samples and represented on average 20% of the OTUs (Supplementary Fig.2). Archaea encompassed a wide diversity of phyla, as observed in recent studies of seepage areas [35], with members of the DPANN and TACK (Proteoarchaeota) being the most abundant, as is common in sediments (e.g. [81, 82], followed by members of the Euryarchaeota (Fig.1B). The latter were largely represented by Thermoplasmatales (Supplementary Fig.3), notably Thermoprofundales (Marine Benthic Group D)

(Supplementary Table 3). These archaea are cosmopolitan in ocean sediments, being likely mixotrophs [83], but they have also been detected in some freshwater lake sediments [84]. Methanogenic archaea were present, albeit not very abundant, at the studied surface sediment horizons except for, occasionally, members of Methanofastidosa (WS2A), which might carry out methanogenesis through methylated thiol reduction [85]. Thaumarchaeota were the dominant TACK members followed, in some samples, by Bathyarchaeota (Supplementary Fig.3). Thaumarchaeota are commonly found in soil and sediments, where they typically oxidize ammonia to nitrite [86, 87]. Finally, Woesearchaeota and Pacearchaeota were the most abundant DPANN members. These archaea have reduced genomes, being likely parasites of other archaea [88, 89]. Given their abundance, as well as that of free-living Thermoplasmatales and Thaumarchaeota, it is tempting to hypothesize that the latter might constitute their potential hosts. This, however, will need to be confirmed by direct observation and further comparative studies.

Bacteria comprised a wide variety of phyla seemingly involved in a complex process of organic matter degradation. Members of the PVC clade (especially Verrucomicrobia, Planctomycetes and Omnitrophica) and Acidobacteria were the most relatively abundant, followed by the FCB clade (notably Bacteroidetes, Latescibacteria, Ignavibacteria and Gemmatimonadetes), Proteobacteria, Chloroflexi and Nitrospinae (Fig.1; Supplementary Fig.4). Bacteroidetes, Verrucomicrobia and Planctomycetes are typically involved in biomass recycling. For instance, Bacteroidetes release carbohydrate-active enzymes (CAZymes), being associated with gut and soil microbiomes [90] but also with microbial mats linked to cyanobacterial primary production [91, 92]. The relatively low percentage of Proteobacteria was unusual when compared to other studies of shallow freshwater lake sediments [93, 94] and even the upper layers of deep-sea sediments [37, 95-97] where proteobacterial abundances often exceed 30%. The values that we found (ca. 15%) are closer to those reported for subseafloor core sediments [42]. This observation extends to the proteobacterial classes, since Baikal sediments were dominated by Delta- and Betaproteobacteria, while Gammaproteobacteria, which is the main group in other freshwater sediment samples [37, 97-99], were only the third most abundant proteobacterial class. Deltaproteobacteria are likely involved in sulfate-reduction and/or other hydrogen-based syntrophic interactions to achieve the mineralization of organics [100-102]. Many biomass-degrading lineages in Lake Baikal sediment showed signatures of deep-subsurface and/or hydrothermally-influenced sites. For instance, among the Planctomycetes, the lineage Phycisphaerae was particularly prevalent (Supplementary Table 3). They are usually found in suboxic sediments [103] and thermophilic members have been isolated from thermal springs around Baikal [104]. Also relatively abundant were the Ignavibacteria, grouping moderate thermophilic, non-photosynthetic relatives of Chlorobi that are facultatively anaerobic and obligate

organotrophs [105], and Latescibacteria, typically involved in hydrocarbon and nutrient cycling in deep-sea hydrothermal sediments [102]. Some thermophilic organisms have indeed been isolated from bottom Baikal sediments in association with gas seeps [106] and communities from its bottom sediments can transform organic matter under thermobaric conditions [34]. This suggests the contribution of thermophilic or seepage-associated microbes to Baikal benthic communities, possibly in association with faulting zones. Both, Nitrospirae, more abundant, and Nitrospinae were present in Baikal sediment and likely contribute to carbon fixation associated with nitrification in interaction with ammonia oxidizing archaea [107, 108]. Therefore, prokaryotic communities in Lake Baikal sediments attest for complex N, S and C cycling.

Lake Baikal sediment communities were not only highly diverse but might be a source of phylogenetic novelty. Thus, many bacterial and archaeal groups included sequences having less than 80% identity to sequences in public databases (Fig.2A). Although sequences affiliated to well-known phyla have average sequence identity higher than 90% to sequences in databases both, CPR and DPANN members displayed average sequence identity of only 87-88% to sequences in databases and many archaeal and bacterial sequences, which we left unclassified, had much lower similarities (Fig.2A). CPR members which, like DPANN, encompass most likely dependent parasites/symbionts [88, 109], were highly diverse in terms of rare (<0.1% CMRs) OTUs, suggesting that they might depend on diverse, not dominant, bacteria. The proportion of DPANN OTUs also increased in the rare OTU fraction (Supplementary Fig.5). Contrasting with the patterns of relative abundance and diversity of rare OTUs, which were rather similar, those for the abundant OTUs differed. Although the percentages of abundant OTUs for the different phyla varied among samples, the diversity of OTUs per phyla was strikingly similar among samples (Supplementary Fig.5A). As relatively abundant OTUs are more likely to correspond to active sediment microbes, this observation strongly suggests that the active component of Baikal benthonic communities is highly stable across latitudinal and depth gradients, although the proportion of the different OTUs varies among samples.

Benthic eukaryotic communities

Microbial eukaryotes are poorly studied in sediments, notably from freshwater systems, despite they are important members in trophic webs [110-112]. We detected a relatively wide diversity of protists in Lake Baikal sediments, with Stramenopiles and Alveolata being the dominant groups (Fig.1). Alveolates included mostly ciliates but also dinoflagellates and Syndiniales (Supplementary Table 4). Stramenopiles included both heterotrophic lineages, notably, labyrinthulids, thraustochytrids, oomycetes and MAST protists but also ochrophyte algae, comprising xanthophytes, chrysophytes and diatoms, which are abundant in the water column. The presence of diatoms and chrysophytes in

Baikal deep sediments has been documented and studied as input biomass for degradation [7, 34, 113]. Other typical photosynthetic eukaryotes, including plant sequences (Archaeplastida) were detected in more or less disparate abundance in shallow but also deep sediments (Fig.1). This highlights the difficulty of discriminating local active eukaryotic communities from external input coming from upper water column levels or continental debris. We also detected a considerable number of OTUs affiliated to Opisthokonta, mostly fungi, frequent in seafloor communities [40, 114], and Rhizaria, notably cercozoans. Detected members of Hacrobia comprised centroheliozoans, cryptophytes, haptophytes and telonemids (Supplementary Table 4).

As compared to prokaryotes, benthic protists may be rare in several lake areas, as obtaining amplification products was difficult for some sediment samples. This was the case of BK22S, taken in a seeping zone (bubbles were visible in the core), even when replicate samples were used, suggesting that protists in this core were not abundant (and they were not diverse). The abundance and diversity patterns of rare and abundant OTUs across phyla globally resembled each other (Supplementary Fig.6). Only plant sequences, that appeared to accumulate in some sediment samples, were relatively more abundant, although not highly diverse. As for prokaryotes, most eukaryotic OTUs shared >90% 18S rRNA gene identity with sequences in databases (Fig.2B). However, members affiliated to Apusomonads and Ancyromonads, Excavata and Hacrobia were more divergent (~85% shared identity) and a significant number of (unclassified) eukaryotic sequences were really divergent (~75% identity with sequences in databases). This suggests that novel protist lineages (unknown from the water column and well-studied environments) likely thrive in these benthic communities.

Marine signature taxa

During the manual revision of OTU phylogenetic assignment (see Methods), we identified several OTUs belonging to typical marine taxa (Fig.3). Salinity is a major driver of microbial community composition [115] and marine-freshwater transitions are deemed to be rare [116]. Indeed, the adaptation to even moderate salt concentrations (e.g. seawater, ~3.5%) elicits wide proteome changes, for instance increased average protein acidity, which translates in lower isoelectric point (pI) [117]. Nonetheless, such transitions are known and the discovery in freshwater systems of prokaryotic and eukaryotic lineages previously thought to dwell exclusively in marine ecosystems is increasing. Among eukaryotes, they include members of the perkinsids [118], haptophytes [119], Bolidophyceas [19, 120] and several Marine Stramenopiles (MAST) clades, such as MAST-2, MAST-12, MAST-3 and possibly MAST-6 [121]. Recently, diplonemids, a cosmopolitan group of oceanic excavates particularly abundant and diverse in the deep ocean [122, 123] have been identified in deep freshwater lakes in Japan [116]. Typical marine prokaryotes have also been recently detected in

freshwater systems, including marine-like *Synechococcus* strains [124], the Chloroflexi lineage SAR202 [125], typically thriving in the dark ocean and involved in sulfur cycling [126], and even members of Pelagibacterales (SAR11), such as the strain LD12, which has the smallest genome for a free-living bacterium (1.16 Mbp [127]).

Our study revealed 285 OTUs belonging to typical marine clades (Supplementary Table 7). Several of them belonged to groups already observed in freshwater systems. In addition to the relatively widespread in freshwater systems MAST clades, which were relatively abundant in Baikal, we also identified the rarely encountered (in freshwater systems) diplonemids and Bolidophyceae and, among bacteria, marine-like *Synechococcus*, SAR202 Chloroflexi and Pelagibacterales (Fig.3). SAR202 and Pelagibacterales have indeed been recently observed in Lake Baikal [128]. However, we also identified lineages never before reported in freshwater ecosystems. These comprised, among bacteria, members of SAR324, a clade of Deltaproteobacteria typically associated with submarine hydrothermal plumes that are able to use dissolved organic sulfur, having flexible metabolism [129, 130]. Among eukaryotes, we identified diverse members of the Syndiniales, which are frequent parasites of marine dinoflagellates [131] and, more surprisingly, Radiolaria. Radiolaria usually have conspicuous silica or strontium sulfate-based skeletons that are easily identifiable [132]. We detected 30 radiolarian OTUs in Baikal sediments, most of which clearly branched among classical members of the Acantharea and Polycystinea in reference phylogenetic trees (Supplementary Fig.7). Despite strict controls made it highly unlikely, to further eliminate the possibility that the observation of these typically marine lineages could derive from some type of hidden cross-contamination of our samples at different steps (collection, handling in the laboratory or sequencing process), we mined for other typical and abundant marine taxa in our datasets. We failed to detect any sequence of the marine picoalgal Prasinophyta and the bacterial genera *Prochlorococcus* and *Alteromonas* (Fig.3). This control reinforces the conclusion that these Baikal typically 'marine' OTUs are indeed autochthonous.

Baikal 'marine' OTUs were not restricted to any specific sampling location (SAR324 was present in all samples). They were moderately diverse, with up to 88 OTUs (SAR202; SAR324 had 52 OTUs and MAST, 54 OTUs). MAST sequences were the most abundant, followed by Radiolaria and Syndiniales (Fig.3). Each Baikal sediment sample harbored between 20 and 80 OTUs attributed to 'marine' clades. The highest 'marine' CMR abundances were found in BK16S (Northern basin; 846m deep), BK04S (Central basin; 0.5m deep) and BK26S (Southern basin; 1 412m deep). The fact that each of these three sampling points was located on a different basin of the lake and having different depth highlights the generalized presence of typical marine taxa in Lake Baikal albeit always with very low frequency. Some of these OTUs might be potentially thriving in sediments, e.g. members of the SAR202 and SAR324 clades (although they are typically planktonic in oceans) eventually involved in

sulfur cycling in association with hydrothermal seepage. However, most likely the detected 'marine' OTUs correspond to true planktoners or surface benthic dwellers that, upon accumulation in sediments, are more easily detected by amplicon sequencing. If this is the case, their presence in the water column may be very low and, in some cases, potentially below the detection threshold both, for classical observations (e.g. radiolarians) or full metagenomic approaches (owing to the difficulty of assembling genomes from rare organisms). Recent 18S rRNA gene metabarcoding studies of surface (1-50 m) plankton in Lake Baikal suggest that some protists having close relatives with marine/brackish species might be glacial relicts [19]. At any rate, our study confirms and extends the presence of several typically marine prokaryotes and eukaryotes at low abundances in Lake Baikal, reinforcing the idea that transition frequency between marine and freshwater habitats is underestimated [133]. This also poses the question of the specific molecular adaptations to very low salinity, as Lake Baikal salinity is extremely low (0.0 PSU) and suggests that oligotrophy and deep waters might be more important drivers than salinity for these lineages.

Benthic communities across latitudinal and vertical gradients

Once we characterized microbial diversity in Lake Baikal sediment samples, we aimed at ascertaining whether the depth or the basin where samples were collected determined benthic microbial community structure. To avoid any bias linked to sequencing depth, we rarefied CMRs to the same amount for all locations (see Methods). An NMDS analysis based on dissimilarity matrices of OTU frequencies showed no obvious pattern discriminating samples according to basin or depth (Fig.4). Although two samples of intermediate depth (100-800 m) appeared to segregate on along axis 1, surface (<100 m) and deep (>800 m) samples appeared mixed. This pattern was almost superimposable to that observed for prokaryotic OTUs, whereas eukaryotic OTUs seemed to segregate better surface from deep samples (Supplementary Fig.8). PERMANOVA analyses confirmed no significant discrimination of prokaryotic and eukaryotic communities at the OTU level by latitude and only marginal significance for depth (prokaryotes, p=0.01; eukaryotes, 0.04) (Supplementary Table 8). Since this marginal effect of depth in determining prokaryotic and eukaryotic communities might be due to the large collective dominance of rare OTUs, we also carried out NMDS and PERMANOVA analyses on dissimilarity matrices at high-rank taxon, rather than OTU, level. This analysis reinforced previous results. Prokaryotic and eukaryotic communities did not appear segregated by depth (Supplementary Fig.9) and correlation between phyla and depth categories were not significant for eukaryotes and only very marginally for prokaryotes (p=0.031; Supplementary Table 8).

Contrasting with the high sample similarity in terms of OTU distribution among phyla (Fig1; Supplementary Fig.5-6), a large percentage of individual OTUs within each sample was unique, especially for eukaryotes (~37% prokaryotic vs. ~80% eukaryotic OTUs) (Fig.5, upper right inset). However, most of these were rare OTUs possibly encompassing both low-frequency benthic active members but also dormant, dispersing and/or decaying forms from the water column and upper sediment layers. Nonetheless, we detected a core of 75 prokaryotic OTUs that were shared by all the sediment samples (Fig.5). Interestingly, the relative abundance of benthic shared OTUs represented between ~25% and ~50% of the total prokaryotic abundance (Fig.5, left-bottom panel). In addition, the phylogenetic affiliation of these OTUs matches well the general phyla distribution observed for prokaryotic communities (Fig.1; Supplementary Fig.5). This strongly suggests that there is a stable core of active benthic prokaryotic communities across basins and depths in Lake Baikal. This core is accompanied by a wide diversity of rare OTUs, some of which are likely inactive and others distribute among the major dominant phyla in Baikal benthos.

Comparative analysis of benthic communities across deep water bodies

How do benthic Baikal communities compare to those of other aquatic ecosystems? Does the presence of Baikal marine taxa indicate intermediate ecological features between freshwater and marine environments? To address these questions, we retrieved 16S rRNA gene metabarcoding data from a selection of sediment samples across a variety of freshwater, brackish and marine ecosystems whose high-rank taxa patterns were available in the literature (Supplementary Tables 9-10). An NMDS plot based on the high-rank taxa dissimilarity matrix showed that Lake Baikal sediment samples clustered away from the other samples, although it was closer to freshwater samples (Fig.6A). To quantify the effect of potential factors accounting for the differences between samples, we performed PERMANOVA analyses on the same dissimilarity matrix. They revealed that the effect size associated with the salinity category was greater (R^2 = 0.53; p-value = 10^{-4}) than that attributed to the amplification and sequencing method (R^2 = 0.42; p-value = 10^{-4}) (Supplementary Table 8). To put these effect sizes in perspective, the effect size of the sampling location (one per study), which is believed to be the most influential variable, was R^2 = 0.72; p-value = 10^{-4}. Acidobacteria, PVC, Nitrospirae and Chloroflexi seemed to drive the segregation of Baikal samples, together with a lower proportion of most proteobacterial classes (except Betaproteobacteria). For a more detailed look, we computed a dendrogram from the dissimilarity matrix (Fig.6B). Lakes Baikal and Erhai clustered together in a freshwater clade characterized mainly by high abundances of PVC and Chloroflexi along with smaller abundances of, especially, Alphaproteobacteria and Gammaproteobacteria. Marine sediment samples clustered together united by a high abundance of proteobacterial classes,

especially Gamma- and Deltaproteobacteria, and the lower prevalence of Nitrospirae. In summary, independently of methodological biases associated to different studies, Lake Baikal harbors distinctive benthic communities that differ from the compared marine, brackish and freshwater communities.

Concluding remarks

Lake Baikal is a unique water body with very low salinity but also high depth and volume, its oligotrophic waters being influenced by hydrothermal seepage. Thus, except for the salinity, its features justify the local denomination of 'Baikal Sea'. Accordingly, Lake Baikal sediments harbor idiosyncratic prokaryotic and eukaryotic benthic communities that differ from those in other freshwater, brackish and marine ecosystems (Fig.6). Baikal sediment communities are extremely diverse, encompassing a wide variety of archaeal, bacterial and eukaryotic taxa (Fig.1). Rare OTUs including plankton dormant or decaying forms and low-frequency members compose a significant portion of Baikal benthic communities. However, there is a significant proportion of likely active communities that share a conserved OTU core (Fig.5) and are most likely active in C, N and S cycles. Furthermore, Lake Baikal benthic microbial communities are remarkably stable across the latitudinal N-S transect and depth gradient (0.5-1 450 m), with no clear differentiation of samples according to basin or depth (Fig.4). This stability is most striking at the level of prokaryotic abundant OTUs (>0.1% CMR; Supplementary Fig.5), which most likely correspond to active benthic members. Beyond light penetration (limited for sediments) and pressure, this possibly reflects the relatively stable temperature of the lake at almost all depths (~ 4°C). The lake benthic communities partially reflect the adaptation to methane seepage and hydrothermal influence, with the presence of members typical of subseafloor sediments and hydrothermal-influenced communities and lower proteobacterial abundances (e.g. [42, 102, 105]). Along with this resemblance, we unequivocally identify several OTUs belonging to typical marine taxa, including some lineages never before detected in freshwater systems, such as the bacterial SAR324 clade, and the eukaryotic Syndiniales and Radiolaria. Members of these 'marine' taxa are present in low frequencies and are likely planktonic, but their accumulation in sediments facilitates their detection by metabarcoding approaches. This indicates that marine-freshwater transitions are more frequent than thought and that oligotrophy, low temperature and/or deep-water darkness might be more important drivers than salinity for the environmental adaptation of some lineages.

Acknowledgments We thank Luis J. Galindo and Anabel Lopez-Archilla for help and discussions during our 2017 limnological cruise, and Philippe Deschamps for technical bioinformatic support. We thank the crew of the R/V G. Titov for their professionalism and efficiency onboard and the director of

the Limnological Institute at Irkusk for logistical assistance. This research was funded by the European Research Council Grants ProtistWorld (322669, PL-G) and PlastEvol (787904, DM) as well as the Russian State grant 0345-2016-0009 (NVA).

Author contributions PLG, DM and NVA designed the work and organized the limnological cruise. PLG, GR, NVA and KEV collected sediment samples. PB and GR purified DNA and carried out PCR reactions for metabarcoding. GR carried out the bioinformatic analysis of amplicon sequences, statistical analyses and wrote an early draft of the manuscript. PLG wrote the final manuscript. All authors read, critically commented and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

[1] Müller J, Oberhänsli H, Melles M, Schwab M, Rachold V, Hubberten HW. Late Pliocene sedimentation in Lake Baikal: implications for climatic and tectonic change in SE Siberia. Palaeogeogr Palaeoclimatol Palaeoecol. 2001; 174**:** 305-326.

[2] Sherstyankin PP, Alekseev SP, Abramov AM, Stavrov KG, De Batist M, Hus R *et al*. Computer-based bathymetric map of Lake Baikal. Dokl Earth Sci. 2006; 408**:** 564-569.

[3] UNDP-GEF. The ecological atlas of the Baikal basin. United Nations Office for Project Sercives (UNOPS): http://baikal.iwlearn.org/en. 2015. p 145.

[4] Moore MV, Hampton SE, Izmest'eva LR, Silow EA, Peshkova EV, Pavlov BK. Climate change and the world's "Sacred Sea"—Lake Baikal, Siberia. BioScience. 2009; 59**:** 405-417.

[5] Schmid M, Budnev NM, Granin NG, Sturm M, Schurter M, Wüest A. Lake Baikal deepwater renewal mystery solved. Geophys Res Lett. 2008; 35**:** L09605-L09605.

[6] Troitskaya E, Blinov V, Ivanov V, Zhdanov A, Gnatovsky R, Sutyrina E *et al*. Cyclonic circulation and upwelling in Lake Baikal. Aquatic Sciences. 2015; 77**:** 171-182.

[7] Shimaraev MN, Domysheva VM. Trends in hydrological and hydrochemical processes in Lake Baikal under conditions of modern climate change. John Wiley & Sons, Ltd: Chichester, UK. 2012. p^pp 43-66.

[8] De Batist M, Klerkx J, Van Rensbergen P, Vanneste M, Poort J, Golmshtok AY *et al*. Active hydrate destabilization in Lake Baikal, Siberia? Terra Nova. 2002; 14**:** 436-442.

[9] Granin NG, Aslamov IA, Kozlov VV, Makarov MM, Kirillin G, McGinnis DF *et al*. Methane hydrate emergence from Lake Baikal: direct observations, modelling, and hydrate footprints in seasonal ice cover. Scientific reports. 2019; 9**:** 19361.

[10] Mats VD, Perepelova TI. A new perspective on evolution of the Baikal Rift. Geoscience Frontiers. 2011; 2**:** 349-365.

[11] Yu Sherbakov D. Molecular phylogenetic studies on the origin of biodiversity in Lake Baikal. Trends Ecol Evol. 1999; 14**:** 92-95.

[12] Stelbrink B, Shirokaya AA, Clewing C, Sitnikova TY, Prozorova LA, Albrecht C. Conquest of the deep, old and cold: an exceptional limpet radiation in Lake Baikal. Biology letters. 2015; 11.

[13] Maksimova EA, Maksimov VN. [Vertical distribution of microbial plankton in the Southern part of Lake Baikal in 1969]. Mikrobiologiia. 1972; 41**:** 896-902.

[14] Maksimov VV, Shchetinina EV, Kraikivskaia OV, Maksimov VN, Maksimova EA. [The classification and the monitoring of the state of mouth riverine and lacustrine ecosystems in lake Baikal based on the composition of local microbiocenoses and their activity]. Mikrobiologiia. 2002; 71: 690-696.
[15] Bel'kova NL, Parfenova VV, Kostopnova T, Denisova L, Zaichikov EF. [Microbial biodiversity in the Lake Baikal water]. Mikrobiologiia. 2003; 72: 239-249.
[16] Glöckner FO, Zaichikov E, Belkova N, Denissova L, Pernthaler J, Pernthaler A *et al*. Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. Applied and environmental microbiology. 2000; 66: 5053-5065.
[17] Kurilkina MI, Zakharova YR, Galachyants YP, Petrova DP, Bukin YS, Domysheva VM *et al*. Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. FEMS Microbiol Ecol. 2016; 92.
[18] Yi Z, Berney C, Hartikainen H, Mahamdallie S, Gardner M, Boenigk J *et al*. High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations. FEMS Microbiol Ecol. 2017; 93: 10.
[19] Annenkova NV, Giner CR, Logares R. Tracing the origin of planktonic protists in an ancient lake. Microorganisms. 2020; 8.
[20] Mikhailov IS, Zakharova YR, Galachyants YP, Usoltseva MV, Petrova DP, Sakirko MV *et al*. Similarity of structure of taxonomic bacterial communities in the photic layer of Lake Baikal's three basins differing in spring phytoplankton composition and abundance. Doklady Biochemistry and biophysics. 2015; 465: 413-419.
[21] Mikhailov IS, Bukin YS, Zakharova YR, Usoltseva MV, Galachyants YP, Sakirko MV *et al*. Co-occurrence patterns between phytoplankton and bacterioplankton across the pelagic zone of Lake Baikal during spring. Journal of microbiology (Seoul, Korea). 2019; 57: 252-262.
[22] Mikhailov IS, Zakharova YR, Bukin YS, Galachyants YP, Petrova DP, Sakirko MV *et al*. Co-occurrence networks among bacteria and microbial eukaryotes of Lake Baikal during a spring phytoplankton bloom. Microbial ecology. 2019; 77: 96-109.
[23] Bashenkhaeva MV, Zakharova YR, Petrova DP, Khanaev IV, Galachyants YP, Likhoshway YV. Sub-ice microalgal and bacterial communities in freshwater Lake Baikal, Russia. Microbial ecology. 2015; 70: 751-765.
[24] Zakharenko AS, Galachyants YP, Morozov IV, Shubenkova OV, Morozov AA, Ivanov VG *et al*. Bacterial communities in areas of oil and methane seeps in pelagic of Lake Baikal. Microbial ecology. 2019; 78: 269-285.
[25] Zakharova YR, Galachyants YP, Kurilkina MI, Likhoshvay AV, Petrova DP, Shishlyannikov SM *et al*. The structure of microbial community and degradation of diatoms in the deep near-bottom layer of Lake Baikal. PloS one. 2013; 8: e59977.
[26] Annenkova NV, Lavrov DV, Belikov SI. Dinoflagellates associated with freshwater sponges from the ancient lake baikal. Protist. 2011; 162: 222-236.
[27] Belikov S, Belkova N, Butina T, Chernogor L, Martynova-Van Kley A, Nalian A *et al*. Diversity and shifts of the bacterial community associated with Baikal sponge mass mortalities. PloS one. 2019; 14: e0213926.
[28] Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov VV *et al*. Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. Applied and environmental microbiology. 2018; 84.
[29] Cabello-Yeves PJ, Zemskaya TI, Zakharenko AS, Sakirko MV, Ivanov VG, Ghai R *et al*. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. Limnol Oceanogr. 2020; n/a.
[30] Butina TV, Bukin YS, Krasnopeev AS, Belykh OI, Tupikin AE, Kabilov MR *et al*. Estimate of the diversity of viral and bacterial assemblage in the coastal water of Lake Baikal. FEMS microbiology letters. 2019; 366.

[31] Potapov SA, Tikhonova IV, Krasnopeev AY, Kabilov MR, Tupikin AE, Chebunina NS *et al*. Metagenomic analysis of virioplankton from the pelagic zone of Lake Baikal. Viruses. 2019; 11.
[32] Sorokovikova EG, Belykh OI, Gladkikh AS, Kotsar OV, Tikhonova IV, Timoshkin OA *et al*. Diversity of cyanobacterial species and phylotypes in biofilms from the littoral zone of Lake Baikal. Journal of microbiology (Seoul, Korea). 2013; 51: 757-765.
[33] Lenk S, Arnds J, Zerjatke K, Musat N, Amann R, Mußmann M. Novel groups of Gammaproteobacteria catalyse sulfur oxidation and carbon fixation in a coastal, intertidal sediment. Environmental microbiology. 2011; 13: 758-774.
[34] Bukin SV, Pavlova ON, Manakov AY, Kostyreva EA, Chernitsyna SM, Mamaeva EV *et al*. The ability of microbial community of Lake Baikal bottom sediments associated with gas discharge to carry out the transformation of organic matter under thermobaric conditions. Frontiers in microbiology. 2016; 7: 690.
[35] Lomakina AV, Mamaeva EV, Galachyants YP, Petrova DP, Pogodaeva TV, Shubenkova OV *et al*. Diversity of archaea in bottom sediments of the discharge areas with oil- and gas-bearing fluids in Lake Baikal. Geomicrobiol J. 2018; 35: 50-63.
[36] Kadnikov VV, Mardanov AV, Beletsky AV, Shubenkova OV, Pogodaeva TV, Zemskaya TI *et al*. Microbial community structure in methane hydrate-bearing sediments of freshwater Lake Baikal. FEMS Microbiol Ecol. 2012; 79: 348-358.
[37] Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM *et al*. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PloS one. 2011; 6: e24570-e24570.
[38] Dang H, Lovell CR. Microbial surface colonization and biofilm development in marine environments. Microbiol Mol Biol Rev. 2015; 80: 91-138.
[39] Rastelli E, Corinaldesi C, Dell'Anno A, Amaro T, Greco S, Lo Martire M *et al*. CO2 leakage from carbon dioxide capture and storage (CCS) systems affects organic matter cycling in surface marine sediments. Marine environmental research. 2016; 122: 158-168.
[40] Orsi WD. Ecology and evolution of seafloor and subseafloor microbial communities. Nature reviews Microbiology. 2018; 16: 671-683.
[41] Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. Global distribution of microbial abundance and biomass in subseafloor sediment. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109: 16213.
[42] Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH. Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105: 10583-11058 10588.
[43] Schauer R, Røy H, Augustin N, Gennerich HH, Peters M, Wenzhoefer F *et al*. Bacterial sulfur cycling shapes microbial communities in surface sediments of an ultramafic hydrothermal vent field. Environmental microbiology. 2011; 13: 2633-2648.
[44] Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF *et al*. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013; 499: 431-437.
[45] Baker BJ, Lazar CS, Teske AP, Dick GJ. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. Microbiome. 2015; 3: 14-14.
[46] Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE *et al*. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015; 521: 173–179.
[47] Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Op Microbiol. 2016; 31: 217-226.
[48] Seitz KW, Dombrowski N, Eme L, Spang A, Lombard J, Sieber JR *et al*. Asgard archaea capable of anaerobic hydrocarbon cycling. Nature communications. 2019; 10: 1822-1822.
[49] Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nature biotechnology. 2018; 36: 190-195.

[50] Wang Y, Sheng HF, He Y, Wu JY, Jiang YX, Tam NFY *et al*. Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. Applied and environmental microbiology. 2012; 78: 8264-8271.
[51] Bower SM, Carnegie RB, Goh B, Jones SRM, Lowe GJ, Mak MWS. Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. J Euk Microbiol. 2004; 51: 325-332.
[52] Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics (Oxford, England). 2011; 27: 2957-2963.
[53] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJournal. 2011; 17: 10-12.
[54] Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016; 4: e2584.
[55] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England). 2012; 28: 3150-3152.
[56] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013; 41: D590-596.
[57] Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L *et al*. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 2013; 41: D597-D604.
[58] Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ *et al*. A new view of the tree of life. Nat Microbiol. 2016; 1: 16048.
[59] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013; 30: 772-780.
[60] Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England). 2009; 25: 1972-1973.
[61] Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015; 32: 268-274.
[62] Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T *et al*. EPA-ng: Massively parallel evolutionary placement of genetic sequences. Systematic biology. 2019; 68: 365-369.
[63] Czech L, Barbera P, Stamatakis A. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics (Oxford, England). 2020.
[64] Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A *et al*. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015; 523: 208-211.
[65] R Development Core Team. R: A language and environment for statistical computing. In: http://www.r-project.org (ed), http://www.r-project.org edn. R Foundation for Statistical Computing: Vienna, Austria. 2017.
[66] Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. 2016.
[67] Wickham H. Reshaping data with the reshape package. Journal of Statistical Software. 2007; 21: 1-20.
[68] Conway JR, Lex A, Gehlenborg N. UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics (Oxford, England). 2017; 33: 2938-2940.
[69] Oksanen J, Blanchet G, Kindt R, Legendre P, O'Hara RB, Simpson GL *et al*. Vegan: Community Ecology Package. R package version 1.17-9. In: http://CRAN.R-project.org/package=vegan (ed). http://CRAN.R-project.org/package=vegan. 2011.
[70] Liu J, Liu X, Wang M, Qiao Y, Zheng Y, Zhang XH. Bacterial and archaeal communities in sediments of the North Chinese Marginal Seas. Microbial ecology. 2015; 70: 105-117.

[71] Mahmoudi N, Robeson MS, Castro HF, Fortney JL, Techtmann SM, Joyner DC *et al*. Microbial community composition and diversity in Caspian Sea sediments. FEMS Microbiology Ecology. 2015; 91: 1-11.
[72] Chen Y, Dai Y, Wang Y, Wu Z, Xie S, Liu Y. Distribution of bacterial communities across plateau freshwater lake and upslope soils. J Environ Sci (China). 2016; 43: 61-69.
[73] Gugliandolo C, Michaud L, Lo Giudice A, Lentini V, Rochera C, Camacho A *et al*. Prokaryotic community in lacustrine sediments of Byers Peninsula (Livingston Island, Maritime Antarctica). Microbial ecology. 2016; 71: 387-400.
[74] Ye Q, Wu Y, Zhu Z, Wang X, Li Z, Zhang J. Bacterial diversity in the surface sediments of the hypoxic zone near the Changjiang Estuary and in the East China Sea. MicrobiologyOpen. 2016; 5: 323-339.
[75] Wan Y, Ruan X, Zhang Y, Li R. Illumina sequencing-based analysis of sediment bacteria community in different trophic status freshwater lakes. MicrobiologyOpen. 2017; 6.
[76] Zeng YX, Yu Y, Li HR, Luo W. Prokaryotic community composition in arctic kongsfjorden and sub-arctic northern bering sea sediments as revealed by 454 pyrosequencing. Frontiers in microbiology. 2017; 8: 2498-2498.
[77] Adams JK, Martins CC, Rose NL, Shchetnikov AA, Mackay AW. Lake sediment records of persistent organic pollutants and polycyclic aromatic hydrocarbons in southern Siberia mirror the changing fortunes of the Russian economy over the past 70 years. Environmental pollution (Barking, Essex : 1987). 2018; 242: 528-538.
[78] Roberts S, Adams JK, Mackay AW, Swann GEA, McGowan S, Rose NL *et al*. Mercury loading within the Selenga River basin and Lake Baikal, Siberia. Environmental pollution (Barking, Essex : 1987). 2020; 259: 113814.
[79] Pedros-Alio C. Marine microbial diversity: can it be determined? Trends in microbiology. 2006; 14: 257-263.
[80] Simon M, Lopez-Garcia P, Deschamps P, Moreira D, Restoux G, Bertolino P *et al*. Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. The ISME journal. 2015; 9: 1941-1953.
[81] Stieglmeier M, Mooshammer M, Kitzler B, Wanek W, Zechmeister-Boltenstern S, Richter A *et al*. Aerobic nitrous oxide production through N-nitrosating hybrid formation in ammonia-oxidizing archaea. The ISME journal. 2014; 8: 1135-1146.
[82] Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD *et al*. Predominant archaea in marine sediments degrade detrital proteins. Nature. 2013; 496: 215-218.
[83] Zhou Z, Liu Y, Lloyd KG, Pan J, Yang Y, Gu J-D *et al*. Genomic and transcriptomic insights into the ecology and metabolism of benthic archaeal cosmopolitan, Thermoprofundales (MBG-D archaea). The ISME journal. 2019; 13: 885-901.
[84] Borrel G, Lehours AC, Crouzet O, Jézéquel D, Rockne K, Kulczak A *et al*. Stratification of Archaea in the deep sediments of a freshwater meromictic lake: vertical shift from methanogenic to uncultured archaeal lineages. PloS one. 2012; 7: e43346.
[85] Nobu MK, Narihiro T, Kuroda K, Mei R, Liu WT. Chasing the elusive Euryarchaeota class WSA2: genomes reveal a uniquely fastidious methyl-reducing methanogen. The ISME journal. 2016; 10: 2478-2487.
[86] Offre P, Spang A, Schleper C. Archaea in biogeochemical cycles. Annual review of microbiology. 2013; 67: 437-457.
[87] Oton EV, Quince C, Nicol GW, Prosser JI, Gubry-Rangin C. Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. The ISME journal. 2016; 10: 85-96.
[88] Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nature reviews Microbiology. 2018; 16: 629-645.
[89] Dombrowski N, Lee JH, Williams TA, Offre P, Spang A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. FEMS microbiology letters. 2019; 366: slo.

[90] Larsbrink J, McKee LS. Bacteroidetes bacteria in the soil: Glycan acquisition, enzyme secretion, and gliding motility. Advances in applied microbiology. 2020; 110: 63-98.
[91] Saghaï A, Zivanovic Y, Moreira D, Benzerara K, Bertolino P, Ragon M *et al*. Comparative metagenomics unveils functions and genome features of microbialite-associated communities along a depth gradient. Environmental microbiology. 2016; 18: gut4990-5004.
[92] Gutierrez-Preciado A, Saghai A, Moreira D, Zivanovic Y, Deschamps P, Lopez-Garcia P. Functional shifts in microbial mats recapitulate early Earth metabolic transitions. Nature ecology & evolution. 2018; 2: 1700-1708.
[93] Bai Y, Shi Q, Wen D, Li Z, Jefferson WA, Feng C *et al*. Bacterial communities in the sediments of Dianchi lake, a partitioned eutrophic waterbody in China. PloS one. 2012; 7: e37796-e37796.
[94] Zhang J, Yang Y, Zhao L, Li Y, Xie S, Liu Y. Distribution of sediment bacterial and archaeal communities in plateau freshwater lakes. Applied microbiology and biotechnology. 2015; 99: 3291-3302.
[95] Bienhold C, Zinger L, Boetius A, Ramette A. Diversity and biogeography of bathyal and abyssal seafloor bacteria. PloS one. 2016; 11: e0148016-e0148016.
[96] Durbin AM, Teske A. Microbial diversity and stratification of South Pacific abyssal marine sediments. Environmental microbiology. 2011.
[97] Walsh EA, Kirkpatrick JB, Rutherford SD, Smith DC, Sogin M, D'Hondt S. Bacterial diversity and community composition from seasurface to subseafloor. ISME Journal. 2016; 10: 979-989.
[98] Dyksma S, Bischof K, Fuchs BM, Hoffmann K, Meier D, Meyerdierks A *et al*. Ubiquitous Gammaproteobacteria dominate dark carbon fixation in coastal sediments. ISME Journal. 2016; 10: 1939-1953.
[99] Hoffmann K, Bienhold C, Buttigieg PL, Knittel K, Laso-Pérez R, Rapp JZ *et al*. Diversity and metabolism of Woeseiales bacteria, global members of marine sediment communities. ISME Journal. 2020: 1-15.
[100] Muyzer G, Stams AJ. The ecology and biotechnology of sulphate-reducing bacteria. Nature reviews Microbiology. 2008; 6: 441-454.
[101] Lovley DR. Syntrophy Goes Electric: Direct Interspecies Electron Transfer. Annual review of microbiology. 2017; 71: 643-664.
[102] Dombrowski N, Seitz KW, Teske AP, Baker BJ. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. Microbiome. 2017; 5: 106.
[103] Spring S, Bunk B, Sproer C, Rohde M, Klenk HP. Genome biology of a novel lineage of planctomycetes widespread in anoxic aquatic environments. Environmental microbiology. 2018; 20: 2438-2455.
[104] Kovaleva OL, Merkel AY, Novikov AA, Baslerov RV, Toshchakov SV, Bonch-Osmolovskaya EA. Tepidisphaera mucosa gen. nov., sp. nov., a moderately thermophilic member of the class Phycisphaerae in the phylum Planctomycetes, and proposal of a new family, Tepidisphaeraceae fam. nov., and a new order, Tepidisphaerales ord. nov. International journal of systematic and evolutionary microbiology. 2015; 65: 549-555.
[105] Podosokorskaya OA, Kadnikov VV, Gavrilov SN, Mardanov AV, Merkel AY, Karnachuk OV *et al*. Characterization of Melioribacter roseus gen. nov., sp. nov., a novel facultatively anaerobic thermophilic cellulolytic bacterium from the class Ignavibacteria, and a proposal of a novel bacterial phylum Ignavibacteriae. Environmental microbiology. 2013; 15: 1759-1771.
[106] Baturina OA, Pavlova ON, Novikova AS, Kabilov MR, Zemskaya TI. Draft Genome Sequence of Thermaerobacter sp. Strain PB12/4term, a Thermophilic Facultative Anaerobic Bacterium from Bottom Sediments of Lake Baikal, Russia. Microbiology resource announcements. 2018; 7.
[107] Pachiadaki MG, Sintes E, Bergauer K, Brown JM, Record NR, Swan BK *et al*. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. Science. 2017; 358: 1046-1050.
[108] Tully BJ, Heidelberg JF. Potential mechanisms for microbial energy acquisition in oxic deep-sea sediments. Applied and environmental microbiology. 2016; 82: 4232-4243.

[109] Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the Tree of Life. Cell. 2018; 172: 1181-1197.
[110] Edgcomb VP. Marine protist associations and environmental impacts across trophic levels in the twilight zone and below. Curr Opin Microbiol. 2016; 31: 169-175.
[111] Forster D, Dunthorn M, Mahe F, Dolan JR, Audic S, Bass D *et al*. Benthic protists: the under-charted majority. FEMS Microbiol Ecol. 2016; 92: fiw120.
[112] Bjorbækmo MFM, Evenstad A, Røsæg LL, Krabberød AK, Logares R. The planktonic protist interactome: where do we stand after a century of research? ISME Journal. 2020; 14: 544-559.
[113] Roberts SL, Swann GEA, McGowan S, Panizzo VN, Vologina EG, Sturm M *et al*. Diatom evidence of 20th century ecosystem change in Lake Baikal, Siberia. PloS one. 2018; 13: e0208765.
[114] Edgcomb VP, Beaudoin D, Gast R, Biddle JF, Teske A. Marine subsurface eukaryotes: the fungal majority. Environmental microbiology. 2011; 13: 172-183.
[115] Lozupone CA, Knight R. Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104: 11436-11440.
[116] Mukherjee I, Hodoki Y, Okazaki Y, Fujinaga S, Ohbayashi K, Nakano SI. Widespread dominance of kinetoplastids and unexpected presence of diplonemids in deep freshwater lakes. Frontiers in microbiology. 2019; 10: 2375.
[117] Cabello-Yeves PJ, Rodriguez-Valera F. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. Microbiome. 2019; 7: 117.
[118] Brate J, Logares R, Berney C, Ree DK, Klaveness D, Jakobsen KS *et al*. Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. The ISME journal. 2010; 4: 1144-1153.
[119] Simon M, Lopez-Garcia P, Moreira D, Jardillier L. New haptophyte lineages and multiple independent colonizations of freshwater ecosystems. Environ Microbiol Rep. 2013; 5: 322-332.
[120] Richards TA, Bass D. Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis. Curr Opin Microbiol. 2005; 8: 240-252.
[121] Simon M, Jardillier L, Deschamps P, Moreira D, Restoux G, Bertolino P *et al*. Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. Environmental microbiology. 2015; 17: 3610-3627.
[122] Lara E, Moreira D, Vereshchaka A, Lopez-Garcia P. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environmental microbiology. 2009; 11: 47-55.
[123] de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R *et al*. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015; 348: 1261605.
[124] Cabello-Yeves PJ, Haro-Moreno JM, Martin-Cuadrado AB, Ghai R, Picazo A, Camacho A *et al*. Novel Synechococcus genomes reconstructed from freshwater reservoirs. Frontiers in microbiology. 2017; 8: 1151.
[125] Mehrshad M, Salcher MM, Okazaki Y, Nakano SI, Simek K, Andrei AS *et al*. Hidden in plain sight-highly abundant and diverse planktonic freshwater Chloroflexi. Microbiome. 2018; 6: 176.
[126] Mehrshad M, Rodriguez-Valera F, Amoozegar MA, Lopez-Garcia P, Ghai R. The enigmatic SAR202 cluster up close: shedding light on a globally distributed dark ocean lineage involved in sulfur cycling. The ISME journal. 2018; 12: 655-668.
[127] Henson MW, Lanclos VC, Faircloth BC, Thrash JC. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. The ISME journal. 2018; 12: 1846-1860.
[128] Zemskaya TI, Cabello-Yeves PJ, Pavlova ON, Rodriguez-Valera F. Microorganisms of Lake Baikal—the deepest and most ancient lake on Earth. Applied microbiology and biotechnology. 2020.
[129] Sheik CS, Jain S, Dick GJ. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. Environmental microbiology. 2014; 16: 304-317.
[130] Landa M, Burns AS, Durham BP, Esson K, Nowinski B, Sharma S *et al*. Sulfur metabolites that facilitate oceanic phytoplankton-bacteria carbon flux. The ISME journal. 2019; 13: 2536-2550.

[131] Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R *et al*. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environmental microbiology. 2008; 10**:** 3349-3365.

[132] Haeckel E. Report on radiolaria collected by H. M. S. Challenger during the years 1873-1876. Rep Sci Res Voyage HMS Challenger 1873-76. 1887; 18**:** 1-1803.

[133] Paver SF, Muratore D, Newton RJ, Coleman ML. Reevaluating the salty divide: phylogenetic specificity of transitions between marine and freshwater systems. mSystems. 2018; 3.

Figure Legends

Fig.1. Sampling points and overall prokaryotic and eukaryotic diversity in Baikal sediments. **A**, Bathymetric map of Lake Baikal showing the sampling sites and depths along the three major basins of the lake extending along the North-South latitude axis. **B**, relative abundance of clean merged reads (CMRs) representing the major prokaryotic taxa for each sampling location. **C**, relative abundance of CMRs corresponding to eukaryotic taxa. The asterisk shows the average diversity derived from two replicates from the same sampling site, after nested PCR amplification.

Fig.2. Boxplot distribution of identity percentages for Lake Baikal sediment 16S/18S rRNA gene sequences against their best hits in public databases. **A**, prokaryotic sequences. **B**, prokaryotic sequences. Sample names are coloured according to basin as in Fig 1.

Fig.3. Marine signature taxa in Lake Baikal sediments. Presence (light blue) / absence (white) matrix of typical marine taxa identified in Lake Baikal sediments. Each row represents a sampling location and each column a taxon. The barcharts represent the sum of the detected CMRs (dark red) and OTUs (light red) per typical marine taxon (top) and sampling location (right).

Fig.4. Comparison of Lake Baikal sediment community structure. Non-metric multidimensional scaling (NMDS) of Bray-Curtis dissimilarities based on OTU frequencies of both prokaryotic and eukaryotic OTUs. Each point represents a different sample. Ellipses enclose all points per depth category: shallow (<100 m), medium (100-800 m), deep (>800 m). Samples from the different Baikal basins are indicated with different marker shapes. BK22S was excluded for eukaryotic sequences (see text). NMDS for only prokaryotic and eukaryotic communities are presented in Supplementary Figure 8.

Fig.5. Core prokaryotic communities in Lake Baikal sediments across latitudinal and depth gradients. UpSet plot (central panel) showing the number, phylogenetic affiliation and relative abundance of OTUs within the core shared by all the lakes (left bar) or all the sediment samples but one (light grey dot; bars on the right). The bottom-left histogram shows the relative proportion (CMRs) of the prokaryotic core community in the total prokaryotic community of each sediment sample. The upper right inset shows the total number of shared prokaryotic and eukaryotic OTUs per groups of sediment samples.

Fig.6. Comparative analysis of prokaryotic community structure in upper-layer sediments from Lake Baikal and other freshwater, brackish and marine systems. **A**, NMDS of sediment samples based on

Bray-Curtis dissimilarities of bacterial high-rank taxa. Colored ellipses and symbols correspond to Baikal (light blue squares), other freshwater sediments (light green squares), brackish (red dots) and marine (dark blue triangles) sediment samples. **B**, Diversity barchart displaying the relative abundance of bacterial sequences in the different sediment samples (left) and the dendrogram (right) resulting from the corresponding clustering analysis based on the Bray-Curtis dissimilarity. Dendrogram leaves represent the NMDS points depicted in (A).

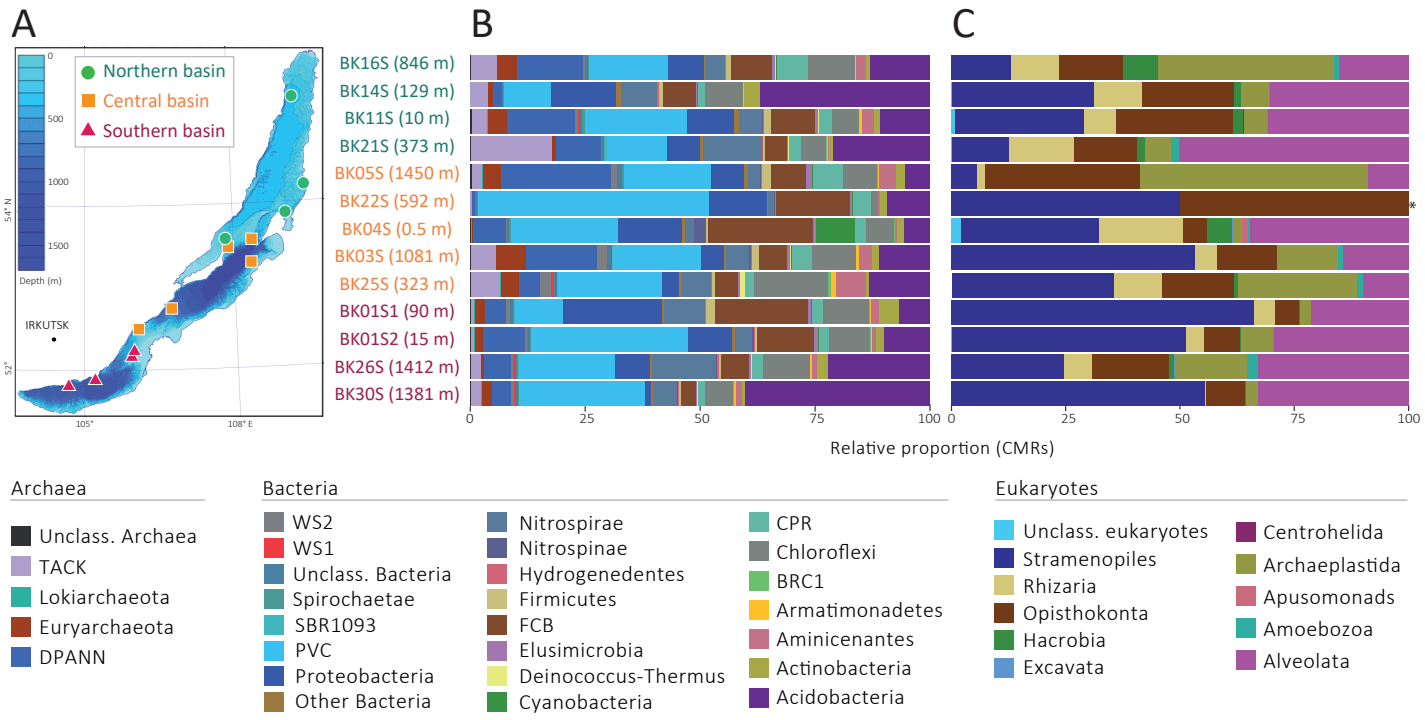


Figure 1. Reboul et al.

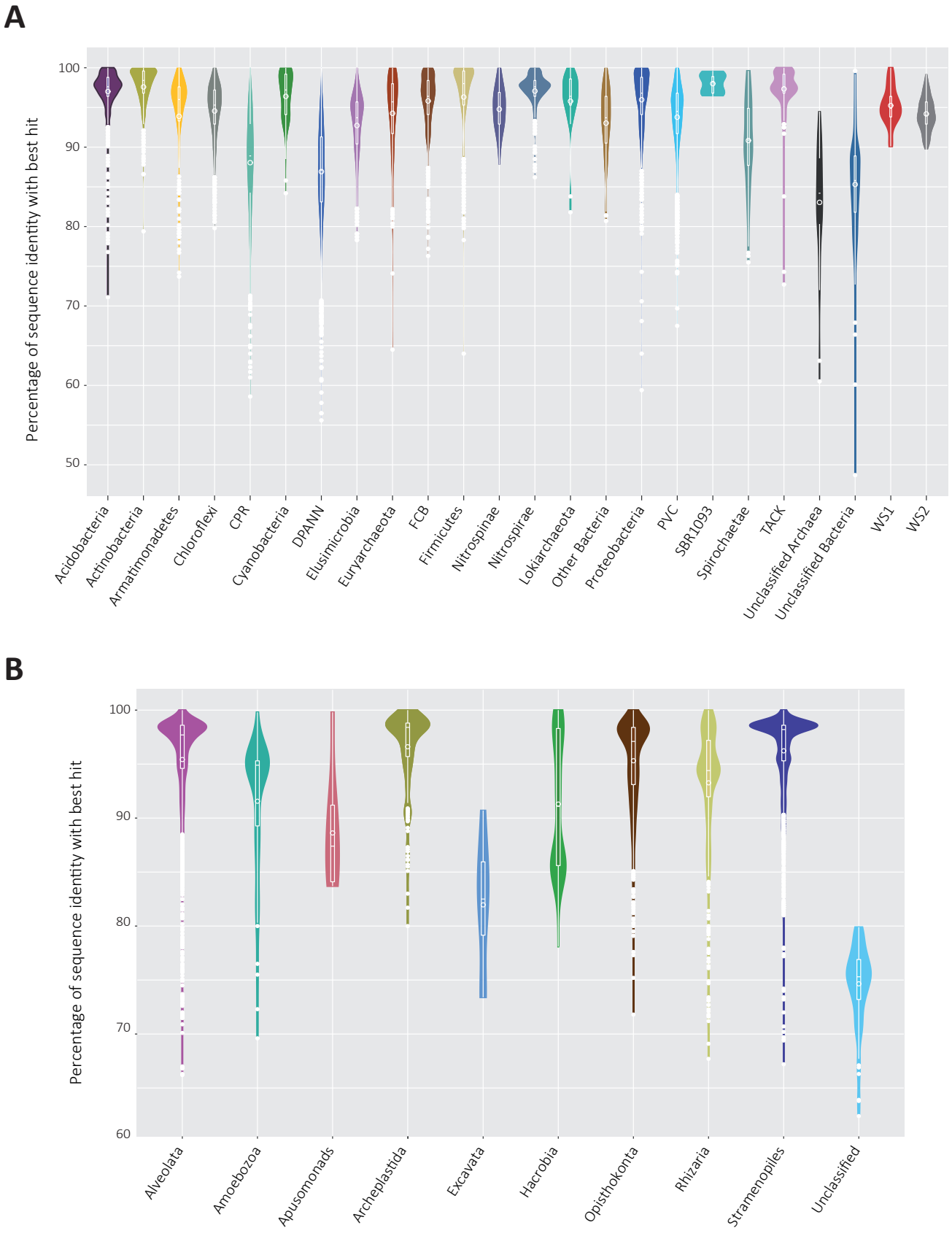


Figure 2. Reboul et al.

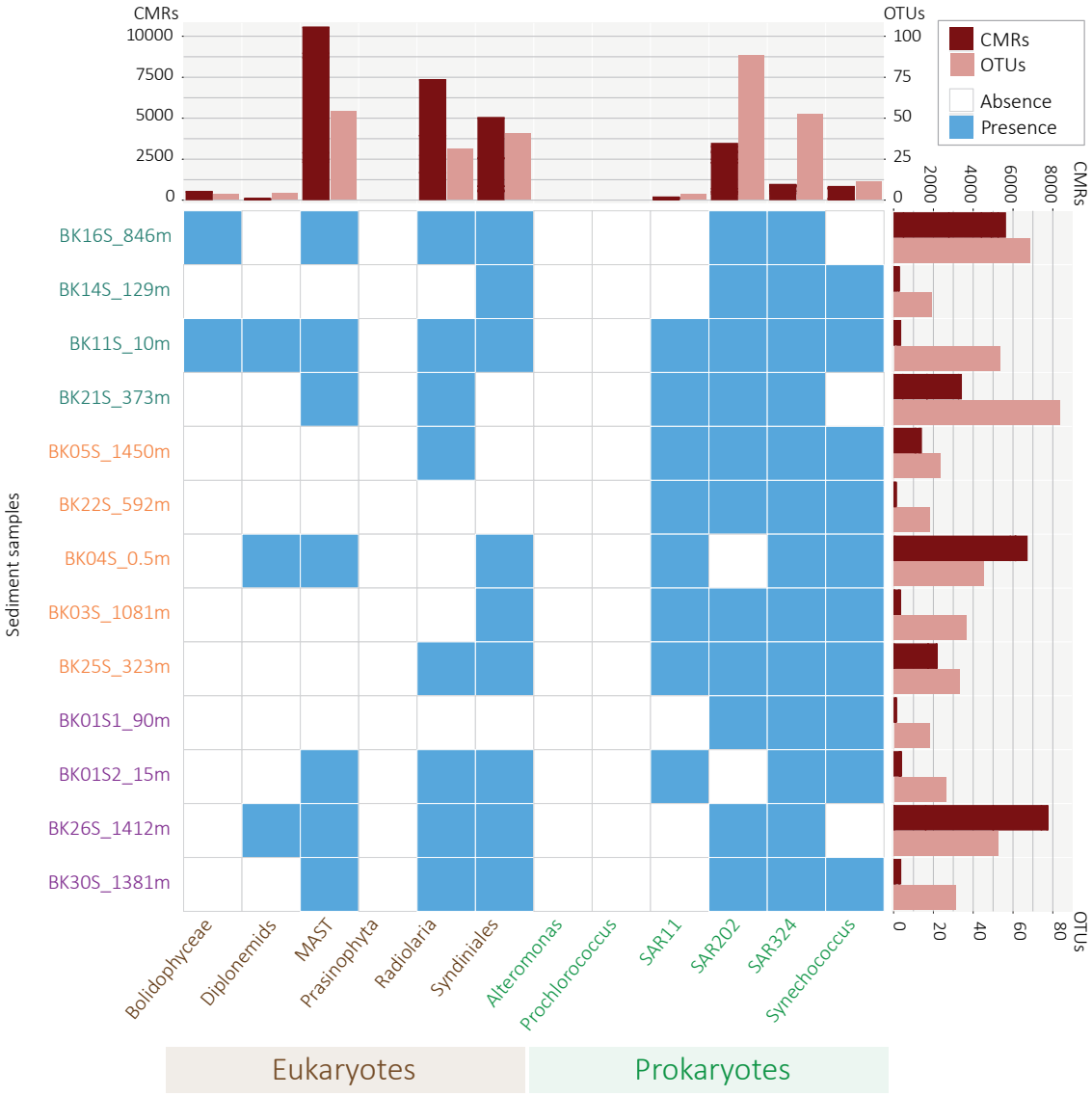


Figure 3. Reboul et al.

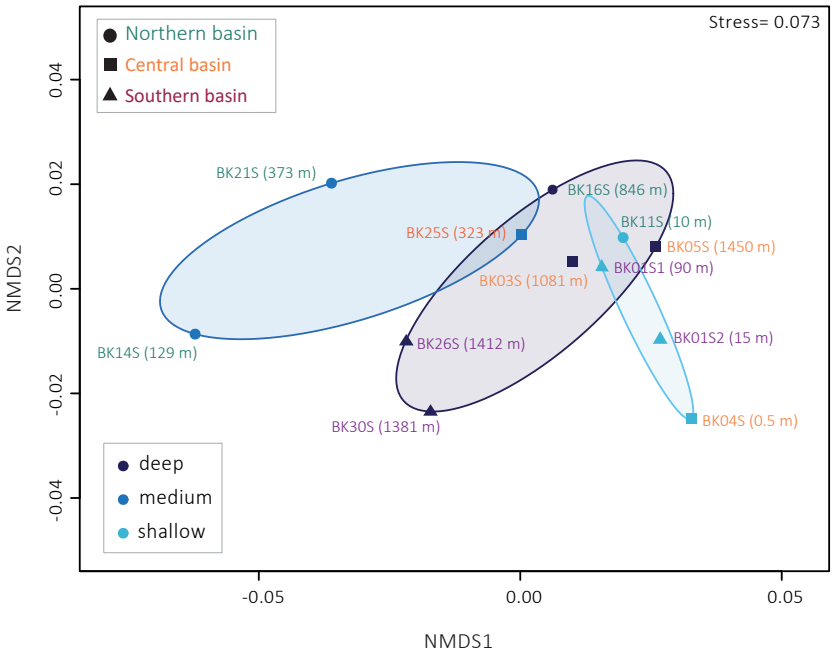


Figure 4. Reboul et al.

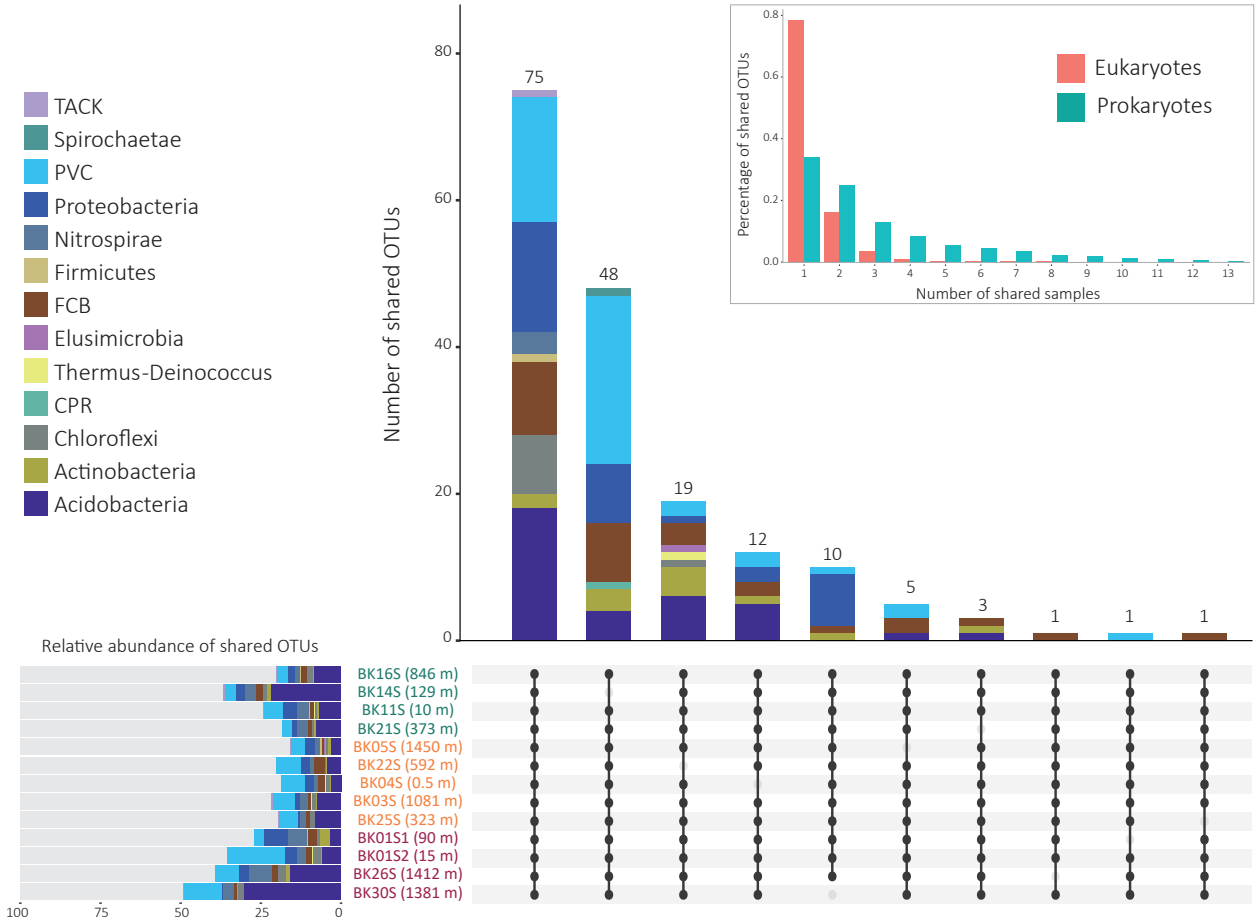


Figure 5. Reboul et al

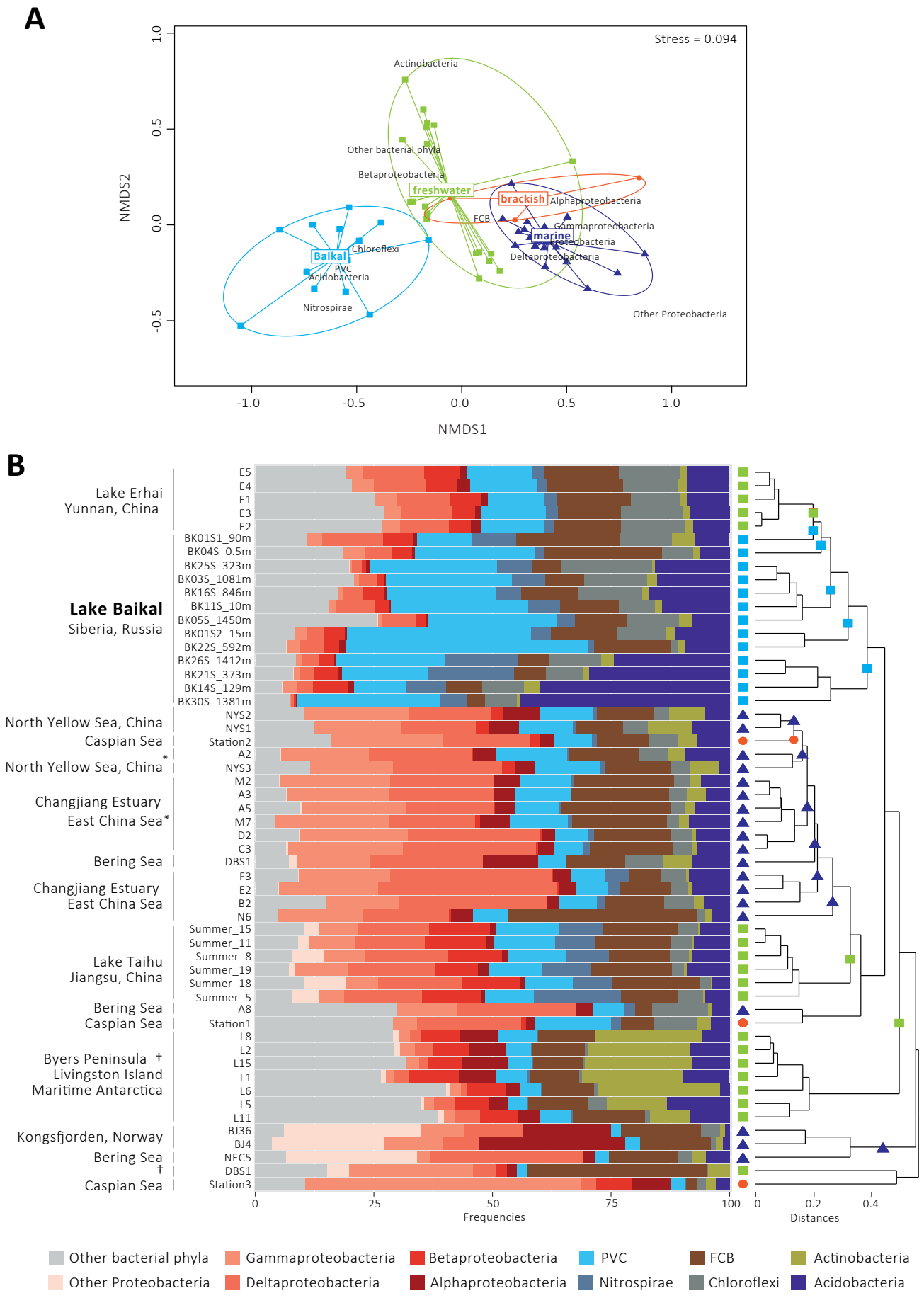
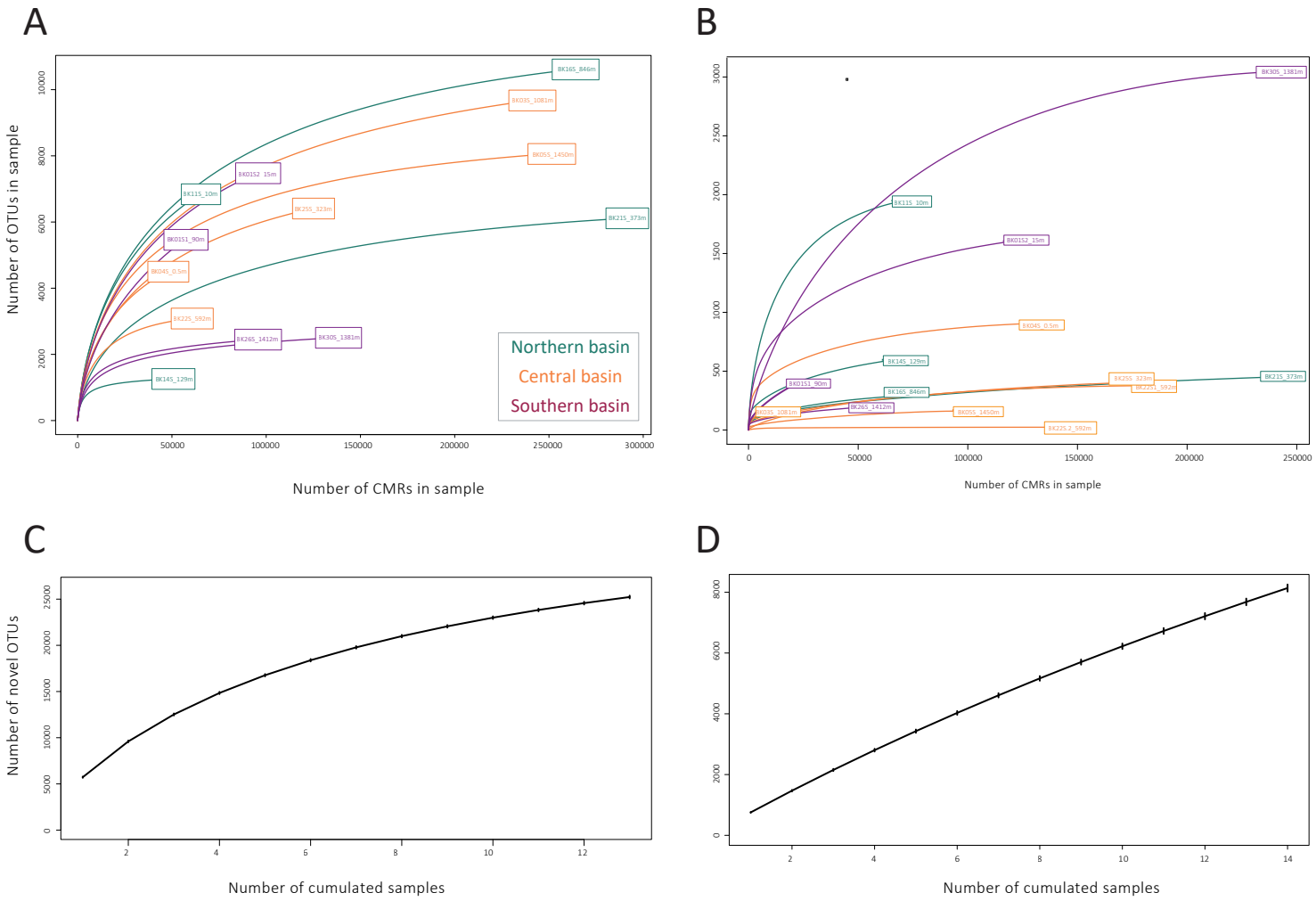
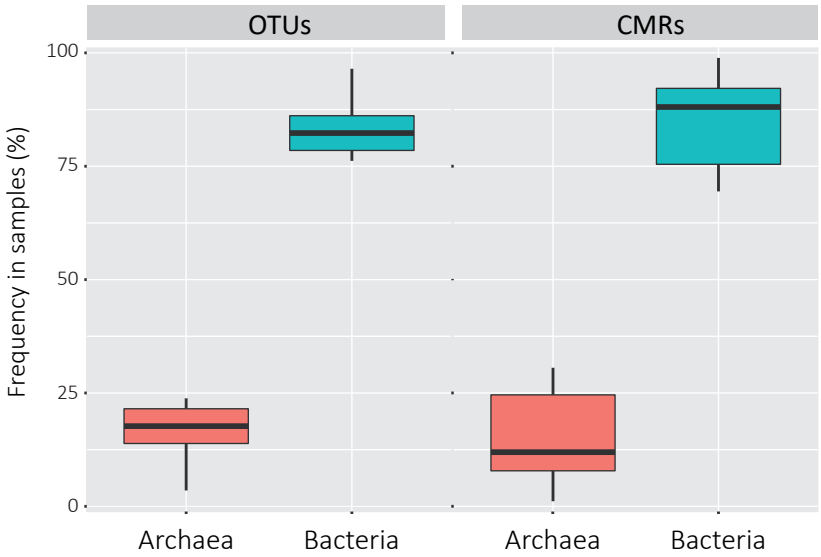


Figure 6. Reboul et al

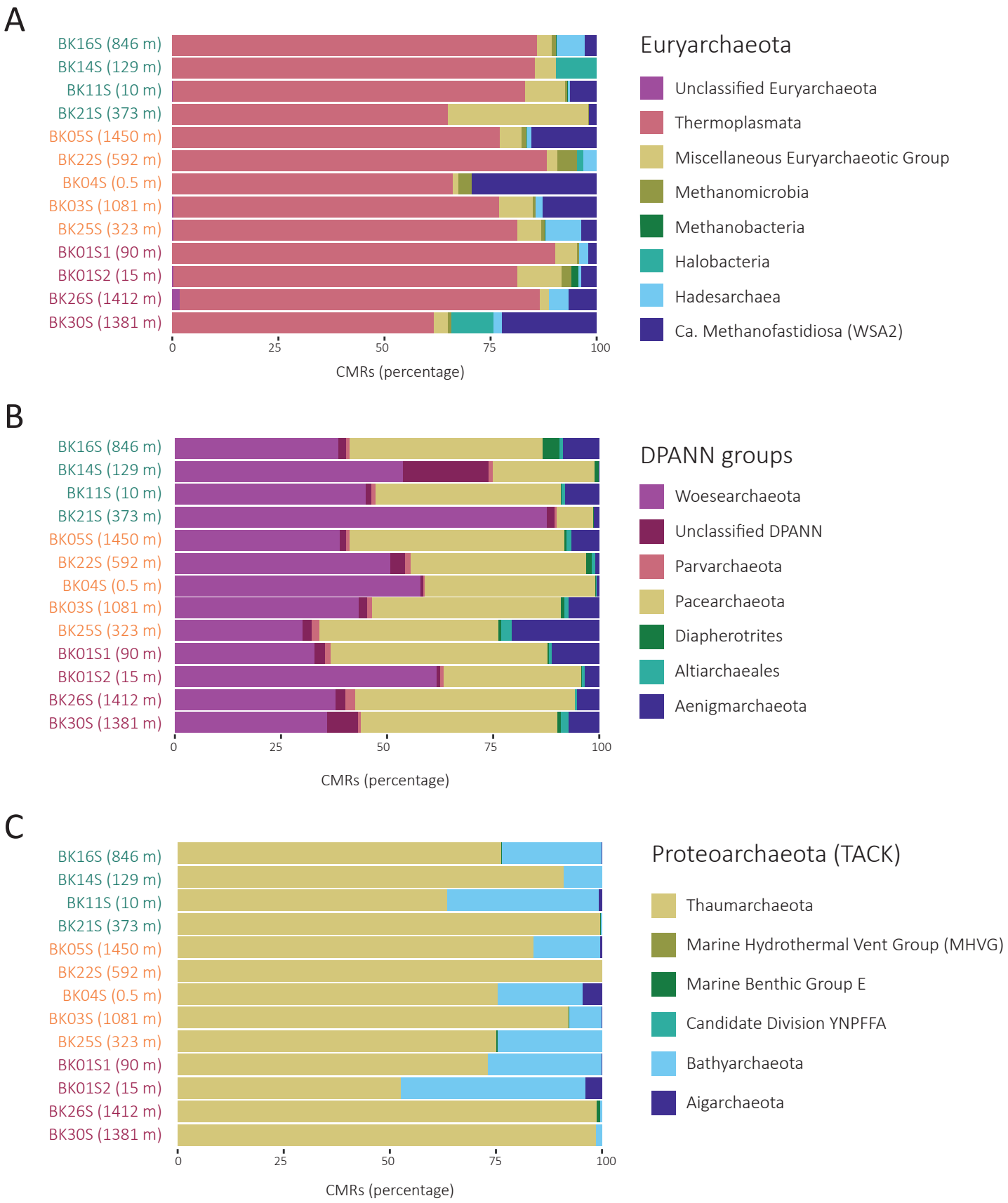
Supplementary information



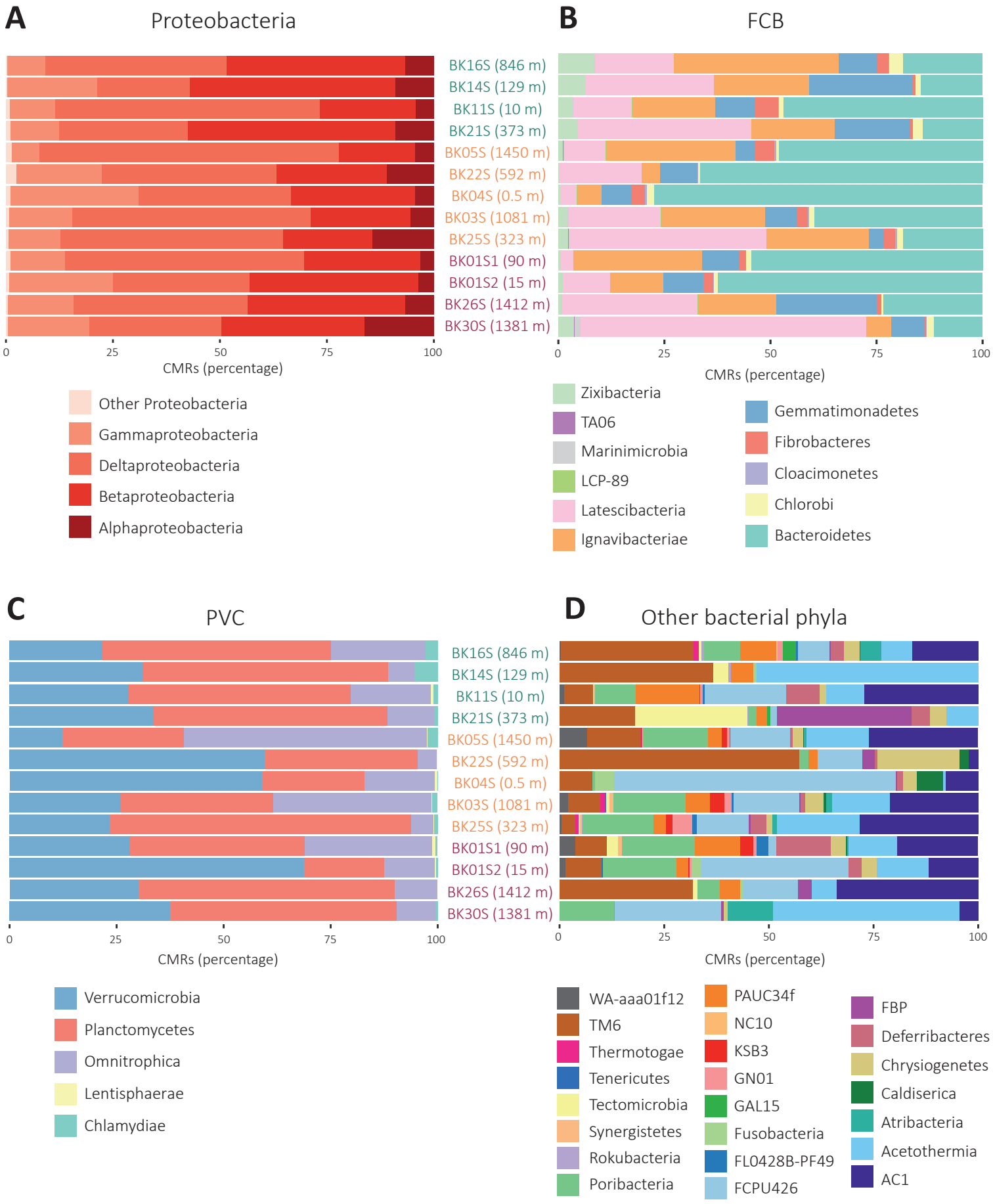
Supplementary Fig.1. Rarefaction and species accumulation curves for 16S and 18S rRNA gene amplicon sequences obtained from Lake Baikal sediment samples. **A**, cumulative count of operational taxonomic units (OTUs) as a function of clean merged reads (CMRs) (rarefaction curve) for prokaryotes. **B**, rarefaction curves for eukaryotes. **C**, cumulative number of OTUs as a function of the number of samples for prokaryotes. **D**, accumulation curve for eukaryotic OTUs. Sample names are colored according to their basin of provenance.



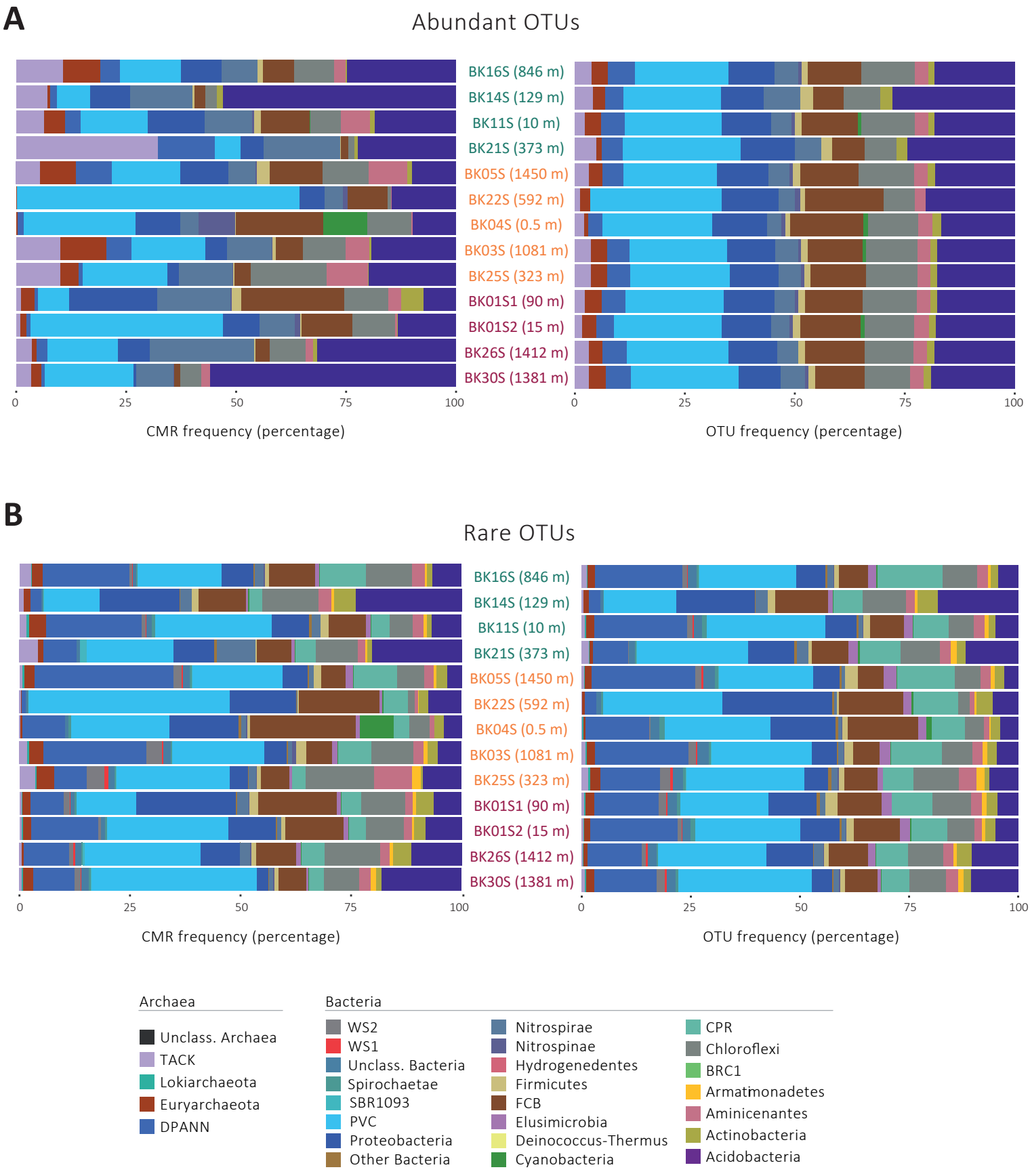
Supplementary Fig.2. Boxplot showing the global distribution of OTUs and CMRs into the two prokaryotic domains Archaea and Bacteria in Lake Baikal sediments.



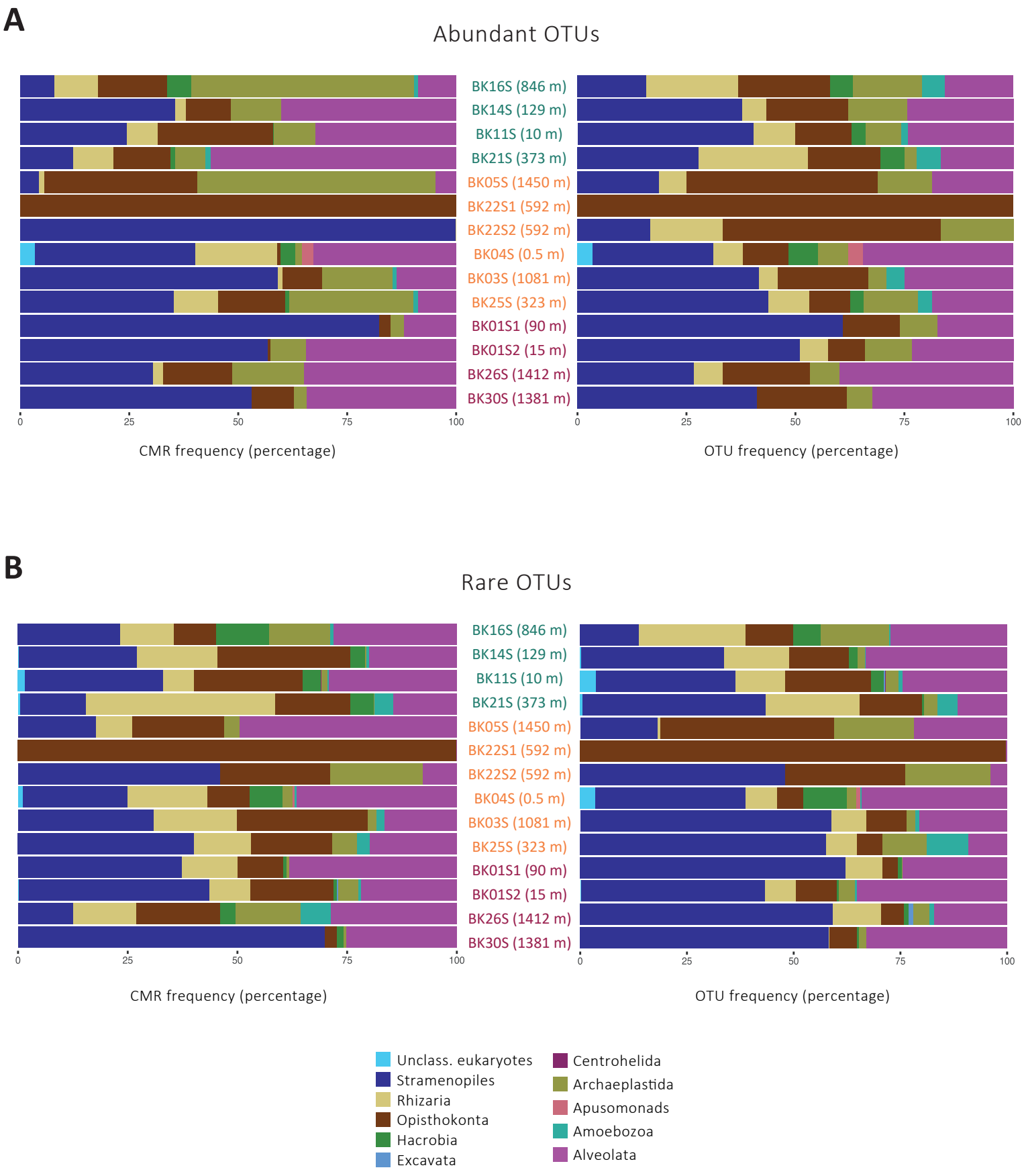
Supplementary Fig.3. Detailed relative proportions of different archaeal clades identified in Lake Baikal sediments. **A,** different members of the Euryarchaeota. **B**, members of the so-called DPANN clade. **C**, members of the Proteoarchaeota or TACK superphylum. The relative abundance of these three clades among prokaryotes is shown in Fig.1. CMR, clean merged read.



Supplementary Fig.4. Detailed relative proportions of selected bacterial clades identified in Lake Baikal sediments. **A**, Proteobacterial classes. **B**, members of the FCB (Fibrobacter-Chlorobi-Bacteroidetes) clade. **C**, members of the PVC (Planctomycetes-Verrucomicrobia-Chlamydiae) clade. **D**, members classed under the category 'other bacteria' in Fig.1.

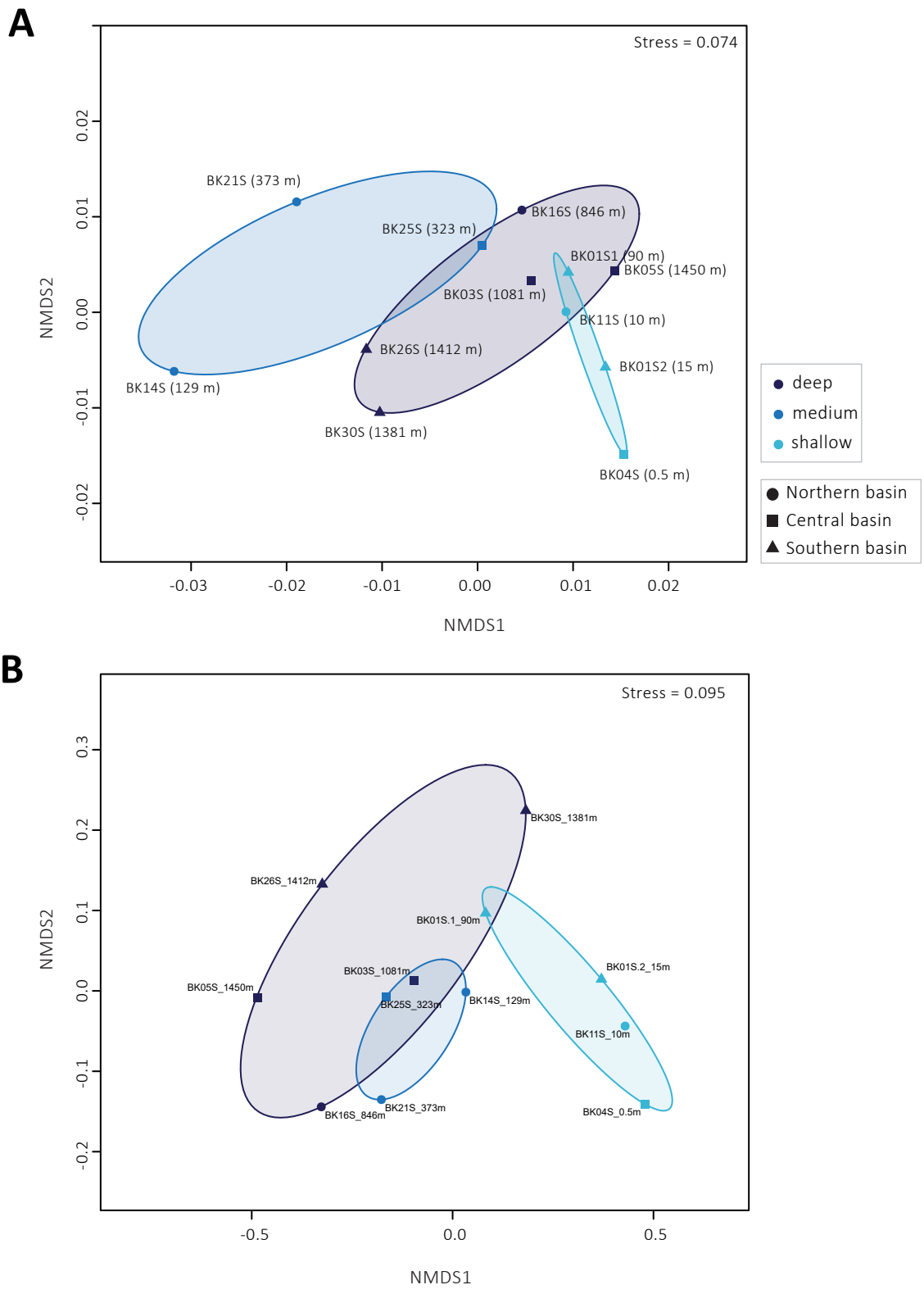


Supplementary Fig.5. Relative abundance and diversity of prokaryotic OTUs in Lake Baikal sediments. **A**, relative proportion (left) and diversity (right) of abundant OTUs (>0.1% CMRs). **B**, relative proportion (left) and diversity (right) of rare OTUs (<0.1% CMRs). CMR, clean merged read.

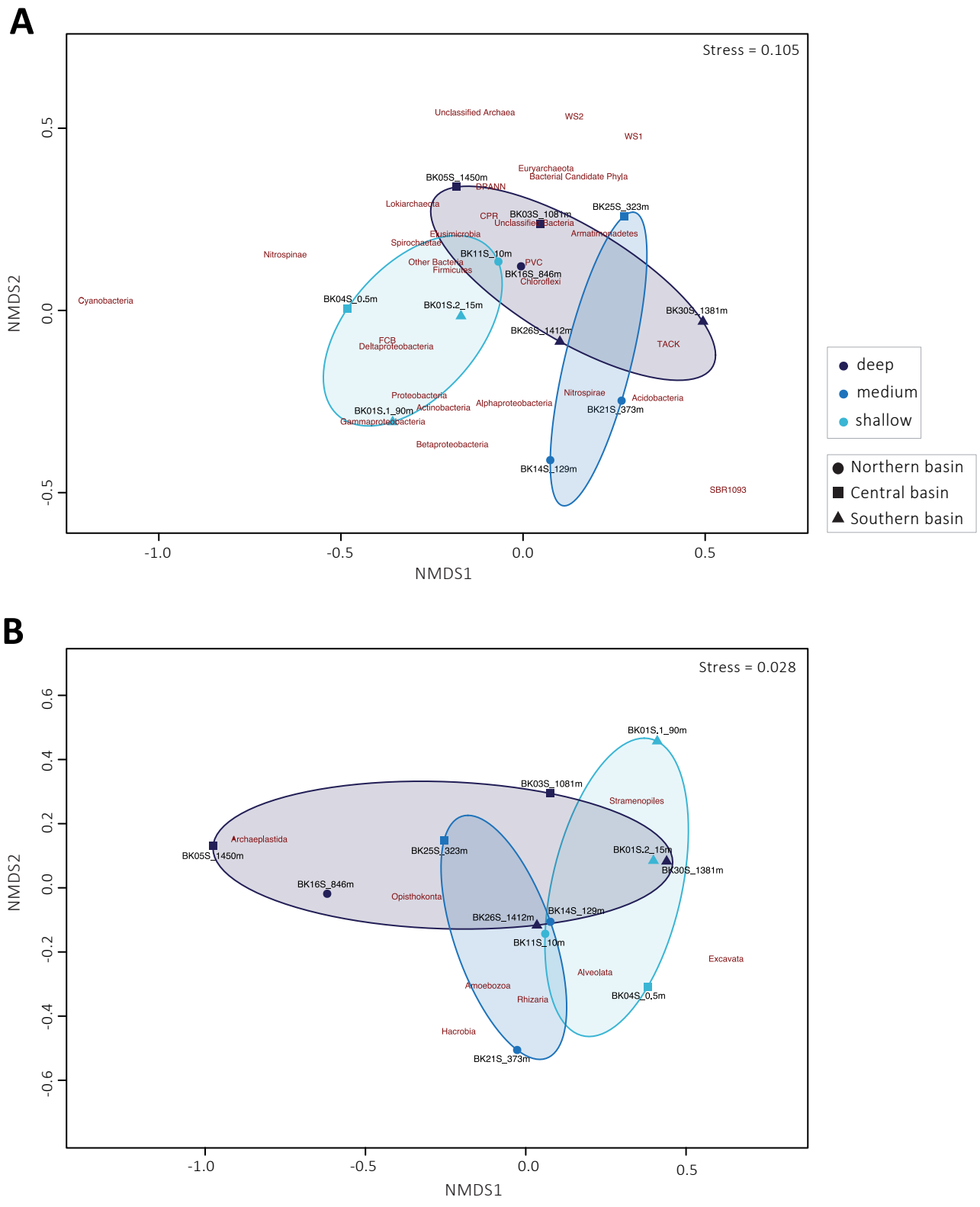


Supplementary Fig.6. Relative abundance and diversity of eukaryotic OTUs in Lake Baikal sediments. **A**, relative proportion (left) and diversity (right) of abundant OTUs (>0.1% CMRs). **B**, relative proportion (left) and diversity (right) of rare OTUs (<0.1% CMRs). CMR, clean merged read.

Supplementary Fig.7. Phylogenetic tree of 18S rRNA gene sequences amplified from Baikal sediment samples affiliating to Radiolaria. Our sequences are in blue. Bootstrap values >50% are indicated at nodes. The scale bar represents the number of substitutions per branch length unit.



Supplementary Fig.8. Comparison of Lake Baikal sediment prokaryotic and eukaryotic community structures based on dissimilarity of OTU matrices. **A**, Non-metric multidimensional scaling (NMDS) of Bray-Curtis dissimilarities based on frequencies of prokaryotic OTUs. **B**, NMDS of eukaryotic OTU matrix dissimilarities. Each point represents a different sample. Ellipses enclose all points per depth category: shallow (<100 m), medium (100-800 m), deep (>800 m). Samples from the different Baikal basins are indicated with different marker shapes. BK22S was excluded for eukaryotic sequences (see text).



Supplementary Fig.9. Comparison of Lake Baikal sediment prokaryotic and eukaryotic community structures based on dissimilarity of high-rank taxa matrices. **A**, NMDS of Bray-Curtis dissimilarities based on frequencies of prokaryotic phyla/classes. **B**, NMDS of eukaryotic taxa. Each point represents a different sample. Ellipses enclose all points per depth category: shallow (<100 m), medium (100-800 m), deep (>800 m). Samples from the different Baikal basins are indicated with different marker shapes. BK22S was excluded for eukaryotic sequences (see text).

Supplementary Table 1. Lake Baikal sediment sample metadata and prokaryotic and eukaryotic alpha diversity indices. Samples were classed according to their depth in shallow (0-100 m), medium (100-800 m) and deep (> 800 m)

Samples	Depth (m)	Depth category	Basin	Latitude	longitude	Temperature (°C)	collection date	Prokaryotic dataset alpha diversity indices						Eukaryotic dataset alpha diversity indices					
								Richness	Chao1	ACE	Shannon	Inverse Simpson	Evenness	Richness	Chao1	ACE	Shannon	Inverse Simpson	Evenness
BK16S_846m	846	deep	Northern	55.06.259 N	109.16.104 E	4.2	02/07/2017	47287	12257.22	12548.25	7.141625	221.7085	0.8074011	3617	225.5	234.650312	3.3262	10.6752	0.719182
BK14S_129m	129	medium	Northern	54.09.926 N	109.31.465 E	4.5	01/07/2017	47287	1363.14	1324.748	5.748503	113.3526	0.80614	3603	376.111111	270.906694	4.4852	49.0222	0.842605
BK11S_10m	10	shallow	Northern	53.45.579 N	109.02.15 E	6.0	30/06/2017	47290	9025.73	9409.718	7.195588	332.7652	0.8241107	3619	2031.09859	2014.50625	5.0382	49.0209	0.773000
BK21S_373m	373	medium	Northern-Central transition	53.30.175 N	107.52.633 E	NA	04/07/2017	47296	7035.577	6904.731	5.911225	71.85421	0.714756	3613	130.375	121.57773	3.1314	8.5362	0.667478
BK05S_1450m	1450	deep	Central	53.31.096 N	108.24.583 E	3.6	29/06/2017	47301	8991.186	9086.944	7.108497	320.2212	0.8199202	3619	132.5	129.08	2.3961	6.4195	0.622352
BK22S.1_592m	592	medium	Northern-Central transition	53.23.524 N	107.53.094 E	4.6	04/07/2017	47288	3255.25	3321.675	6.487509	172.2678	0.8103274	/	/	/	/	/	/
BK22S.2_592m	592	medium	Northern-Central transition	53.23.524 N	107.53.094 E	4.6	04/07/2017	/	/	/	/	/	/	/	/	/	/	/	/
BK04S_0.5m	0.5	shallow	Central	53.14.596 N	108.30.874 E	18.5	29/06/2017	47282	5663.326	5940.309	6.819175	243.6826	0.8107074	3611	481.9375	457.24582	4.7240	41.40111	0.800320
BK03S_1081m	1081	deep	Central	52.41.4014 N	106.44.208 E	4.0	29/06/2017	47256	10895.84	11232.18	6.941362	190.1485	0.7924655	3618	243.5	235.587434	3.1767	11.10596	0.662398
BK25S_323m	323	medium	Southern-Central transition	52.29.854 N	106.05.288 E	NA	04/07/2017	47253	7667.917	7818.66	6.616539	159.3584	0.7773943	3615	140.2	128.62552	3.1471	11.1286	0.708387
BK01S_1_90m	90	shallow	Southern-Central transition	52.15.70 N	106.02.90 E	6.5	28/06/2017	47286	8065.588	8527.433	6.650706	191.8113	0.7796562	3625	516.66667	574.38406	3.5752	16.7737	0.688461
BK01S_2_15m	15	shallow	Southern-Central transition	52.13.70 N	106.09.90 E	9.0	28/06/2017	47251	9585.266	10058.66	6.876209	166.0523	0.7898247	3610	859.0098	898.62375	4.6205	29.8595	0.733536
BK26S_1412m	1412	deep	Southern	51.52.628 N	105.15.294 E	3.9	05/07/2017	47310	2748.947	2633.66	5.947021	75.605	0.7719483	3615	102.5	106.718058	3.2438	16.2828	0.753665
BK30S_1381m	1381	deep	Southern	51.47.817 N	104.46.449 E	NA	07/07/2017	47265	2632.101	2604.098	5.57636	60.40059	0.7291925	3607	1502.03704	1852.0187	3.7377	18.4660	0.641228

Supplementary Table 2. Sequence data for 16S and 18S rRNA amplicons obtained from Lake Baikal sediment samples. CMR, clean merged reads; OTU, operational taxonomic unit.

Kingdom	Phyla	RAW DATA				RAREFIED DATA			
		CMRs (#)	CMRs (%)	OTUs (#)	OTUs (%)	CMRs (#)	CMRs (%)	OTUs (#)	OTUs (%)
16S rRNA gene amplicons									
Archaea	DPANN	207079	11.6723	12060	16.2055	55558	9.0383	9626	16.4736
Archaea	Euryarchaeota	55246	3.1140	1082	1.4539	15513	2.5237	872	1.4923
Archaea	Lokiarchaeota	526	0.0296	91	0.1223	198	0.0322	77	0.1318
Archaea	TACK	103976	5.8607	617	0.8291	24131	3.9257	471	0.8061
Archaea	Unclassified Archaea	1782	0.1004	93	0.1250	515	0.0838	77	0.1318
Bacteria	Acidobacteria	276976	15.6121	4604	6.1866	97734	15.8996	3679	6.2961
Bacteria	Actinobacteria	25631	1.4447	1568	2.1070	10616	1.7270	1218	2.0844
Bacteria	Armatimonadetes	7109	0.4007	682	0.9164	2421	0.3939	540	0.9241
Bacteria	Bacterial Candidate Phyla	52120	2.9378	1802	2.4214	16612	2.7025	1378	2.3583
Bacteria	Chloroflexi	148614	8.3768	5318	7.1460	50005	8.1350	4200	7.1877
Bacteria	CPR	68023	3.8342	7828	10.5188	19524	3.1762	5695	9.7462
Bacteria	Cyanobacteria	5360	0.3021	181	0.2432	4407	0.7169	151	0.2584
Bacteria	Deferribacteres	264	0.0149	59	0.0793	117	0.0190	39	0.0667
Bacteria	Elusimicrobia	10852	0.6117	1643	2.2078	3365	0.5474	1191	2.0382
Bacteria	FBP	507	0.0286	13	0.0175	84	0.0137	10	0.0171
Bacteria	FCB	141260	7.9623	6324	8.4978	61341	9.9791	5268	9.0155
Bacteria	FCPU426	1288	0.0726	145	0.1948	575	0.0935	124	0.2122
Bacteria	Firmicutes	18129	1.0219	1280	1.7200	6010	0.9777	935	1.6001
Bacteria	FL0428B-PF49	32	0.0018	9	0.0121	15	0.0024	7	0.0120
Bacteria	GAL15	41	0.0023	2	0.0027	9	0.0015	2	0.0034
Bacteria	GN01	90	0.0051	12	0.0161	27	0.0044	8	0.0137
Bacteria	New Baikal Group	282	0.0159	59	0.0793	98	0.0159	44	0.0753
Bacteria	Nitrospinae	4797	0.2704	133	0.1787	2432	0.3956	108	0.1848
Bacteria	Nitrospirae	113421	6.3931	995	1.3370	37411	6.0861	760	1.3006
Bacteria	PAUC34f	464	0.0262	49	0.0658	204	0.0332	39	0.0667
Bacteria	Proteobacteria	134874	7.6023	6948	9.3363	57799	9.4029	5502	9.4159
Bacteria	PVC	360967	20.3463	17722	23.8138	137371	22.3479	13994	23.9488
Bacteria	SBR1093	2872	0.1619	24	0.0322	613	0.0997	20	0.0342
Bacteria	Spirochaetae	3397	0.1915	547	0.7350	1377	0.2240	426	0.7290
Bacteria	Synergistetes	21	0.0012	8	0.0107	8	0.0013	6	0.0103
Bacteria	Tenericutes	3	0.0002	2	0.0027	2	0.0003	2	0.0034
Bacteria	Unclassified Bacteria	11599	0.6538	1522	2.0452	3745	0.6092	1151	1.9698
Bacteria	WA-aaa01f12	185	0.0104	16	0.0215	45	0.0073	11	0.0188
Bacteria	WS1	1936	0.1091	176	0.2365	625	0.1017	137	0.2345
Bacteria	WS2	14389	0.8111	805	1.0817	4186	0.6810	665	1.1381
TOTAL		**1774112**	**100**	**74419**	**100**	**614693**	**100**	**58433**	**100**
18S rRNA gene amplicons									
Eukaryota	Alveolata	387133	23.7711	2924	27.6815	11153	22.0400	746	25.8400
Eukaryota	Amoebozoa	9091	0.5582	94	0.8899	292	0.5770	31	1.0738
Eukaryota	Apusozoa	1966	0.1207	9	0.0852	45	0.0890	4	0.1386
Eukaryota	Archaeplastida	170827	10.4893	373	3.5312	6347	12.5430	134	4.6415
Eukaryota	Centrohelida	9	0.0006	1	0.0095	0	0.0000	0	0.0000
Eukaryota	Excavata	168	0.0103	9	0.0852	7	0.0140	5	0.1732
Eukaryota	Hacrobia	22449	1.3784	207	1.9597	798	1.5770	97	3.3599
Eukaryota	Opisthokonta	365294	22.4301	1518	14.3709	10158	20.0740	439	15.2061
Eukaryota	Rhizaria	107524	6.6023	769	7.2801	3390	6.6990	342	11.8462
Eukaryota	Stramenopiles	560532	34.4183	4553	43.1033	18304	36.1720	1057	36.6124
Eukaryota	Unclassified	3595	0.2207	106	1.0035	109	0.2150	32	1.1084
TOTAL		**1628588**	**100**	**10563**	**100**	**50603**	**100**	**2887**	**100**

Supplementary Table 3. Identification, phylogenetic affinity and relative abundance of prokaryotic OTUs identified in Lake Baikal sediments

Supplementary Table 4. Identification, phylogenetic affinity and relative abundance of eukaryotic OTUs identified in Lake Baikal sediments.

….

Supplementary Table 7. Prokaryotic and eukaryotic OTUs of typical marine ecosystems

(too large to be displayed in pdf; available in excel format)

Supplementary Table 5. Phylum-level matrix of total and rarefied 16S rDNA clean merged reads (CMRs) in Lake Baikal sediment samples

		raw count data - CMR												rarefied count data - CMR											
kingdom	**phyla**	**BK16S_846m**	**BK14S_129m**	**BK11S_10m**	**BK21S_373m**	**BK05S_1450m**	**BK22S_592m**	**BK04S_0.5m**	**BK03S_1081m**	**BK25S_323m**	**BK01S.1_90m**	**BK01S.2_15m**	**BK26S_1412m**	**BK30S_1381m**	**BK16S_846m**	**BK14S_129m**	**BK11S_10m**	**BK21S_373m**	**BK05S_1450m**	**BK22S_592m**	**BK04S_0.5m**	**BK03S_1081m**	**BK25S_323m**	**BK01S.1_90m**	**BK01S.2_15m**	**BK26S_1412m**	**BK30S_1381m**
Archaea	DPANN	37599	1023	9627	29387	59563	384	3416	37558	9583	2590	8808	5659	5902	6752	949	7082	4746	11297	307	3416	7445	2150	2170	4383	2794	2067
Archaea	Euryarchaeota	11920	577	2711	2117	10323	128	159	15264	5138	1337	1673	711	3188	2124	547	2005	337	1941	99	159	3011	1922	1118	813	355	1082
Archaea	Lokiarchaeota	78	0	68	0	104	0	25	93	63	28	56	7	4	13	0	46	0	17	0	25	18	22	24	28	4	1
Archaea	TACK	14537	1867	2201	52000	5753	157	65	12965	7978	451	721	2095	3186	2624	1766	1607	8372	1089	124	65	2564	3036	375	363	1046	1100
Archaea	Unclassified Archaea	285	0	179	39	668	0	20	335	75	28	45	16	92	50	0	132	7	119	0	20	65	30	25	20	9	38
Bacteria	Acidobacteria	34025	18498	7077	62096	13656	5605	2683	26661	16740	3818	9600	21083	55434	6159	17579	5126	10064	2540	4415	2683	5175	6396	3211	4794	10555	19037
Bacteria	Actinobacteria	2684	1638	791	2553	4970	960	691	3640	553	2558	1576	2258	759	477	1549	584	405	930	760	691	737	207	2143	765	1111	257
Bacteria	Armatimonadetes	580	155	302	862	1011	79	39	1202	1397	200	262	381	639	103	147	221	149	194	65	39	250	534	172	136	182	229
Bacteria	Bacterial Candidate Phyla	7419	782	2454	2523	11057	353	416	8717	11034	1265	1231	1779	3090	1337	740	1811	399	2128	285	416	1709	4193	1057	603	879	1055
Bacteria	Chloroflexi	26308	4055	3807	16315	18843	832	2875	23114	20279	5686	8542	9581	8377	4739	3852	2770	2640	3540	648	2875	4485	7675	4768	4333	4789	2891
Bacteria	CPR	17907	817	1718	7032	16481	2433	1203	10469	2195	1426	2132	2319	1891	3227	775	1235	1171	3088	1924	1203	2057	822	1176	1023	1176	647
Bacteria	Cyanobacteria	373	45	15	269	95	113	3961	162	17	5	129	31	145	69	42	12	49	16	89	3961	30	7	4	62	19	47
Bacteria	Deferribacteres	30	0	51	65	16	1	4	12	29	43	13	0	0	5	0	33	10	4	1	4	3	15	36	6	0	0
Bacteria	Elusimicrobia	1623	130	434	1217	3047	280	293	1791	391	322	646	487	191	279	123	323	190	568	222	293	342	148	268	315	231	63
Bacteria	FBP	2	0	0	483	0	4	1	2	4	0	0	8	3	1	0	0	74	0	3	1	0	2	0	0	3	0
Bacteria	FCB	23854	3504	6146	14097	19174	9571	10757	14263	6478	11447	11739	5580	4650	4281	3325	4503	2280	3599	7535	10757	2835	2421	9517	5859	2817	1612
Bacteria	FCPU426	72	0	126	25	294	14	191	170	93	6	151	32	114	16	0	92	4	55	12	191	39	40	3	67	16	40
Bacteria	Firmicutes	2744	480	1090	690	5141	90	233	4125	653	1215	589	564	515	507	460	797	121	977	69	233	817	257	1027	281	281	183
Bacteria	FLO428B-PF49	4	0	3	0	2	0	0	5	9	9	0	0	0	1	0	3	0	0	0	0	1	4	6	0	0	0
Bacteria	GAL15	30	0	0	11	0	0	0	0	0	0	0	0	0	7	0	0	2	0	0	0	0	0	0	0	0	0
Bacteria	GN01	11	0	4	0	15	0	0	19	36	3	2	0	0	1	0	3	0	4	0	0	1	14	3	1	0	0
Bacteria	New Baikal Group	35	0	11	60	51	26	9	48	10	12	16	0	4	6	0	8	11	9	18	9	12	3	11	10	0	1
Bacteria	Nitrospinae	319	0	122	2	331	168	1226	1247	225	74	569	176	338	54	0	85	0	63	129	1226	236	86	63	293	88	109
Bacteria	Nitrospirae	10859	3931	3323	37815	6878	780	947	12090	7847	4940	3537	13222	7252	1996	3736	2428	6129	1306	622	947	2294	2996	4113	1762	6596	2486
Bacteria	PAUC34f	82	30	98	38	68	3	0	61	23	36	12	12	1	15	28	70	9	13	3	0	11	8	31	7	8	1
Bacteria	Proteobacteria	20727	7009	6545	21382	18190	7496	6624	11886	4400	12126	9186	7396	1907	3740	6671	4773	3439	3456	5899	6624	2363	1701	10122	4609	3746	656
Bacteria	PVC	44945	5098	14412	37956	47215	30287	10965	46089	28388	6035	32324	19652	37601	8124	4842	10567	6129	8931	23847	10965	9113	10797	5052	16207	9844	12953
Bacteria	SBR1093	22	18	16	2255	31	0	0	77	97	22	1	226	107	3	17	12	367	4	0	0	17	37	17	1	106	32
Bacteria	Spirochaetae	476	44	298	49	899	71	194	439	166	162	280	127	192	92	41	231	7	183	63	194	81	66	137	145	66	71
Bacteria	Synergistetes	1	0	2	0	3	0	0	9	3	3	0	0	0	0	0	1	0	1	0	0	2	1	3	0	0	0
Bacteria	Tenericutes	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
Bacteria	Unclassified Bacteria	1391	107	515	1078	2435	202	247	2287	1046	319	436	591	945	264	103	382	170	469	161	247	454	408	257	214	300	316
Bacteria	WA-aaa01f12	2	0	7	0	134	0	0	21	3	12	6	0	0	0	0	4	0	25	0	0	5	0	9	2	0	0
Bacteria	WS1	95	0	74	9	443	0	5	304	580	37	24	217	148	20	0	52	2	83	0	5	58	217	31	13	98	46
Bacteria	WS2	1094	0	360	48	3196	0	33	4835	2868	436	380	371	767	192	0	266	7	630	0	33	981	1080	355	194	179	269
TOTAL		**262133**	**49788**	**64587**	**292473**	**250087**	**60037**	**47282**	**239962**	**124402**	**56651**	**94687**	**94581**	**137442**	**47278**	**47292**	**47264**	**47290**	**47279**	**47300**	**47282**	**47212**	**47295**	**47304**	**47310**	**47298**	**47289**

Supplementary Table 6. Phylum-level matrix of total and rarefied 18S rDNA clean merged reads (CMRs) in Lake Baikal sediment samples

	raw count data - CMR											rarefied count data - CMR										
phyla	**BK16S_846m**	**BK14S_129m**	**BK11S_10m**	**BK21S_373m**	**BK05S_1450m**	**BK22S.1_592m**	**BK22S.2_592m**	**BK04S_0.5m**	**BK03S_1081m**	**BK25S_323m**	**BK01S.1_90m**	**BK01S.2_15m**	**BK26S_1412m**	**BK30S_1381m**	**BK16S_846m**	**BK14S_129m**	**BK11S_10m**	**BK21S_373m**	**BK05S_1450m**	**BK22S.1_592m**	**BK22S.2_592m**	**BK04S_0.5m**	**BK03S_1081m**	**BK25S_323m**	**BK01S.1_90m**	**BK01S.2_15m**	**BK26S_1412m**	**BK30S_1381m**
Alveolata	10678	20603	21956	122021	8999	4	40	45282	518	16879	4648	36428	18016	81061	549	1075	1103	1804	321	0	1	1240	518	337	790	1032	1158	1225
Amoebozoa	623	205	162	4003	0	0	0	3911	41	2173	0	267	1226	0	30	10	8	61	0	0	0	12	41	47	0	10	73	0
Apusozoa	0	0	0	0	0	0	0	1966	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0
Archaeplastida	26873	4089	3685	14231	50294	0	172	2514	478	44716	486	8295	8691	6303	1408	219	189	203	1803	0	3	72	478	964	81	245	591	91
Centrohelida	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Excavata	0	0	40	0	0	0	0	34	0	0	0	86	8	0	0	0	2	0	0	0	0	1	0	0	0	3	1	0
Hacrobia	5359	1123	1621	4256	0	0	0	7143	0	1274	61	430	606	576	274	58	91	58	0	0	0	218	0	27	12	12	38	10
Opisthokonta	9689	13689	18298	33224	34185	180637	223	6887	480	26665	1174	9768	9279	21096	499	725	912	509	1231	3612	4	200	480	558	205	309	610	304
Rhizaria	7471	7074	5097	34592	1826	0	29	24124	171	17981	986	4762	3393	18	393	385	256	515	64	0	2	662	171	397	162	139	244	0
Stramenopiles	9025	21255	20183	30629	5592	0	143581	39434	1930	60754	14434	63449	13440	136826	465	1145	1014	464	194	0	3608	1087	1930	1286	2371	1861	905	1974
Unclassified	0	31	599	146	0	0	0	2759	0	0	0	60	0	0	0	1	28	2	0	0	0	76	0	0	0	2	0	0
TOTAL	**69718**	**68069**	**71650**	**243102**	**100896**	**180641**	**144045**	**130534**	**3618**	**170442**	**21789**	**123545**	**54659**	**245880**	**3618**	**3618**	**3603**	**3616**	**3613**	**3612**	**3618**	**3613**	**3618**	**3616**	**3621**	**3613**	**3620**	**3604**

Supplementary Table 8. Statistical PERMANOVA and ANOSIM analyses.

Prokaryotes - PERMANOVA (without BK22S)

Bray-Curtis dissimilarities based on rarefied prokaryotic OTUs standardized using Wisconsin method ~ basin (3 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
basin	2	0.76060	0.38030	0.97126	0.17752	0.51090	
residuals	9	3.52400	0.39156		0.82248		
total	11	4.28460			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied prokaryotic OTUs standardized using Wisconsin method ~ depth (3 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
depth	2	1.01050	0.50527	1.38890	0.23585	0.01070	*
residuals	9	3.27410	0.36379		0.76415		
total	11	4.28460			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

OTU-based analyses

Bray-Curtis dissimilarities based on rarefied prokaryotic Phyla and Proteobacterial classes ~ basin (3 categories)

number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
basin	2	0.11447	0.05724	0.95839	0.17558	0.50440	
residuals	9	0.53748	0.59720		0.82442		
total	11	0.65195			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied prokaryotic Phyla and Proteobacterial classes ~ depth (3 categories)

number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
depth	2	0.20367	0.10183	2.04440	0.31239	0.03130	*
residuals	9	0.44829	0.04981		0.68761		
total	11	0.65196			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Phyla-based analyses

Prokaryotes - ANOSIM (without BK22S)

Bray-Curtis dissimilarities based on rarefied prokaryotic OTUs standardized using Wisconsin method

	grouping	Basin (3 categories)		Depth (3 categories)		
Dissimilarity: bray		ANOSIM statistic R:	-0.04630	ANOSIM statisti	0.38860	
Permutation: free		Significance:	0.60270	Significance:	0.0116	*

Number of permutations: 9999
SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

OTU-based analyses

Bray-Curtis dissimilarities based on rarefied prokaryotic Phyla and Proteobacterial classes

	grouping	Basin (3 categories)		Depth (3 categories)		
Dissimilarity: bray		ANOSIM statistic R:	0.02778	ANOSIM statisti	0.28110	
Permutation: free		Significance:	0.38190	Significance:	0.0316	*

Number of permutations: 9999
SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Phyla-based analyses

Baikal and other studies - PERMANOVA on bacterial phyla + proteobacterial classes

Bray-Curtis dissimilarities based on bacterial Phyla and proteobacterial classes abundances ~ Salinity (4 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
Salinity	3	2.6482	0.88273	18.97400	0.53237	0.0001	***
residuals	50	2.3262	0.04652		0.46763		
total	53	4.9744			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on bacterial Phyla and proteobacterial classes abundances ~ sequencing methodology (5 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
Seq Meth	4	2.0987	0.52468	8.94030	0.42191	0.0001	***
residuals	49	2.8757	0.05869		0.57809		
total	53	4.9744			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on bacterial Phyla and proteobacterial classes abundances ~ Study (8 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
Study	7	3.6190	0.51701	17.54700	0.72753	0.0001	***
residuals	46	1.3554	0.02946		0.27247		
total	53	4.9744			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on bacterial Phyla and proteobacterial classes abundances ~ Salinity * Sequencing methodology

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
Salinity	3	2.6482	0.88230	29.95960	0.53237	0.0001	***
Seq Meth	4	0.97080	0.24271	8.23740	0.19517	0.0001	***
residuals	46	1.3554	0.02946		0.27247		
total	53	4.9744			1.0000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on bacterial Phyla and proteobacterial classes abundances ~ Salinity * Study

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
Salinity	3	2.6482	0.88230	29.95960	0.53237	0.0001	***
Study	4	0.97080	0.24271	8.23740	0.19517	0.0001	***
residuals	46	1.3554	0.02946		0.27247		
total	53	4.9744			1.0000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Eukaryotes - PERMANOVA (without BK22S samples)

Bray-Curtis dissimilarities based on rarefied eukaryotic OTUs standardized using Wisconsin method ~ basin (3 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
basin	2	0.9737	0.48683	1.02460	0.18547	0.26030	
residuals	9	4.2761	0.47513		0.81453		
total	11	5.2498			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied eukaryotic OTUs standardized using Wisconsin method ~ depth (3 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
depth	2	1.0204	0.51019	1.08570	0.19437	0.04570	*
residuals	9	4.2294	0.46993		0.80563		
total	11	5.2498			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied eukaryotic phyla ~ basin (3 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
basin	2	0.2435	0.12176	1.70730	0.27504	0.16530	
residuals	9	0.6419	0.07132		0.72496		
total	11	0.8854			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied eukaryotic phyla ~ depth (3 categories)

Number of permutations: 9999
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
depth	2	0.2162	0.10810	1.45370	0.24417	0.22170	
residuals	9	0.6692	0.07436		0.75583		
total	11	0.8854			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Eukaryotes - ANOSIM (without BK22S)

Bray-Curtis dissimilarities based on rarefied eukaryotic OTUs standardized using Wisconsin method

	grouping	Basin (3 categories)		Depth (3 categories)		
Dissimilarity: bray		ANOSIM statistic R:	0.06713	ANOSIM statistic R:	0.30350	
Permutation: free		Significance:	0.26110	Significance:	0.0311	*

Number of permutations: 9999
SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied eukaryotic phyla

	grouping	Basin (3 categories)		Depth (3 categories)		
Dissimilarity: bray		ANOSIM statistic R:	0.11570	ANOSIM statistic R:	0.02800	
Permutation: free		Significance:	0.16610	Significance:	0.37580	

Number of permutations: 9999
SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

PK+EK OTUs - PERMANOVA (without BK22S)

Bray-Curtis dissimilarities based on rarefied prokaryotic and eukaryotic OTUs standardized using Wisconsin method ~ basi

Number of permutations: 9999 (3 categories)
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
basin	2	0.77730	0.38863	0.98191	0.17912	0.49010	
residuals	9	3.56210	0.39579		0.82088		
total	11	4.33940			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Bray-Curtis dissimilarities based on rarefied prokaryotic and eukaryotic OTUs standardized using Wisconsin method ~ dept

Number of permutations: 9999 (3 categories)
Terms added sequentially (first to last)

	Df	SumOfSqs	MeanSqs	F.Model	R2	Pr(>F)	signi
depth	2	1.00170	0.50085	1.35050	0.23084	0.01190	*
residuals	9	3.33770	0.37085		0.76916		
total	11	4.33940			1.00000		

SIGNIF CODES: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1

Supplementary Table 9. Metadata of sediment samples used for comparison with Lake Baikal sediments

Sample name	Ecosystem type	Sediment setting	salinity (‰)	depth (m)	Sequence type_primers	Reference
Chen_E1	lake	freshwater	0.00	10	miseq_515F_907R	Chen et al. 2016
Chen_E2	lake	freshwater	0.00	18	miseq_515F_907R	Chen et al. 2016
Chen_E3	lake	freshwater	0.00	23	miseq_515F_907R	Chen et al. 2016
Chen_E4	lake	freshwater	0.00	15	miseq_515F_907R	Chen et al. 2016
Chen_E5	lake	freshwater	0.00	8	miseq_515F_907R	Chen et al. 2016
Wan_Summer_5	lake	freshwater	0.00	2	miseq_515F_806R	Wan et al. 2017
Wan_Summer_8	lake	freshwater	0.00	0.5	miseq_515F_806R	Wan et al. 2017
Wan_Summer_11	lake	freshwater	0.00	0.5	miseq_515F_806R	Wan et al. 2017
Wan_Summer_15	lake	freshwater	0.00	2	miseq_515F_806R	Wan et al. 2017
Wan_Summer_18	lake	freshwater	0.00	0.5	miseq_515F_806R	Wan et al. 2017
Wan_Summer_19	lake	freshwater	0.00	0.5	miseq_515F_806R	Wan et al. 2017
Fang_Site1	lake	brackish	2.50	42	miseq_341F_805R	Fang et al. 2015
Fang_Site2	lake	brackish	2.50	60	miseq_341F_805R	Fang et al. 2015
Fang_Site4	lake	brackish	2.50	89	miseq_341F_805R	Fang et al. 2015
Fang_Site5	lake	brackish	2.50	63	miseq_341F_805R	Fang et al. 2015
Fang_Site6	lake	brackish	2.50	47	miseq_341F_805R	Fang et al. 2015
Ji_averageSed	lake	freshwater	0.00	2	miseq_515F_907R	Ji et al. 2019
Liu_NYS1	sea	marine	32.00	57	pyro_344F_915R	Liu et al. 2015
Liu_NYS2	sea	marine	31.70	51	pyro_344F_915R	Liu et al. 2015
Liu_NYS3	sea	marine	32.00	63.5	pyro_344F_915R	Liu et al. 2015
Mahmoudi_Station1	sea	brackish	11.40	600	miseq_515F_806R	Mahmoudi et al. 2015
Mahmoudi_Station2	sea	brackish	11.30	205	miseq_515F_806R	Mahmoudi et al. 2015
Mahmoudi_Station3	sea	brackish	11.30	141	miseq_515F_806R	Mahmoudi et al. 2015
Ye_A2	sea	marine	33.42	34.3	miseq_515F_907R	Ye et al. 2016
Ye_A3	sea	marine	33.16	37.4	miseq_515F_907R	Ye et al. 2016
Ye_B2	sea	marine	31.84	41.1	miseq_515F_907R	Ye et al. 2016
Ye_A5	sea	marine	34.34	47.3	miseq_515F_907R	Ye et al. 2016
Ye_M2	sea	marine	31.42	26	miseq_515F_907R	Ye et al. 2016
Ye_A8	sea	marine	34.52	118	miseq_515F_907R	Ye et al. 2016
Ye_M7	sea	marine	32.93	66	miseq_515F_907R	Ye et al. 2016
Ye_N6	sea	marine	32.43	63.4	miseq_515F_907R	Ye et al. 2016
Ye_C3	sea	marine	34.37	38	miseq_515F_907R	Ye et al. 2016
Ye_D2	sea	marine	34.38	46	miseq_515F_907R	Ye et al. 2016
Ye_E2	sea	marine	34.37	58	miseq_515F_907R	Ye et al. 2016
Ye_F3	sea	marine	33.92	55	miseq_515F_907R	Ye et al. 2016
Gugliandolo_L1	lake	freshwater	0.05	5.5	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_L2	lake	freshwater	0.05	0.5	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_L8	lake	freshwater	0.02	4.5	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_L11	lake	freshwater	0.03	2.5	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_L15	lake	freshwater	0.07	7.8	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_L5	lake	freshwater	0.20	0.5	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_L6	lake	freshwater	0.10	0.5	pyro_341F_805R	Gugliandolo et al. 2016
Gugliandolo_S1	estuary	freshwater	0.21	0.3	pyro_341F_805R	Gugliandolo et al. 2016
Zeng_NEC5	ocean	marine	32.57	62	pyro_8F_533R	Zeng et al. 2017
Zeng_DBS1	ocean	marine	34.61	2420	pyro_8F_533R	Zeng et al. 2017
Zeng_BJ4	ocean	marine	NA	350	pyro_8F_533R	Zeng et al. 2017
Zeng_BJ36	ocean	marine	NA	28	pyro_8F_533R	Zeng et al. 2017

Supplementary Table 10. Bacterial phylum-level matrix of sediment samples from studies used for comparison with Lake Baikal sediments.

	Chen_E1	Chen_E2	Chen_E3	Chen_E4	Chen_E5	Wan_Summer_S	Wan_Summer_8	Wan_Summer_11	Wan_Summer_15	Wan_Summer_18	Wan_Summer_19	Fang_Site1	Fang_Site2	Fang_Site4	Fang_Site5	Fang_Site6	Ji_average Sed	Liu_NYS1	Liu_NYS2	Liu_NYS3	Mahmoudi_Station1	Mahmoudi_Station2	Mahmoudi_Station3	Ye_A2	Ye_A3	Ye_B2	Ye_A5	Ye_M2	Ye_A8	Ye_M7	Ye_N6	Ye_C3	Ye_D2	Ye_E2	Ye_F3	Gugliandolo_L1	Gugliandolo_L2	Gugliandolo_L8	Gugliandolo_L11	Gugliandolo_L15	Gugliandolo_L5	Gugliandolo_L6	Gugliandolo_S1	Zeng_NIECS	Zeng_DBS1	Zeng_B14	Zeng_B186
Acidobacteria	9.1	7.8	8.3	9.1	9.1	5.1	8.08	7.58	6.19	3.76	9.13	4.47	6.32	4.05	5.49	4.71	8.4	8	5.2	2.4	4	7	3	9.3	5	5.5	7.4	6.1	3.8	8.6	3.9	7	7.1	7.9	8.6	9.85	8.27	5.99	8.31	8.07	13.19	1.98	0.02	5	8	1.5	1
Actinobacteria	1.3	1.7	1.7	1.7	1.3	0.5	0.61	0.52	0.43	0.31	0.48	11.53	10.36	8.27	7.2	8.18	2.8	9.2	7.6	6	3	4	2	3.7	4.1	2	2	2.3	0.7	2.1	1.4	1.1	0.8	0.8	0.7	21.34	21.6	22.28	8.55	22.55	12.74	25.77	4.64	3	6	1.5	1
Armatimonadetes	0.9	0.9	0.9	0.3	0.9	NA	NA	NA	NA	NA	NA	0.17	0.55	0.98	0.46	0.45	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.06	0.03	0.06	0.06	0.03	0.01	0.14	NA	NA	NA	NA	NA
Bacterial Candidate Phyla	2.6	2.6	2.6	2.6	3	0.4	0.83	0.33	0.23	1.32	0.2	0.55	0.74	0.84	1.85	1.34	NA	NA	NA	NA	7	2	0.5	0.5	0.2	1.4	0.4	0.2	9.2	0.2	0.1	0.2	0.4	0.4	1	0.03	0.01	0.01	0.01	0.05	0.05	0.12	0.04	NA	NA	NA	NA
Chloroflexi	10.4	13.5	13	11.7	13	5.18	4.05	4.35	4.25	2.25	2.66	1.68	1.72	1.71	2.52	1.76	8.3	5.2	3.2	4	9	6	2	6.1	3.7	6.4	3	3.3	11.9	2.6	1.5	3.9	4.5	3.7	5.1	0.51	0.68	0.29	0.99	0.21	3.88	0.91	0.03	3	8	1	4
CPR	NA	NA	NA	NA	NA	0.7	0.65	1.13	1.05	0.55	0.76	0.58	0.76	0.67	1.54	0.7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.49	1.6	4.6	3.97	1.27	1.4	1.6	1.24	NA	NA	NA	NA
Cyanobacteria	0.5	0.5	0.5	0.5	0.5	0.48	0.67	0.95	2.35	0.34	0.7	0.15	0.27	0.13	0.56	0.19	5	NA	NA	NA	NA	NA	NA	0.7	1.3	0.7	0.9	1.3	0.1	0.5	0.5	1.9	1.4	0.2	0.1	4.16	7.87	2.46	10.2	9.33	0.66	0.31	0.03	NA	NA	NA	NA
Elusimicrobia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.19	0.12	0.16	0.3	0.18	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
FCB	15.8	14	13.2	16.2	15.7	12.16	12.71	15.98	15.89	18.43	17.03	15.81	12.63	15.11	16.44	15.02	10.1	10	12	14	7	11	2.25	14.7	20.6	12.3	23.2	21.3	3.5	19.9	39.9	17.6	16.3	10.8	7	10.57	11.02	11.81	15.22	10.55	12.68	10.95	38.06	14.5	12.5	15	17
Firmicutes	0.5	1.3	1.3	0.9	0.9	1.19	2.34	1.23	1.76	1.78	0.83	2	2.94	2.8	2.63	2.86	3.6	5.6	5.2	4.8	NA	NA	NA	0.4	1.2	0.7	3.3	0.1	0.5	0.1	1.2	0.9	0.4	0.1	0.1	0.45	0.81	0.39	0.79	0.31	1.8	0.26	3.36	NA	NA	NA	NA
Nitrospinae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nitrospirae	2.6	2.3	2.6	2	2.6	18.3	8.95	8.38	9.22	8.34	10.31	2.39	2.75	3.83	3.86	3.71	2.3	0.8	0.8	0.8	2	1	0.25	0.4	0.2	1.9	0.5	0.4	2.5	1	0.2	1.2	1.2	2.5	4.8	0.49	0.31	0.29	0.36	0.26	0.36	0.13	0	0	0	0	0
Other Bacteria	17.8	19.2	19.6	13	11.2	1.58	1.31	2.18	1.94	2.38	2.02	0.08	0.12	0.09	0.21	0.08	10.3	6.8	5.2	6.8	22	14	10	1.6	1.7	2	2.6	1.1	3.1	2	1.1	1.6	2	1.8	2.1	NA	NA	NA	NA	NA	NA	NA	NA	4.5	5	2.5	5
Proteobacteria	23.7	20.7	20.5	25	25.5	40.79	45.68	41.29	40.36	46.52	42.15	46.16	46.72	45.99	43.02	45.1	46.1	43.2	49.6	47.2	30	47	77	45.4	48.5	49.3	45.5	50.9	41.3	49.7	41.2	56.5	54.1	63.1	57.5	24.13	23.43	22	21.46	21.51	18.43	15.61	40.03	65	52.5	74.5	69
Alphaproteobacteria	1.6	2.1	1.4	2.9	1.4	0.67	1.4	1.68	1.37	0.85	2.43	NA	NA	NA	NA	NA	NA	6.3	7.9	4.1	2	3	8.2	4.8	4.6	2.5	4.3	5.6	3.4	6.4	4.7	3.5	2.8	3.7	3.8	7.73	7.89	8	4.59	9.84	4.09	3.03	2.15	2.3	11.5	30.8	18.5
Betaproteobacteria	6.3	4.5	4.8	6.5	7.6	12.59	11.27	12.85	12.84	12.29	8.99	NA	NA	NA	NA	NA	NA	2.9	2.5	2.7	1	2	7.5	0.2	0.1	0.1	0.3	0.1	0.1	1	0.4	0.3	0.4	0.3	0.3	10.64	7.4	8.04	8.06	7.06	7.11	8.36	2.15	0.1	0.1	0.1	0.1
Gammaproteobacteria	4.6	3.8	3.6	4.7	3.7	5.44	8.47	9.76	8.31	6.75	11.17	NA	NA	NA	NA	NA	NA	18.4	21.8	17.4	5	23.5	58	18.5	21.4	15.3	22	23.3	12.7	24.1	17.9	24.2	22.8	20.8	19.3	1.76	3.08	2.4	2.43	2.69	1.97	2.34	26.27	2.7	15.4	12.3	9.2
Deltaproteobacteria	11.2	10.3	10.7	10.9	12.8	16.37	17.73	14.76	14.94	17.75	18.31	NA	NA	NA	NA	NA	NA	15.6	17.4	23	22	18.5	3.3	21.7	22	31.2	18.4	21.8	25	18.1	18.1	28.2	27.8	37.9	34	3.02	3.93	2.49	5.23	1.91	4.56	1.05	4.89	32.3	23.8	7.7	12.3
Other Proteobacteria	0	0	0	0	0	5.69	6.81	2.24	2.9	8.88	1.25	NA	NA	NA	NA	NA	NA	0	0	0	0	0	0	0.2	0.4	0.2	0.5	0.1	0.1	0.1	0.1	0.3	0.3	0.4	0.1	0.98	1.13	1.07	1.15	0.01	0.7	0.83	4.57	27.6	1.7	23.6	28.9
PVC	11.8	13.3	13.6	14	13.7	10.37	12.23	12.91	13.29	10.16	11.15	4.15	8.08	4.06	5.28	5.39	3.1	11.2	11.2	14	16	8	3	15.2	11.5	7.8	9.1	10.7	6.5	12.1	7.2	6.2	6.9	6.6	7.3	6.61	5.33	8.3	6.59	5.02	3.84	4.46	2.11	3	6	3	2
SBR1093	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Spirochaetae	3	2.2	2.2	3	2.6	0.86	0.43	0.51	0.57	1.42	0.55	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.3	0.5	1.3	1.1	0.4	2.5	0.1	0.9	0.3	0.6	0.6	0.7	0.03	0.04	0.02	0.13	0.03	0.01	0.01	0.05	NA	NA	NA	NA
Unclassified Bacteria	NA	NA	NA	NA	NA	2.39	1.46	2.66	2.47	2.44	2.03	10.09	5.92	11.31	8.64	10.33	NA	NA	NA	NA	NA	NA	NA	1.7	1.5	8.7	1	1.9	14.4	1.1	0.9	1.6	4.3	1.5	5	20.28	19	21.5	23.36	20.81	30.95	37.75	10.39	2	2	1	1
WS1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
WS2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

CHAPTER

4

COMPARATIVE METAGENOMIC OF BAIKAL LAKE SEDIMENTS

4.1 Context and objectives

In this chapter, I will present my study of lake Baikal upper layer sediments using a metagenomic approach, one I detailed in Section 1.2.2.3. The first goal of this study was to verify the metabarcoding conclusions stated in the last chapter (Chapter 3) about the stability of the communities. To note that for this study, we sequenced only the 7 deepest sediments of the metabarcoding study. The second goal of this study is to gain insight into the metabolic capabilities of these ecosystems and identifying the key players. The following manuscript has been recently written and is a first draft which needs to be improved.

4.2 Manuscript draft version

Metabolic traits of Lake Baikal sediment microbial communities inferred from comparative metagenomics

Guillaume Reboul[1], , David Moreira[1], Paola Bertolino[1], Nataliia V. Annenkova[2] and Purificación López-Garcia[1]

[1] *Ecologie Systématique Evolution, Centre National de la Recherche Scientifique - CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France*

[2] *Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia*

*For correspondence: puri.lopez@u-psud.fr

Running title: Metagenomics of lake Baikal sediments

Keywords: anaerobic metabolism, Thaumarchaeota, S cycling, N cycle, MAGs

Manuscript in preparation

Abstract

Lake Baikal is the deepest and the most voluminous lake on earth. While the planktonic communities of this unique freshwater reservoir have been relatively well studied, its sediments remain poorly and partially explored. Here we apply shotgun metagenomics on the upper layers (0–3cm) of sediment collected across a latitudinal and vertical gradient in Lake Baikal to unravel the metabolic potential of resident microbial communities. We identified the phylogenetic diversity of sediment microbial communities based on both raw reads and universal single copy genes from assembled contigs. Based on KEGG Orthologs (KOs), we were able to identify metabolic pathways for carbon fixation and nutrient cycling potentially active in Baikal sediments as well as their respective dominant players. Archaea, especially Thaumarchaeota (TACK superphylum) and their associated metabolic pathways were well represented. Proteobacterial classes were also abundant and involved in different metabolic processes. We also recovered metagenome assembled genomes (MAGs) from our comprehensive sediment sample collection. Closely related MAGs were shared across sampling sites, notably those of Thaumarchaeota.

Introduction

Lake Baikal, UNESCO world Heritage Site since 1966 (UNESCO website) is the oldest, deepest and largest (by volume) freshwater lake such that, in a sense, it constitutes a freshwater sea. Lake Baikal is divided into three basins: the Northern, Central and Southern basins of the lake are delimited by the Academician ridge and the Selenga river (major inflow river) delta respectively (Mats and Perepelova, 2011; Touchart, 2012). The lake is approximately 30 million years old and has a maximum depth of ca. 1,650 meters (ca. 750 m average) and a water volume of ca. 23 000 km^3 corresponding to ca. 20% of the Earth's unfrozen water (Sherstyankin *et al.*, 2006). Coastal downwelling and deep-water ventilation occur in the lake as a consequence of the winter freezing period (which still takes place despite recent climate change (Hampton *et al.*, 2008)) and strong wind regime (Schmid *et al.*, 2008; Moore *et al.*, 2009). Most of the lake water column remains constantly cold (ca. 4°C), ultra-oligotrophic and oxygen-rich (dissolved oxygen levels often above 10 $mg.L^{-1}$)(Schmid *et al.*, 2008; Moore *et al.*, 2009; Shimaraev and Domysheva, 2012; Troitskaya *et al.*, 2015). Low water temperature and high pressure due to the depth are known to facilitate solid phase methane; indeed, lake Baikal is the only lake in which methane hydrates have been discovered and studied (De Batist *et al.*, 2002; Granin *et al.*, 2019).

Being an ancient lake, Baikal harbors many endemic fauna and flora species and thus has been a valuable source for biodiversity and ecological studies in the past century. 1455 out of the 2595 animal species described are endemic(Yu Sherbakov, 1999; Moore *et al.*, 2009), including some undergoing adaptive radiations (Stelbrink *et al.*, 2015). Planktonic microbial species have also been relatively well studied using classical observation and cultural approaches (Maksimova and Maksimov, 1972, 1975; Maksimova, Maksimov and Vorbieva, 1974; Maksimov *et al.*, 2002; Bel'kova *et al.*, 2003) as well as with early molecular approaches such as clone libraries (Glockner *et al.*, 2000). Recent advances in sequencing techniques have allowed deeper sequencing of gene markers and allowed wider comparisons of microbial communities associated to this ecosystem. The diversity of pelagic, surface water, near bottom water, sediment, methane-seep associated communities of bacteria, archaea and protists have since been unveiled using marker gene approaches (rRNA genes and metabolic marker genes) (Bashenkhaeva *et al.*, 2015; Mikhailov *et al.*, 2015, 2019; Kurilkina *et al.*, 2016; Yi *et al.*, 2017; Zakharenko *et al.*, 2019; Annenkova, Giner and Logares, 2020) (Reboul et al, submitted). These approaches have also been used to carry out analyses centered in taxonomic level centered or host associated communities analyses have been carried out especially on diatoms (Zakharova *et al.*, 2013; Roberts *et al.*, 2018) and dinoflagellates (Annenkova, Lavrov and Belikov, 2011) or bacteria associated with metazoan like endemic fish (Denikina *et al.*, 2016) species and sponges (Kulakova *et al.*, 2018; Belikov *et al.*, 2019). Only very recently have total DNA metagenomic approaches been applied to study Baikal ecosystems. In particular, these techniques have been used to investigate sponge viruses (Butina *et al.*, 2020), sub-ice communities (Cabello-Yeves *et al.*, 2017), deep-water communities (Cabello-Yeves *et al.*, 2019) as well as virioplankton in coastal (Butina *et al.*, 2019) or pelagic (Potapov *et al.*, 2019; Coutinho *et al.*, 2020) waters.

In general, microbial communities associated to sediments have been less explored than planktonic ones, especially in lacustrine environments. In lake Baikal, notable exceptions are areas with methane seep and oil-bearing fluids, which have been sampled and studied to

highlight specific prokaryotic lineages associated with these particular environments that show similarity to seabed communities (Kadnikov *et al.*, 2012; Bukin *et al.*, 2016; Lomakina *et al.*, 2018). Except when targeting the methane seep, studies of Baikal lake sediments have been geographically restricted to the Southern basin of the lake as it is more accessible (Yi *et al.*, 2017). Benthic communities, however, can be complex and phylogenetically diverse (Zinger *et al.*, 2011). Despite accounting for only a small portion of earth's living biomass (Kallmeyer *et al.*, 2012), they are of ecological importance for organic matter remineralization and crucial role in carbon storage (Dang and Lovell, 2016; Rastelli *et al.*, 2016; Orsi, 2018). They are also the source, at least in oceans, of divergent archaeal or bacterial lineages (Biddle *et al.*, 2008; Schauer *et al.*, 2011; Rinke *et al.*, 2013; Baker *et al.*, 2015, 2016; Spang *et al.*, 2015; Solden, Lloyd and Wrighton, 2016; Orsi, 2018; Seitz *et al.*, 2019).

To get knowledge about the diversity and community structure of Lake Baikal sediment communities, we recently investigated the prokaryotic and eukaryotic components of several sediment samples collected at different depths and basins using metabarcoding analyses. Our results suggested that microbial community structure was rather stable across depth and latitude (Reboul et al, submitted). Here, we have carried out metagenomic analyses of sediments occupying the greatest depths in the three Baikal basins as well as the transition areas between basins. Our study reveals which are the major metabolic pathways operating at the sediment ecosystem level. In addition, the reconstruction of metagenome-assembled genomes allows to predict the metabolic potential of dominant community members.

Materials & methods

Lake Baikal sampling details

Lake Baikal samples were collected during a joint French-Russian research cruise carried out between June 28th and July 7th, 2017. Seven sediment push cores were retrieved from each basin at its deepest point and also at the transitional zone between basins (from 320 to 1450 meters) along a North-South transect. Two sediment samples were taken from the northern and the Southern basin and three from the central basin (**Figure 1**; Supplementary Table 1). The physicochemical parameters of the lake waters close to the bottom were measured when possible *in situ* with a CTD probe. For this study, we collected the upper layer of the sediment core (ca. 0-3 cmbsf; Supplementary Table 1), including the water interface. (Supplementary Table 1). Sediment samples from the chosen sites were fixed in ethanol (>80% v/v) and stored at -20°C until processed.

DNA purification and sequencing

After ethanol elimination, ~2 g sediment samples were left to rehydrate for 2-4h at 4°C and DNA was extracted using the Power Soil™ DNA purification kit (Qiagen, Hilden, Germany). Sequencing was performed using paired-end (2x125 bp) Illumina HiSeq by Eurofins Genomics (Ebersberg, Germany). Raw sequences have been deposited in GenBank under the BioProject number: XXXXXXXX.

Metagenomic data cleaning, assembly, ORF and function predictions

Metagenomic raw data were quality checked using FASTQC v0.11.5 (Andrews, 2010) and then trimmed of the low-quality bases at their extremities using trimmomatic v0.38 (Bolger, Lohse and Usadel, 2014) (options *PE -phred33 -summary -baseout ILLUMINACLIP:Trimmomatic-0.38/adapters/TruSeq3-PE-2.fa:2:30:10 HEADCROP:10 SLIDINGWINDOW:4:30 MINLEN:35*). Trimmomatic outputs fastq files containing respectively the R1 only, R2 only and paired-end reads which passed the criteria of trimming. After validation of the trimming by another FASTQC run, we selected for further treatment and analyses only the paired-end reads fastq files outputted by trimmomatic (around 80% of initial raw reads for each sample, see Supplementary Table 2). Trimmed reads were then assembled into contigs using metaSpades v3.13.0 (Nurk *et al.*, 2017) options *-k 21,25,31,35,41,45,51,55*. Coverage was computed for the contigs using the following pipeline: first BWA v0.7.17 (Li and Durbin, 2009) was used to index the contigs (options *index -a bwtsw*) and then mapping the reads used for the assembly using the *mem* algorithm which created SAM files. Samtools v1.9 (Li *et al.*, 2009) was then used to create and index the BAM files using the *sort* and the *index* options, respectively, and finally, BEDTOOLS v2.28.0 (Quinlan and Hall, 2010) was used on the indexed BAM files with the *genomecov* option to compute the coverage per contigs; awk was used to rearrange the output into an easy-to-use format. Open Reading Frames (ORF) on contigs were predicted using prokka tool v1.12 (Seemann, 2014) in the meta mode and following options: *--metagenome –rfam –addmrna –addgenes –mincontiglen 200*. KOFAMSCAN v1.2 (Aramaki *et al.*, 2019) was used to assign KEGG Orthologs (KOs) (Kanehisa *et al.*, 2016) to all predicted proteins by search against the new database of profile hidden Markov models (KOfam) of KEGG (Aramaki *et al.*, 2019).

Phylogenetic assignment of metagenomic sequences

We assigned the trimmed reads and the predicted proteins to different taxa using Kaiju v1.7.1 (Menzel and Krogh, 2015) options *OPTIONS* and *OPTIONS2* respectively. The taxonomic classification was manually adapted to add, remove or clarify some taxonomic levels or group names; details of this can be found in Supplementary Table 3.

Raw reads The affiliated kaiju taxonomy of the reads was parsed using in-house awk scripts to sum up the number of reads per taxonomic levels.

USiCGs Universal Single-Copy Genes (USiCGs) according to the list on which is based the MUSiCC software (Manor and Borenstein, 2015) were retrieved along with the average coverage of the contig they belong to and their taxonomy affiliated by kaiju.

Taxonomy was adapted as previously described and analyses were performed using R v3.6.3 'Holding the Windsock' (R Core Team, 2017) scripts. Barcharts and heatmaps were generated using ggplot2 v3.3.0 (Wickham, 2016) and ComplexHeatmap v2.3.4 (Gu, Eils and Schlesner, 2016) R packages, respectively. Bootstrap values were computed using the R package pvclust v2.2.0 (Suzuki and Shimodaira, 2006). The seed was set to 123456789 for replicability and 10000 iterations were performed to compute the p-values. Clustering results and the corresponding p-values were manually added on the heatmap dendrograms using InkScape (Inkscape Project, no date). To assess clustering quality, we used Approximately Unbiased (AU) p-values computed by multiscale bootstrap resampling which attempts to correct for possible sampling biases, although we also report the uncorrected Bootstrap Probability (BP) p-values computed by normal bootstrap resampling (https://github.com/shimo-lab/pvclust). All reported values are AU values unless specified otherwise.

Prediction of metabolic pathways

Coverage of manually selected KOs corresponding to key KOs within metabolic pathways of interest were fetched from the KOFAMSCAN prediction on contigs as well as the USiCGs KO list (Supplementary Table 4). Then these key KOs were normalized following the MUSiCC software v1.0.3 (Manor and Borenstein, 2015) procedure using the available python scripts and the options *-v -c use_generic -n -perf*. These normalized values were then grouped by metabolic pathways and the median of their normalized coverage was used to draw the heatmaps and compare metabolic pathway abundances within sample and between samples. When not all diagnostic KOs were predicted for a specific pathway, then the absent ones were not considered to compute the median for the respective metabolic pathway in the corresponding sample.

MAG binning and metrics computation

Metagenome-assembled genomes (MAGs) were binned from the assembled contigs using the METABAT software v2.13-29-g2e72973 (Kang *et al.*, 2015, 2019). To do so, we first mapped all the raw reads of all the samples to the contigs assembled from their own set of raw reads (as previously described, using a BWA v0.7.17 (Li and Durbin, 2009) (options *index -a bwtsw*) to create a SAM file and using Samtools v1.9 (Li *et al.*, 2009) to transform the SAM file into a BAM file). Thus, we obtained a coverage profile of every contig by all the samples as BAM files. We then summarized the different coverage profiles for each contig using metabat script jgi_summarize_bam_contig_depths. Only contigs greater than 1500 bp were selected for the binning process with metabat2 (Kang *et al.*, 2015, 2019) option *-m 1500* and 500 bins were retrieved (Supplementary Table 2). In order to estimate the completeness, contamination, taxonomic affiliation and other metrics on our constituted MAGs, we applied checkM tool

v1.0.11 (Parks *et al.*, 2015) using the following options: *lineage_wf –ali –nt -x fa*. GC content of the MAGs was obtained locally by computing the average GC content of all constituted contigs of the respective MAGs. We only considered for the further analyses 303 MAGs which were selected if their completeness was >30% and contamination <10% based on checkM output (Supplementary Table 2). MAG families were defined as groups of MAGs satisfying two conditions: (1) they are clustered together using UPGMA on the coverage rate profile across samples, and (2) they are placed in the same group on the phylogenomic tree.

Results and discussion

We analyzed metagenomes from deep sediments sampled across Lake Baikal basins (Northern, Central and Southern (Mats and Perepelova, 2011; Touchart, 2012)) (Figure 1). Samples were collected in summer, but the bottom sediments were surrounded by cold water of approximately 4°C (Supplementary Table 1). From each sample, we were able to sequence between 40 (BK30S) and 110 (BK25S) million reads, ~80% of which were high quality paired-end reads (hereafter raw reads (RR)) which were merged and selected for further analyses (Supplementary Table 2).

Overview of sediment community structure

The vast majority of the raw reads were prokaryotic, with 90%–95% bacteria and ~5–10% archaea (Raw Reads on **SuppFig2**). Eukaryotic reads only accounted for XX-XX%. In terms of the inferred number of organisms, bacterial species accounted for 80–85%, archaea for 10–18% and 2–4% of unclassified USiCGs (USiCGs on **SuppFig2**). For diversity analyses at the phyla level (or classes for Proteobacteria), we used the affiliation of raw reads (RR) (**SuppFig1**) as a proxy for relative abundance in terms of total DNA corresponding to each taxon, and predicted USiCGs (**Figure 2**), as a proxy for genomes (see Mat&Met for details).

In general, the sediment communities were dominated, within archaea, by TACK members (mostly Thaumarchaeota) (2-6% RR – 6-15% USiCGs), and, within bacteria, by FCB (8-15% RR – 6-16% USiCGs), Chloroflexi (5-12% RR – 4-21% USiCGs), Acidobacteria (3-13% RR – 3-22% USiCGs) and Proteobacteria phyla members, notably Delta and Betaproteobacteria"' (**Figure2; Figure 2**). We investigated whether, superimposed on this general conserved structure, specific shared tendencies could be identified between the samples based on their phyla relative abundances. We therefore applied hierarchical clustering (see M&M for full procedure) on both RR and USiCGs (**Figure2.B & Figure 2.B**), which yielded 2 (out of 5) significant (p-value >=0.9) clusters using RR data and 4 (out of 5) significant clusters using USiCGs data. The only cluster found significant with both approaches was the one containing the two deep samples from the Southern basin (BK26S and BK30S); this is the only clear evidence for local differences among the 7 samples based on depth or geographical position. The aggregation of the sample BK21S with the cluster formed by samples BK26S and BK30S to make C1, one of the two main clusters, was supported using the USiCGs dataset but failed to reach significance in the RR dataset. These three samples had a lower relative abundance of Chloroflexi, FCB and Deltaproteobacteria and a higher relative abundance of Acidobacteria, Betaproteobacteria and Rokubacteria than the samples of the other major cluster C2 (with the exception of BK16S which had a high abundance of Acidobacteria) (Supp.Fig1B). This confirms previous studies based on metabarcoding diversity data suggesting that communities were rather stable across sediment samples (Reboul et al, submitted)

The differences between analyses using total DNA (RR analyses) and estimated species (USiCGs analyses) relative abundances is most likely due due to the bias of genome sizes in the abundant phyla (**SuppFig3**). For example, Thaumarchaeota (accounting for 90% and 80% TACK RR and USiCGs affiliations, respectively) are known to have relatively small genomes on average (Walker *et al.*, 2010; Stieglmeier, Alves and Schleper, 2014) and here, they account for 2 to 6 percent of the total sequenced DNA amount (**SuppFig1**) they have been predicted for 6 to 15% of the total predicted organisms thriving in these sediments (**Figure 2**). Actinobacteria, FCB, Chloroflexi and PVC have larger genome sizes than the average prokaryotic size in our samples, while genome sizes of Thaumarchaeota, Nitrospirae, Rokubacteria and

Betaproteobacteria were smaller (**SuppFig3**). The prokaryotic diversity in lake Baikal sediments is relatively uniform, with no particular phyla dominating the samples. Within this balance, relatively better represented phyla were the bacterial phyla Acidobacteria, Chloroflexi, Proteobacteria (mainly Delta- and Betaproteobacteria), FCB and PVC groups as well as a high proportion of archaea mostly assigned to TACK (**Figure2 and Figure 2**). This is in agreement with our recently reported results using metabarcoding (Reboul et al, submitted).

Carbon fixation in Baikal sediments

Primary producers are at the base of the trophic chain. We investigated the ability to fix carbon and the phylogenetic diversity of potential autotrophs in Baikal sediments. To do so, we looked for genes belonging to the most common carbon fixation pathways and the predicted associated taxa. We first retrieved key genes from the six main carbon fixation pathways known to date (reviewed in (Berg, 2011; Fuchs, 2011; Hügler and Sievert, 2011), see details in M&M): the Calvin-Benson reductive pentose phosphate cycle (Calvin Cycle), the reductive citric acid cycle (rTCA Arnon-Buchanan cycle), the reductive acetyl-coa (Wood-Ljungdahl) pathway, the 3-hydroxypropionate (Fuchs-Holo) bi-cycle (HP-bicycle), the 3-hydroxypropionate/4-hydroxybutyrate cycle (HP-HB cycle) and the dicarboxylate/4-hydroxybutyrate cycle (DC-HB cycle). Each of these cycles is known to be carried out by specific taxa in which they were originally discovered. However, recent studies have broadened this view by attributing cycles, albeit with some modifications, to previously unrelated taxonomic groups (Könneke *et al.*, 2014; Mall *et al.*, 2018; Nunoura *et al.*, 2018). In the majority of our samples, there was a clear dominance of the HP-HB cycle (**Figure3.A**, Supplementary Table 5), usually carried out by Thaumarchaeota (only TACK superphylum lineage recovered) (**Figure3.B**) over other carbon fixation pathways. Although discovered in Crenarchaeota, also member of the TACK superphyla, recent studies show evidences for a modified HP-HB cycle in Thaumarchaeota (Berg *et al.*, 2007; Könneke *et al.*, 2014).This phylum was found important for carbon fixation based CO_2 in oceans (Herndl *et al.*, 2005; Offre, Spang and Schleper, 2013). The remaining two samples (BK05S and BK25S), were dominated by the Wood-Ljungdahl pathway (WLP). WLP was represented in four major taxonomicgroups: PVC, Chloroflexi, Deltaproteobacteria and Euryarchaeota (**Figure3.B**). Indeed, Deltaproteobacteria, Euryarchaeota and PVC superphyla members include autotrophic organisms known to fix carbon via the WLP (Ragsdale and Pierce, 2008; Berg, 2011; Hügler and Sievert, 2011). Moreover, around 30 Chloroflexia high-quality Metagenome-Assembled genomes (MAGs) were recently recovered from marine sub-seafloor and most of them harbor the WLP for C fixation (Fincker *et al.*, 2020). The Calvin cycle was detected to a higher degree in samples dominated by the HP–HB cycle than samples dominated by the WLP cycle (**Figure3.A**). This cycle is usually linked to photosynthesis and therefore its relatively small presence might be attributed to organism residuals which have sunk from the euphotic zone to the lake bed. In our samples, Calvin cycle genes were essentially found in Beta- and Gammaproteobacteria as well as the NC10 phylum (**Figure3.B**). This presence of NC10 phylum is intriguing as those organisms are not known to be photosynthetic. Recent studies also identified Calvin pathway genes in NC10 members, apparently linked to nitrite-dependent anaerobic methane oxidation (Rasigraf *et al.*, 2014).. Finally, the Arnon-Buchanan cycle (reverse Krebs cycle or rTCA) was only marginally represented in our sediments. We found this pathway carried out by Nitrospirae phylum, which is in line with previous observations (Lücker *et al.*, 2010). Our analysis also detected

genes associated with the HP-bicycle represented in Deltaproteobacteria and Rokubacteria. TheHP-bicycle pathway is typically associated with photosynthetic green-nonsulfur bacteria (Chloroflexi (Zarzycki *et al.*, 2009)). However, we detect only one out of the seven HP-bicycle enzymes in Baikal sediments (table???). Therefore, the product of this gene might be involved in other metabolic pathways and/or in a modified carbon fixation pathway. Assuming that carbon fixation can be extended to the respectively associated lineages, this would imply that autotrophic organisms in Baikal sediments might account for up to 35% (in BK30S) to 60% (BK26S) of their respective communities (Supplementary Table 5) with the HP-HB cycle or the WLP as dominant pathways. Although these values seem to suggest that C fixation is important in these sediments, some of these genes might be involved in anaplerotic central metabolism reactions and a more quantitative measure of C fixation implemented.

Energy metabolism pathways

We next investigated the main energy-harvesting strategies in Baikal sediment-associated microbial communities. We looked for the presence of six major metabolic pathways: nitrification (ammonia oxidation to nitrite and nitrate), dissimilatory nitrate reduction (DNR – nitrate reduction to nitrite), sulfur oxidation (SOX system), denitrification (nitrate and nitrite reduction), dissimilatory sulfate reduction (DSR), and hydrogen-metabolism ([NiFe]-hydrogenase; NFH). In each sample, 45%–65% of the population was assigned to one of these energy metabolism pathways (Supplementary Table 5). Overall, the communities employ metabolic pathways driven by nitrogen- and sulfur-derived compounds; we found no evidence for methanogenesis, which is a likely carried out in deeper sediment layers (cite ref on redox zonation in sediments). The relatively high abundance of sulfate reducers indicates that the analyzed layers are rich in sulfate, which is known toinhibit methanogenesis (Kristjansson, Schönheit and Thauer, 1982) (**Figure4,** Supplementary Table 5).

We found evidence for Nitrification, DNR, SOX system, Denitrification, and DSR in all the samples but [NiFe]-hydrogenase (NFH) was only predicted in five out of the seven samples. For all but one (BK16S) samples, Nitrification was the major energy pathway (**Figure4.A,** Supplementary Table 5). The most frequent taxonomic assignations of the key KOs for nitrification were Thaumarchaeota (~50%) then unclassified archaea (~25%) and Nitrospirae (~10%) (**Figure4.B**). Nitrification in archaea has been known for some time and is typically associated to Thaumarchaeota (Walker *et al.*, 2010; Offre, Spang and Schleper, 2013). DNR was detected in Chloroflexi, Nitrospirae, Betaproteobacteria, Deltaproteobacteria and Rokubacteria in all sediments. Most of the SOX system enzymes were phylogenetically assigned to Betaproteobacteria, as previously observed (Friedrich *et al.*, 2005). Denitrification was detected in TACK superphyla (Thaumarchaeota) (~50%) and, to a lesser extent, in Deltaproteobacteria, unclassified bacteria and Nitrospirae. A potential involvement of Thaumarchaeota in denitrification was already proposed but subsequently questioned because of the expression of the genes in aerobic conditions. Thus it is mainly hypothesized that Thaumarchaeota have lost their denitrification capabilities (Kozlowski *et al.*, 2016; Kimble *et al.*, 2018). Another possibility as suggested for marine habitats by Pachiadaki *et al.*, 2017, is a possible symbiosis of the Thaumarchaeota and nitrite-oxidizing bacteria such as Nitrospirae which will contribute to dark carbon fixation. In contrast to the other pathways, the taxonomic representation of DSR was less consistent across samples, with Betaproteobacteria the only group detected across all samples.

UPGMA clustering of the samples (see M&M for details) divided the samples' metabolic expression profiles, butthe two main groups C1 (higher SOX system and lower NFH detected

levels) and C2 (relatively more homogenized profiles) were not significant (Figure4.A). These non-significant groups mirror the groups detected after clustering the USiCGs taxonomical profiles described above (Figure 2.B). This similarity stems from the samples BK21S, BK26S, and BK30s (C1 in the USiCGs clustering), being more abundant in Betaproteobacteria and Rokubacteria, groups that were majoritarily assigned to the SOX system key KOs in our communities (Figure4.B). In terms of smaller, local clusters, UPGMA clustered the samples from the southern basin (BK26S and BK30S, higher SOX system and lower Denitrification and DNR levels) and the samples from the central basin (BK03S, BK05S and BK25S, lower SOX system and higher Denitrification and DNR levels) but not the values from the northern basin. Overall, Lake Baikal top layer sediments are inhabited by species using mainly nitrogen- and sulfur- based energy metabolism pathways to thrive, and the relative abundances of these pathways vary according to the basin of origin.

Thaumarchaeota metagenome-assembled genomes

A chart integrating our findings for energy metabolism (Figure4) as well as the assimilative pathways (SuppFig4) summarizing the nutrient cycles and their key players is depicted in Figure5. In the above analyses, taxonomy was assessed at the level of phyla (or classes for Proteobacteria), but the diversity of organisms within phyla remains unexplored. Thaumarchaeota was the only phylum detected to employ the HP/HB cycle, the most abundant carbon fixation pathway in our samples. Thaumarchaeota also represented a large share of the nitrification (and denitrification) energy metabolism pathways, also relatively important in our samples, and it was also the phylum dominating the Assimilatory Sulfate Reduction (ASR) pathway. Therefore, given the importance of Thaumarchaeota in the ecosystem of the lake Baikal upper layer sediments, we searched to better assess the phylogenetic diversity of thaumarchaeotal genomes in Baikal sediments. To do so, we reconstructed 304 metagenome-assembled genomes (MAGs) from our datasets (see M&M). In many cases, we observed MAGs initially obtained from different samples displaying the same coverage range profiles across the samples and also grouped together on inferred phylogenomic analyses. We refer to these MAGs as MAG families hereafter. In the case of Thaumarchaeota, we detected three MAG families (LBSSTF1-3; Figure 6) which were all assigned to the family Nitrosopumilaceae. This assignment was further corroborated by the genome sizes (smaller than 2 Mb even for the most complete MAGs within the family) and the GC content (~ 35%) which are in line with the known values for Nitrosopumilaceae (Walker *et al.*, 2010; Stieglmeier, Alves and Schleper, 2014). Overall, the metabolic predictions for Thaumarchaeota are due to multiple species of the family Nitrosopumilaceae, which were well distributed in sediment samples across the latitudinal gradient.

Concluding remarks

In this study we applied a metagenomic approach to explore the diversity and the metabolic potential of the microbial communities associated with sediments in the biggest liquid freshwater reservoir on Earth. This is the first analysis of this type for lake Baikal sediments across a latitudinal N-S transect (**Figure1**). Matagenomes were dominated by prokaryotic sequences with a relatively abundant archaea (ca. XX%) in line with previous metabarcoding results (Reboul et al., submitted). As in the case of metabarcoding analyses, metagenome comparisons showed a relatively stable pattern across the sediments with no evidence for a geographical or depth link (**Figure2**). FCB and Acidobacteria were abundant, being likely involved in the degradation of more or less complex organics. Chloroflexi were abundant and had a role in nitrate reduction and carbon fixation *via* the WLP. As in many ecosystems, Proteobacteria were one of the most abundant phyla; however, its abundance did not exceed 30%, which is unusual for upper layer sediments at the limit with the water body. Among them, Delta- and Beta-proteobacteria appeared important players in the carbon fixation (WLP and Calvin cycle, respectively) and nutrient cycles (denitrification, DNR, and DSR and DSR, ASR, SOX system, respectively). Thaumarchaeota were predicted to play a role in carbon fixation through the HP-HB cycle and the nutrient cycles (50% of the denitrification and nitrification processes) in lake Baikal sediments (**Figure3; Figure4; Figure5**). This was accounted for by the family Nitrosopumilaceae as revealed by the MAGs analysis (**Figure 6**) and notably the phylogenomically inferred two novel Thaumarchaeota genus.

Overall, this study sheds light on the metabolic potential of the microbial communities inhabiting the first sediment layers at the interface with the water column and constitute a basis for understanding their ecology.

Acknowledgements

We thank Luis J. Galindo and Anabel Lopez-Archilla for help during our 2017 limnological cruise, Philippe Deschamps for technical bioinformatic support and Miguel Iniesto for discussions on microbial metabolism. We thank the crew of the R/V G. Titov for their work onboard and the director of the Limnological Institute at Irkusk for logistical assistance. This research was funded by the European Research Council Grants ProtistWorld (322669, PL-G) and PlastEvol (787904, DM) as well as the Russian State grant 0345-2016-0009 (NVA).

Author contributions

PLG, DM and NVA designed the work and organized the limnological cruise. PLG and GR collected sediment samples. PB and GR purified DNA. GR carried out the bioinformatic analysis of metagenomic sequences, statistical analyses. GR wrote an early draft of the manuscript. PLG wrote the final manuscript. All authors read, critically commented and approved the final manuscript

Figure Legends

Figure 1 Bathymetric map of the Baikal lake showing the sampling sites in the different basins following the North-South latitude gradient: Northern Basin (green filled circles), Central Basin (orange squares), Southern Basin (purple triangles).

Figure 2 Diversity of microbial communities in Baikal upper layer sediments as estimated by the predicted USiCGs. (A) Diversity barchart of all sampling locations. Numerical values indicated for communities detected greater than 3%. (B) Diversity presented in matrix form with the corresponding UPGMA clustering (above chart): red: Approximately Unbiased (AU) p-value; green: Bootstrap Probability (BP) p-value.

Figure 3 Dominant carbon fixation pathways in upper layers (0-3 cm) of deep Baikal sediments. (A) Relative abundance of major identified carbon fixation pathway genes in matrix form with the corresponding UPGMA clustering (above chart): red: Approximately Unbiased (AU) p-value; green: Bootstrap Probability (BP) p-value. (B) For each pathway, a barchart depicting the phyla (or classes for Proteobacteria) in which the pathway was detected. Note that the most abundantly detected cycle, the HP-HB cycle, was detected exclusively in TACK; the Wood-Ljungdahl pathway: by Chloroflexi, Deltaproteobacteria, and Euryarchaeota; the Calvin cycle was represented mostly by Betaproteobateria, Gammaproteobacteria, and NC10; the rTCA cycle: by Nitrospirae.

Figure 4 Energy metabolic pathways (A) levels of the tested energy metabolic pathways in matrix form with the corresponding UPGMA clustering (above chart): red: Approximately Unbiased (AU) p-value; green: Bootstrap Probability (BP) p-value. (B) For each pathway, a barchart depicting the phyla (or classes for Proteobacteria) in which the pathway was detected. Note that Nitrification and Denitrification are dominated by TACK, while the SOX system is mostly detected in Betaproteobacteria.

Figure 5 A chart summarizing the detected nutrient cycles and their key players in the ecosystem of surface layer sediments in lake Baikal. Line width corresponds to the levels at which a pathway is detected (per genome).

Figure 6. Thaumarcheota MAG families. (A) Phylogenomic tree of the Thaumarcheota-reconstructed MAG families. All three MAG families were detected under the family Nitrosopumilaceae using the last version of GTDB marker gene sets and reference genomes. Note that LBSSTF1 and LBSSTF2 branched with only one genome from the GTDB and formed new genus clades of the family Nitrosopumilaceae of the Thaumarchaeota phylum, and LBSSTF3 was also part of a non-named clade but along many other genomes. (B) Left: coverage rate of each MAG (rows) in each sample (columns). White rhombuses denote for each MAG the sample from which it was originally reconstructed. UPGMA clustering: red: Approximately Unbiased (AU) p-value; green: Bootstrap Probability (BP) p-value. Note the closely matching coverage profiles of MAGs clustered together. MAGs significantly (AU p-value) clustered together and grouped together in the phylogenomic tree were retained as MAG families. Middle left: size of each reconstructed MAG. Middle right: GC content of each reconstructed MAG. Note the similar GC content across MAGs within the same family. Right: completeness. Note that within each MAG family, MAGs smaller in size show lower completeness, reinforcing

the notion that they are fragments of the genomes of the same species as other MAGs in the MAG family. Supplementary Figure Legends

Supplementary Figure 1 Diversity of microbial communities in Baikal upper layer sediments as estimated by the raw reads (RR). (A) Diversity barchart of all sampling locations. Numerical values indicated for communities detected greater than 3%. (B) Diversity presented in matrix form with the corresponding UPGMA clustering (above chart): red: Approximately Unbiased (AU) p-value; green: Bootstrap Probability (BP) p-value.

Supplementary Figure 2 Kingdom representation in Baikal upper layer sediments using raw reads or predicted USiCGs. Violin plots representing the relative abundances of each kingdom. Note that the vast majority of the detected communities were prokaryotic. Note that as a whole, the total amount of DNA (raw reads) tends to under-represent the relative abundance of Archaea as estimated by UsiCGs due to their smaller genomes, and that the opposite is true for Bacteria.

Supplementary Figure 3 Abundance of microbial communities as estimated by the USiCGs coverage (a proxy for number of individuals, x axis) and the count of USiCGs (a proxy for total amount of DNA, y axis). Symbol shape: sample location. Symbol color: phylum (or class for Proteobacteria). Where possible, ovals highlight all the points for the same phylum or class for illustration purposes. The diagonal line would denote phyla for which the estimated abundance would be the same with both methods (RR and USiCGs), i.e. one with an average genome size for that sample. Note that Actonobacteria, FCB, Chloroflexi, and PVC are located above the diagonal, indicating that their genomes are larger than the average for our samples, while TACK, Nitrospirae, Rokubactera, and Betaproteobacteria are below the diagonal, indicating smaller genomes. Assessing diversity with RR without correcting for USiCGs would therefore tend to overestimate the former communities and underestimate the latter.

Supplementary Figure 4 Assimilatory metabolic pathways in Baikal upper layer sediments. (A) Expression levels of the Assimilatory Nitrate Reduction (ANR) pathway and the Assimilatory Sulfate Reduction (ASR) pathway in matrix form. (B) For each pathway, a barchart depicting the phyla (or classes for Proteobacteria) in which the pathway was detected. Note that ANR was dominated by Rokubacteria, while ASR was mostly represented by TACK.

Bibliography

Andrews, S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Annenkova, N. V., Giner, C. R. and Logares, R. (2020) 'Tracing the Origin of Planktonic Protists in an Ancient Lake', *Microorganisms*. Multidisciplinary Digital Publishing Institute, 8(4), p. 543. doi: 10.3390/microorganisms8040543.

Annenkova, N. V, Lavrov, D. V and Belikov, S. I. (2011) 'Dinoflagellates Associated with Freshwater Sponges from the Ancient Lake Baikal', *Protist*, 162(2), pp. 222–236. doi: 10.1016/j.protis.2010.07.002.

Aramaki, T. *et al.* (2019) 'KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold', *Bioinformatics*. Edited by A. Valencia. doi: 10.1093/bioinformatics/btz859.

Baker, B. J. *et al.* (2015) 'Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria', *Microbiome*. BioMed Central, 3(1), p. 14. doi: 10.1186/s40168-015-0077-6.

Baker, B. J. *et al.* (2016) 'Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea', *Nature Microbiology*. Nature Publishing Group, 1(3), pp. 1–7. doi: 10.1038/nmicrobiol.2016.2.

Bashenkhaeva, M. V *et al.* (2015) 'Sub-Ice Microalgal and Bacterial Communities in Freshwater Lake Baikal, Russia', *Microbial Ecology*. Microb Ecol, 70(3), pp. 751–765. doi: 10.1007/s00248-015-0619-2.

De Batist, M. *et al.* (2002) 'Active hydrate destabilization in Lake Baikal, Siberia?', *Terra Nova*, 14(6), pp. 436–442. doi: 10.1046/j.1365-3121.2002.00449.x.

Bel'kova, N. L. *et al.* (2003) 'Microbial biodiversity in the water of lake baikal', *Microbiology*. Springer, 72(2), pp. 203–213. doi: 10.1023/A:1023224215929.

Belikov, S. *et al.* (2019) 'Diversity and shifts of the bacterial community associated with Baikal sponge mass mortalities', *PLoS ONE*. Edited by B. A. Wilson. Public Library of Science, 14(3), p. e0213926. doi: 10.1371/journal.pone.0213926.

Berg, I. A. *et al.* (2007) 'A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in archaea', *Science*. American Association for the Advancement of Science, 318(5857), pp. 1782–1786. doi: 10.1126/science.1149976.

Berg, I. A. (2011) 'Ecological aspects of the distribution of different autotrophic CO2 fixation pathways', *Applied and Environmental Microbiology*. American Society for Microbiology, pp. 1925–1936. doi: 10.1128/AEM.02473-10.

Biddle, J. F. *et al.* (2008) 'Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment', *Proceedings of the National Academy of Sciences of the United States of America*, 105(30), pp. 10583–10588. doi: 10.1073/pnas.0709942105.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bukin, S. V *et al.* (2016) 'The ability of microbial community of Lake Baikal bottom sediments associated with gas discharge to carry out the transformation of organic matter under thermobaric conditions', *Frontiers in Microbiology*. Frontiers Media SA, 7(MAY), p. 690. doi: 10.3389/fmicb.2016.00690.

Butina, T. V. *et al.* (2020) 'Metavirome datasets from two endemic Baikal sponges Baikalospongia bacillifera', *Data in Brief*. Elsevier, 29, p. 105260. doi: 10.1016/j.dib.2020.105260.

Butina, T. V *et al.* (2019) 'Estimate of the diversity of viral and bacterial assemblage in the

549 coastal water of Lake Baikal', *FEMS Microbiology Letters*. Oxford Academic, 366(9). doi:
550 10.1093/femsle/fnz094.
551 Cabello-Yeves, P. J. *et al.* (2017) 'Genomes of novel microbial lineages assembled from the
552 sub-ice waters of Lake Baikal', *Applied and Environmental Microbiology*. Edited by H. L. Drake.
553 American Society for Microbiology, 84(1), p. AEM.02132-17. doi: 10.1128/AEM.02132-17.
554 Cabello-Yeves, P. J. *et al.* (2019) 'Microbiome of the deep Lake Baikal, a unique oxic
555 bathypelagic habitat', *Limnology and Oceanography*. John Wiley & Sons, Ltd, p. lno.11401.
556 doi: 10.1002/lno.11401.
557 Coutinho, F. H. *et al.* (2020) 'New Viral Biogeochemical Roles Revealed Through
558 Metagenomic Analysis of Lake Baikal', *bioRxiv*. Cold Spring Harbor Laboratory, p.
559 2020.04.02.019802. doi: 10.1101/2020.04.02.019802.
560 Dang, H. and Lovell, C. R. (2016) 'Microbial Surface Colonization and Biofilm Development in
561 Marine Environments', *Microbiology and Molecular Biology Reviews*. American Society for
562 Microbiology, 80(1), pp. 91–138. doi: 10.1128/mmbr.00037-15.
563 Denikina, N. *et al.* (2016) 'Genetic diversity of Diplomonadida in fish of the genus Coregonus
564 from Southeastern Siberia', *Acta Parasitologica*. Springer, 61(2), pp. 299–306. doi:
565 10.1515/ap-2016-0040.
566 Fincker, M. *et al.* (2020) 'Metabolic strategies of marine subseafloor Chloroflexi inferred from
567 genome reconstructions', *Environmental Microbiology*. John Wiley & Sons, Ltd, pp. 1462-
568 2920.15061. doi: 10.1111/1462-2920.15061.
569 Friedrich, C. G. *et al.* (2005) 'Prokaryotic sulfur oxidation', *Current Opinion in Microbiology*.
570 Elsevier Current Trends, pp. 253–259. doi: 10.1016/j.mib.2005.04.005.
571 Fuchs, G. (2011) 'Alternative pathways of carbon dioxide fixation: Insights into the early
572 evolution of life?', *Annual Review of Microbiology*. Annual Reviews, pp. 631–658. doi:
573 10.1146/annurev-micro-090110-102801.
574 Glockner, F. O. *et al.* (2000) 'Comparative 16S rRNA analysis of lake bacterioplankton reveals
575 globally distributed phylogenetic clusters including an abundant group of actinobacteria',
576 *Applied and Environmental Microbiology*. American Society for Microbiology, 66(11), pp.
577 5053–5065. doi: 10.1128/AEM.66.11.5053-5065.2000.
578 Granin, N. G. *et al.* (2019) 'Methane hydrate emergence from Lake Baikal: direct
579 observations, modelling, and hydrate footprints in seasonal ice cover', *Scientific Reports*.
580 Nature Publishing Group, 9(1), p. 19361. doi: 10.1038/s41598-019-55758-8.
581 Gu, Z., Eils, R. and Schlesner, M. (2016) 'Complex heatmaps reveal patterns and correlations
582 in multidimensional genomic data', *Bioinformatics*. Oxford Academic, 32(18), pp. 2847–2849.
583 doi: 10.1093/bioinformatics/btw313.
584 Hampton, S. E. *et al.* (2008) 'Sixty years of environmental change in the world's largest
585 freshwater lake - Lake Baikal, Siberia', *Global Change Biology*. John Wiley & Sons, Ltd, 14(8),
586 pp. 1947–1958. doi: 10.1111/j.1365-2486.2008.01616.x.
587 Herndl, G. J. *et al.* (2005) 'Contribution of Archaea to total prokaryotic production in the deep
588 atlantic ocean', *Applied and Environmental Microbiology*. American Society for Microbiology,
589 71(5), pp. 2303–2309. doi: 10.1128/AEM.71.5.2303-2309.2005.
590 Hügler, M. and Sievert, S. M. (2011) 'Beyond the Calvin cycle: Autotrophic carbon fixation in
591 the ocean', *Annual Review of Marine Science*. Annual Reviews, 3(1), pp. 261–289. doi:
592 10.1146/annurev-marine-120709-142712.
593 Inkscape Project (no date) 'Inkscape'. Available at: https://inkscape.org.
594 Kadnikov, V. V *et al.* (2012) 'Microbial community structure in methane hydrate-bearing
595 sediments of freshwater Lake Baikal', *FEMS Microbiology Ecology*, 79(2), pp. 348–358. doi:

10.1111/j.1574-6941.2011.01221.x.
Kallmeyer, J. *et al.* (2012) 'Global distribution of microbial abundance and biomass in subseafloor sediment', *Proceedings of the National Academy of Sciences of the United States of America*, 109(40), pp. 16213–16216. doi: 10.1073/pnas.1203849109.
Kanehisa, M. *et al.* (2016) 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research*. Oxford Academic, 44(D1), pp. D457–D462. doi: 10.1093/nar/gkv1070.
Kang, D. D. *et al.* (2015) 'MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities', *PeerJ*. PeerJ Inc., 2015(8), p. e1165. doi: 10.7717/peerj.1165.
Kang, D. D. *et al.* (2019) 'MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies', *PeerJ*. PeerJ Inc., 2019(7), p. e7359. doi: 10.7717/peerj.7359.
Kimble, J. C. *et al.* (2018) 'A potential central role of Thaumarchaeota in N-Cycling in a semi-arid environment, Fort Stanton Cave, Snowy River passage, New Mexico, USA', *FEMS microbiology ecology*. Oxford University Press, 94(11). doi: 10.1093/femsec/fiy173.
Könneke, M. *et al.* (2014) 'Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(22), pp. 8239–8244. doi: 10.1073/pnas.1402028111.
Kozlowski, J. A. *et al.* (2016) 'Pathways and key intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria and Thaumarchaeota', *ISME Journal*. Nature Publishing Group, 10(8), pp. 1836–1845. doi: 10.1038/ismej.2016.2.
Kristjansson, J. K., Schönheit, P. and Thauer, R. K. (1982) 'Different Ks values for hydrogen of methanogenic bacteria and sulfate reducing bacteria: An explanation for the apparent inhibition of methanogenesis by sulfate', *Archives of Microbiology*. Springer, 131(3), pp. 278–282. doi: 10.1007/BF00405893.
Kulakova, N. V *et al.* (2018) 'Brown Rot Syndrome and Changes in the Bacterial Community of the Baikal Sponge Lubomirskia baicalensis', *Microbial Ecology*. Microb Ecol, 75(4), pp. 1024–1034. doi: 10.1007/s00248-017-1097-5.
Kurilkina, M. I. *et al.* (2016) 'Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing', *FEMS Microbiology Ecology*, 92(7). doi: 10.1093/femsec/fiw094.
Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.
Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*. Oxford Academic, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.
Lomakina, A. V. *et al.* (2018) 'Diversity of Archaea in Bottom Sediments of the Discharge Areas With Oil- and Gas-Bearing Fluids in Lake Baikal', *Geomicrobiology Journal*, 35(1), pp. 50–63. doi: 10.1080/01490451.2017.1315195.
Lücker, S. *et al.* (2010) 'A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 107(30), pp. 13479–13484. doi: 10.1073/pnas.1003860107.
Maksimov, V. V. *et al.* (2002) 'The classification and the monitoring of the state of mouth riverine and lacustrine ecosystems in Lake Baikal based on the composition of local microbiocenoses and their activity', *Microbiology*. Springer, 71(5), pp. 595–600. doi:

10.1023/A:1020571122456.
Maksimova, E. A. and Maksimov, V. N. (1972) 'Vertikal'noe raspredelenie mikrobial'nogo planktona v techenie 1969 g. v iuzhnom Baĭkale.', *Mikrobiologiya*, 41(5), pp. 896–902. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4643939 (Accessed: 6 September 2020).
Maksimova, E. A. and Maksimov, V. N. (1975) 'Vremia generatsii chistykh kul'tur geterotrofnykh mikroorganizmov Baĭkala', *Mikrobiologiya*, 44(6), pp. 1098–1102. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1240574 (Accessed: 6 September 2020).
Maksimova, E. A., Maksimov, V. N. and Vorbieva, E. I. (1974) 'Content of heterotrophic microorganisms in water of Lake Baikal (Russian)', *Mikrobiologiya*, 43(1), pp. 124–128. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4407453 (Accessed: 6 September 2020).
Mall, A. *et al.* (2018) 'Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium', *Science*. American Association for the Advancement of Science, 359(6375), pp. 563–567. doi: 10.1126/science.aao2410.
Manor, O. and Borenstein, E. (2015) 'MUSiCC: A marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome', *Genome Biology*. BioMed Central, 16(1), p. 53. doi: 10.1186/s13059-015-0610-8.
Mats, V. D. and Perepelova, T. I. (2011) 'A new perspective on evolution of the Baikal Rift', *Geoscience Frontiers*. Elsevier, 2(3), pp. 349–365. doi: 10.1016/j.gsf.2011.06.002.
Menzel, P. and Krogh, A. (2015) 'Kaiju : Fast and sensitive taxonomic classification for metagenomics', *bioRxiv*. Nature Publishing Group, 7, pp. 1–9. doi: 10.1101/031229.
Mikhailov, I. S. *et al.* (2015) 'Similarity of structure of taxonomic bacterial communities in the photic layer of Lake Baikal's three basins differing in spring phytoplankton composition and abundance', *Doklady Biochemistry and Biophysics*. Dokl Biochem Biophys, 465(1), pp. 413–419. doi: 10.1134/S1607672915060198.
Mikhailov, I. S. *et al.* (2019) 'Co-occurrence Networks Among Bacteria and Microbial Eukaryotes of Lake Baikal During a Spring Phytoplankton Bloom', *Microbial Ecology*, 77(1), pp. 96–109. doi: 10.1007/s00248-018-1212-2.
Moore, M. V. *et al.* (2009) 'Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia', *BioScience*, 59(5), pp. 405–417. doi: 10.1525/bio.2009.59.5.8.
Nunoura, T. *et al.* (2018) 'A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile', *Science*. American Association for the Advancement of Science, 359(6375), pp. 559–563. doi: 10.1126/science.aao3407.
Nurk, S. *et al.* (2017) 'MetaSPAdes: A new versatile metagenomic assembler', *Genome Research*. Cold Spring Harbor Laboratory Press, 27(5), pp. 824–834. doi: 10.1101/gr.213959.116.
Offre, P., Spang, A. and Schleper, C. (2013) 'Archaea in biogeochemical cycles', *Annual Review of Microbiology*. Annu Rev Microbiol, 67, pp. 437–457. doi: 10.1146/annurev-micro-092412-155614.
Orsi, W. D. (2018) 'Ecology and evolution of seafloor and subseafloor microbial communities', *Nature Reviews Microbiology*. Nature Publishing Group, pp. 671–683. doi: 10.1038/s41579-018-0046-8.
Pachiadaki, M. G. *et al.* (2017) 'Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation.', *Science (New York, N.Y.)*. Science, 358(6366), pp. 1046–1051. doi: 10.1126/science.aan8260.
Parks, D. H. *et al.* (2015) 'CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

690 Potapov, S. A. *et al.* (2019) 'Metagenomic analysis of virioplankton from the pelagic zone of
691 lake baikal', *Viruses*. Multidisciplinary Digital Publishing Institute, 11(11), p. 991. doi:
692 10.3390/v11110991.
693 Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing
694 genomic features', *Bioinformatics*. Oxford Academic, 26(6), pp. 841–842. doi:
695 10.1093/bioinformatics/btq033.
696 R Core Team (2017) 'R: A Language and Environment for Statistical Computing'. Vienna,
697 Austria. Available at: https://www.r-project.org/.
698 Ragsdale, S. W. and Pierce, E. (2008) 'Acetogenesis and the Wood-Ljungdahl pathway of CO2
699 fixation', *Biochimica et Biophysica Acta - Proteins and Proteomics*. Elsevier, pp. 1873–1898.
700 doi: 10.1016/j.bbapap.2008.08.012.
701 Rasigraf, O. *et al.* (2014) 'Autotrophic carbon dioxide fixation via the Calvin-Benson-Bassham
702 cycle by the denitrifying methanotroph "Candidatus Methylomirabilis oxyfera"', *Applied and
703 Environmental Microbiology*. American Society for Microbiology (ASM), 80(8), pp. 2451–
704 2460. doi: 10.1128/AEM.04199-13.
705 Rastelli, E. *et al.* (2016) 'CO2 leakage from carbon dioxide capture and storage (CCS) systems
706 affects organic matter cycling in surface marine sediments', *Marine Environmental Research*.
707 Elsevier, 122, pp. 158–168. doi: 10.1016/j.marenvres.2016.10.007.
708 Rinke, C. *et al.* (2013) 'Insights into the phylogeny and coding potential of microbial dark
709 matter', *Nature*. Nature Publishing Group, 499(7459), pp. 431–437. doi:
710 10.1038/nature12352.
711 Roberts, S. L. *et al.* (2018) 'Diatom evidence of 20th century ecosystem change in Lake Baikal,
712 Siberia', *PLoS ONE*. Edited by B. Yang. Public Library of Science, 13(12), p. e0208765. doi:
713 10.1371/journal.pone.0208765.
714 Schauer, R. *et al.* (2011) 'Bacterial sulfur cycling shapes microbial communities in surface
715 sediments of an ultramafic hydrothermal vent field', *Environmental Microbiology*, 13(10), pp.
716 2633–2648. doi: 10.1111/j.1462-2920.2011.02530.x.
717 Schmid, M. *et al.* (2008) 'Lake Baikal deepwater renewal mystery solved', *Geophysical
718 Research Letters*. John Wiley & Sons, Ltd, 35(9), p. L09605. doi: 10.1029/2008GL033223.
719 Seemann, T. (2014) 'Prokka: Rapid prokaryotic genome annotation', *Bioinformatics*. Oxford
720 Academic, 30(14), pp. 2068–2069. doi: 10.1093/bioinformatics/btu153.
721 Seitz, K. W. *et al.* (2019) 'Asgard archaea capable of anaerobic hydrocarbon cycling', *Nature
722 Communications*. Nature Publishing Group, 10(1), p. 1822. doi: 10.1038/s41467-019-09364-x.
723 Sherstyankin, P. P. *et al.* (2006) 'Computer-based bathymetric map of Lake Baikal', *Doklady
724 Earth Sciences*, 408(4), pp. 564–569. doi: 10.1134/S1028334X06040131.
725 Shimaraev, M. N. and Domysheva, V. M. (2012) 'Trends in Hydrological and Hydrochemical
726 Processes in Lake Baikal under Conditions of Modern Climate Change', in *Climatic Change
727 and Global Warming of Inland Waters: Impacts and Mitigation for Ecosystems and Societies*.
728 Chichester, UK: John Wiley & Sons, Ltd, pp. 43–66. doi: 10.1002/9781118470596.ch3.
729 Solden, L., Lloyd, K. and Wrighton, K. (2016) 'The bright side of microbial dark matter: Lessons
730 learned from the uncultivated majority', *Current Opinion in Microbiology*. Elsevier Current
731 Trends, pp. 217–226. doi: 10.1016/j.mib.2016.04.020.
732 Spang, A. *et al.* (2015) 'Complex archaea that bridge the gap between prokaryotes and
733 eukaryotes', *Nature*. Nature Publishing Group, 521(7551), pp. 173–179. doi:
734 10.1038/nature14447.
735 Stelbrink, B. *et al.* (2015) 'Conquest of the deep, old and cold: An exceptional limpet radiation
736 in Lake Baikal', *Biology Letters*. The Royal Society, 11(7), p. 20150321. doi:

10.1098/rsbl.2015.0321.
Stieglmeier, M., Alves, R. J. E. and Schleper, C. (2014) 'The phylum thaumarchaeota', in *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 347–362. doi: 10.1007/978-3-642-38954-2_338.
Suzuki, R. and Shimodaira, H. (2006) 'Pvclust: An R package for assessing the uncertainty in hierarchical clustering', *Bioinformatics*. Oxford Academic, 22(12), pp. 1540–1542. doi: 10.1093/bioinformatics/btl117.
Touchart, L. (2012) 'Baikal, lake', in *Encyclopedia of Earth Sciences Series*, pp. 83–91. doi: 10.1007/978-1-4020-4410-6_50.
Troitskaya, E. *et al.* (2015) 'Cyclonic circulation and upwelling in Lake Baikal', *Aquatic Sciences*. Springer, 77(2), pp. 171–182. doi: 10.1007/s00027-014-0361-8.
Walker, C. B. *et al.* (2010) 'Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 107(19), pp. 8818–8823. doi: 10.1073/pnas.0913533107.
Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.
Yi, Z. *et al.* (2017) 'High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations', *FEMS microbiology ecology*. Narnia, 93(8). doi: 10.1093/femsec/fix073.
Yu Sherbakov, D. (1999) 'Molecular phylogenetic studies on the origin of biodiversity in Lake Baikal', *Trends in Ecology and Evolution*. Trends Ecol Evol, 14(3), pp. 92–95. doi: 10.1016/S0169-5347(98)01543-2.
Zakharenko, A. S. *et al.* (2019) 'Bacterial Communities in Areas of Oil and Methane Seeps in Pelagic of Lake Baikal', *Microbial Ecology*. Springer, 78(2), pp. 269–285. doi: 10.1007/s00248-018-1299-5.
Zakharova, Y. R. *et al.* (2013) 'The Structure of Microbial Community and Degradation of Diatoms in the Deep Near-Bottom Layer of Lake Baikal', *PLoS ONE*, 8(4). doi: 10.1371/journal.pone.0059977.
Zarzycki, J. *et al.* (2009) 'Identifying the missing steps of the autotrophic 3-hydroxypropionate CO2 fixation cycle in Chloroflexus aurantiacus', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 106(50), pp. 21317–21322. doi: 10.1073/pnas.0908356106.
Zinger, L. *et al.* (2011) 'Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems', *PLoS ONE*. Public Library of Science, 6(9), p. e24570. doi: 10.1371/journal.pone.0024570.

Figure 1

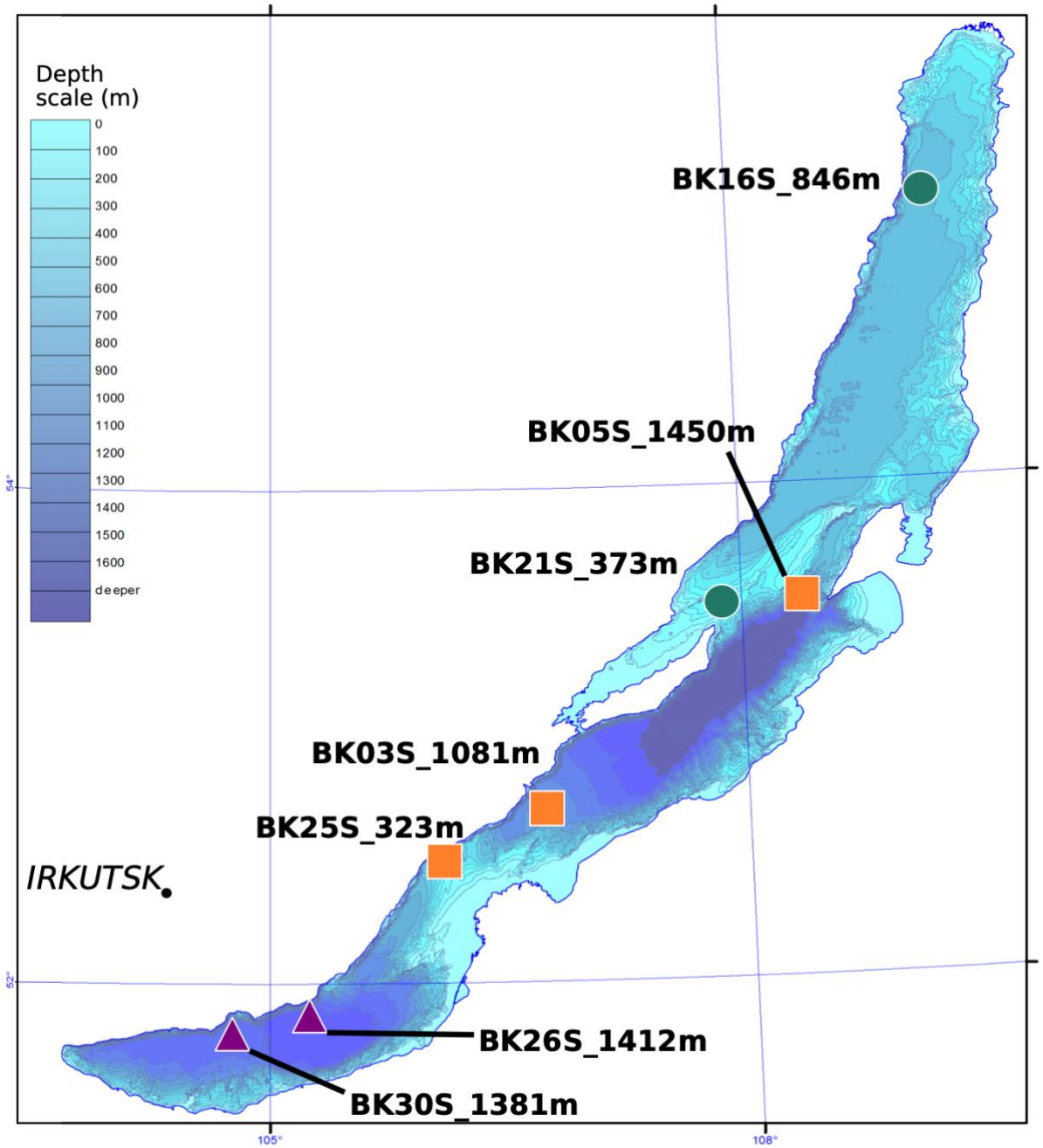
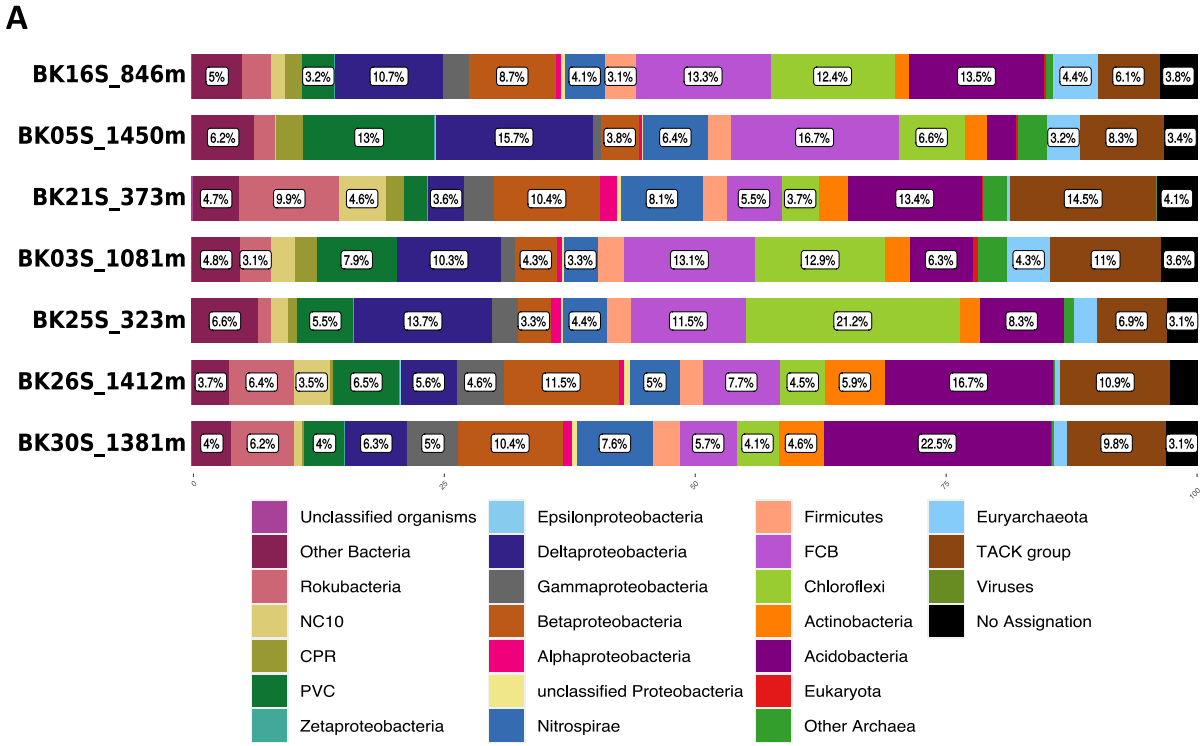

Depth
scale (m)
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
deeper
BK16S_846m
BK05S_1450m
BK21S_373m
BK03S_1081m
BK25S_323m
IRKUTSK
BK26S_1412m
BK30S_1381m
54°
52°
105°
108°

Figure 2


A
BK16S_846m
BK05S_1450m
BK21S_373m
BK03S_1081m
BK25S_323m
BK26S_1412m
BK30S_1381m
Unclassified organisms
Other Bacteria
Rokubacteria
NC10
CPR
PVC
Zetaproteobacteria
Epsilonproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Betaproteobacteria
Alphaproteobacteria
unclassified Proteobacteria
Nitrospirae
Firmicutes
FCB
Chloroflexi
Actinobacteria
Acidobacteria
Eukaryota
Other Archaea
Euryarchaeota
TACK group
Viruses
No Assignation

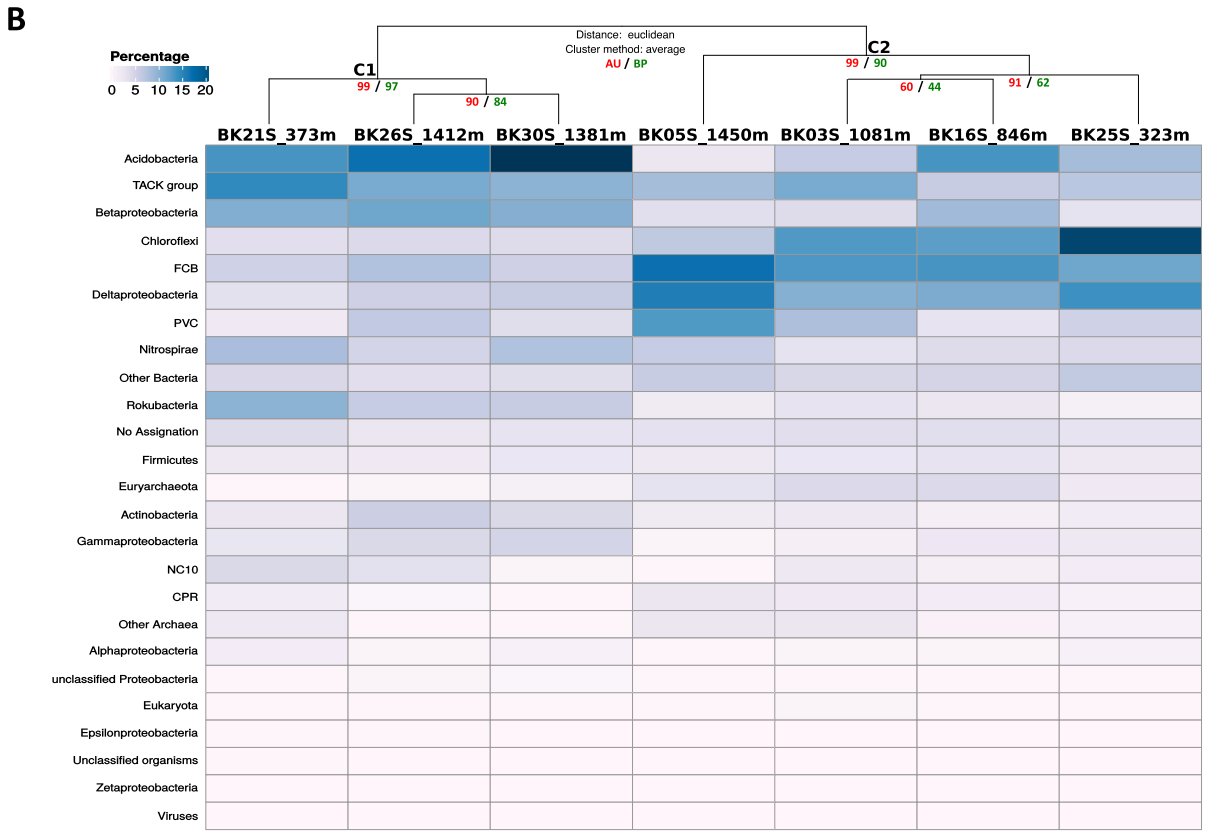

B
Percentage
0 5 10 15 20
Distance: euclidean
Cluster method: average
AU / BP
C1
99 / 97
90 / 84
C2
99 / 90
60 / 44
91 / 62
BK21S_373m
BK26S_1412m
BK30S_1381m
BK05S_1450m
BK03S_1081m
BK16S_846m
BK25S_323m
Acidobacteria
TACK group
Betaproteobacteria
Chloroflexi
FCB
Deltaproteobacteria
PVC
Nitrospirae
Other Bacteria
Rokubacteria
No Assignation
Firmicutes
Euryarchaeota
Actinobacteria
Gammaproteobacteria
NC10
CPR
Other Archaea
Alphaproteobacteria
unclassified Proteobacteria
Eukaryota
Epsilonproteobacteria
Unclassified organisms
Zetaproteobacteria
Viruses

Figure 3

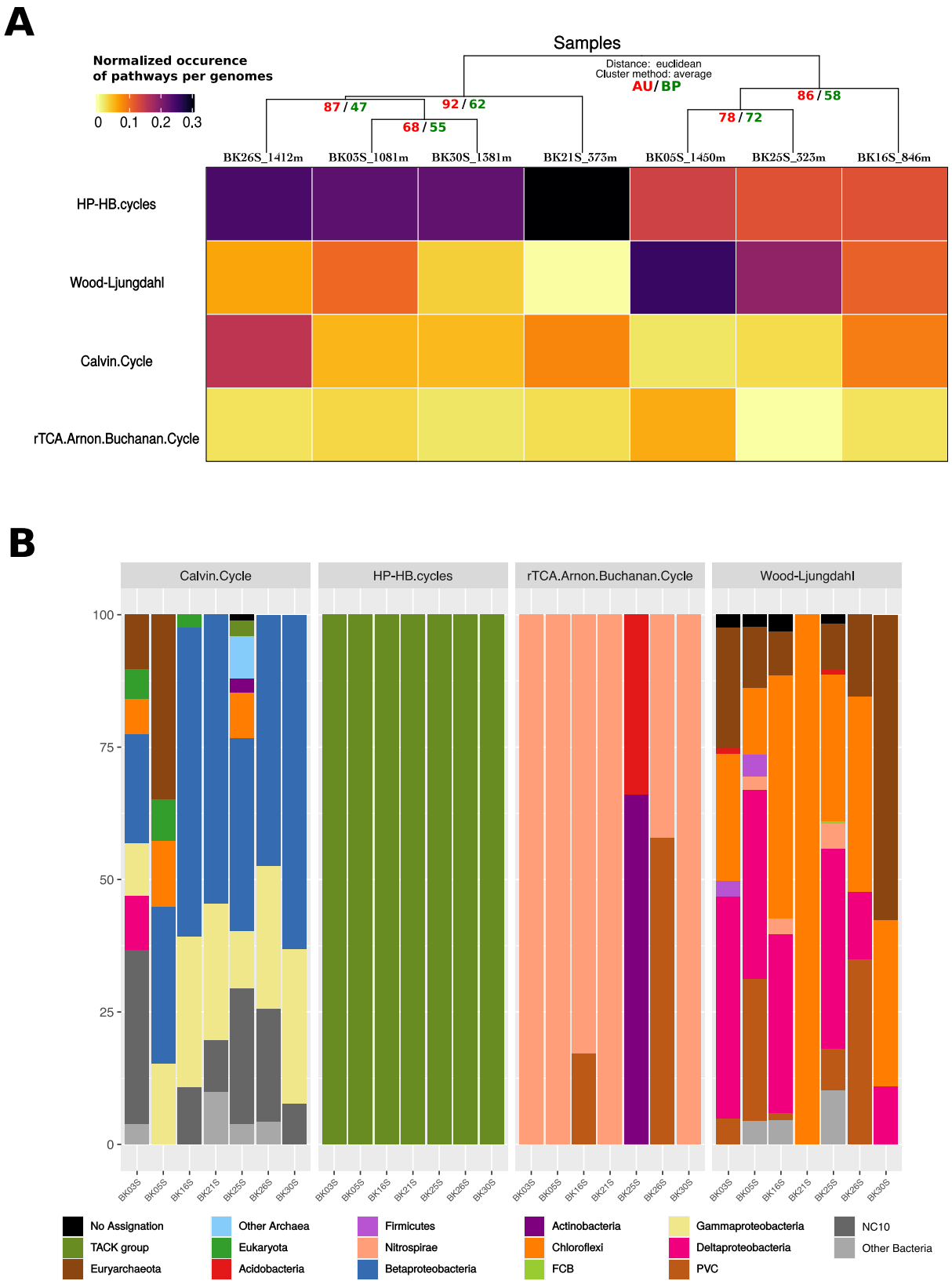
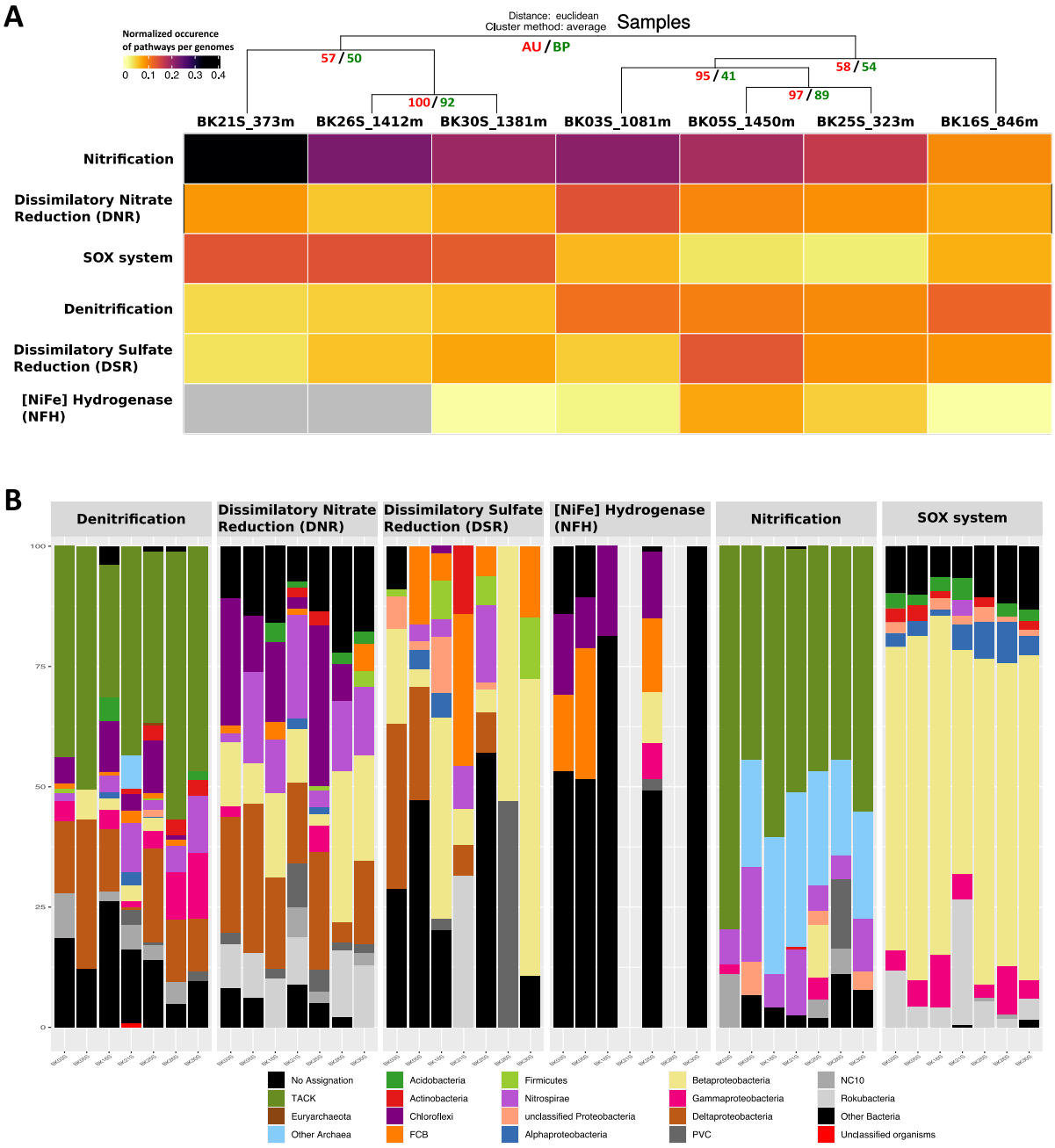

A
Samples
Normalized occurence of pathways per genomes
0 0.1 0.2 0.3
Distance: euclidean
Cluster method: average
AU / BP
87 / 47
92 / 62
68 / 55
86 / 58
78 / 72
BK26S_1412m
BK03S_1081m
BK30S_1381m
BK21S_373m
BK05S_1450m
BK25S_323m
BK16S_846m
HP-HB.cycles
Wood-Ljungdahl
Calvin.Cycle
rTCA.Arnon.Buchanan.Cycle
B
Calvin.Cycle
HP-HB.cycles
rTCA.Arnon.Buchanan.Cycle
Wood-Ljungdahl
100
75
50
25
0
BK03S
BK05S
BK16S
BK21S
BK25S
BK26S
BK30S
No Assignation
TACK group
Euryarchaeota
Other Archaea
Eukaryota
Acidobacteria
Firmicutes
Nitrospirae
Betaproteobacteria
Actinobacteria
Chloroflexi
FCB
Gammaproteobacteria
Deltaproteobacteria
PVC
NC10
Other Bacteria

Figure 4


A
Normalized occurence of pathways per genomes
0 0.1 0.2 0.3 0.4
Distance: euclidean
Cluster method: average
Samples
AU / BP
57 / 50
100 / 92
95 / 41
97 / 89
58 / 54
BK21S_373m
BK26S_1412m
BK30S_1381m
BK03S_1081m
BK05S_1450m
BK25S_323m
BK16S_846m
Nitrification
Dissimilatory Nitrate Reduction (DNR)
SOX system
Denitrification
Dissimilatory Sulfate Reduction (DSR)
[NiFe] Hydrogenase (NFH)
B
Denitrification
Dissimilatory Nitrate Reduction (DNR)
Dissimilatory Sulfate Reduction (DSR)
[NiFe] Hydrogenase (NFH)
Nitrification
SOX system
100
75
50
25
0
No Assignation
TACK
Euryarchaeota
Other Archaea
Acidobacteria
Actinobacteria
Chloroflexi
FCB
Firmicutes
Nitrospirae
unclassified Proteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
PVC
NC10
Rokubacteria
Other Bacteria
Unclassified organisms

803 *Figure 5*
804
805

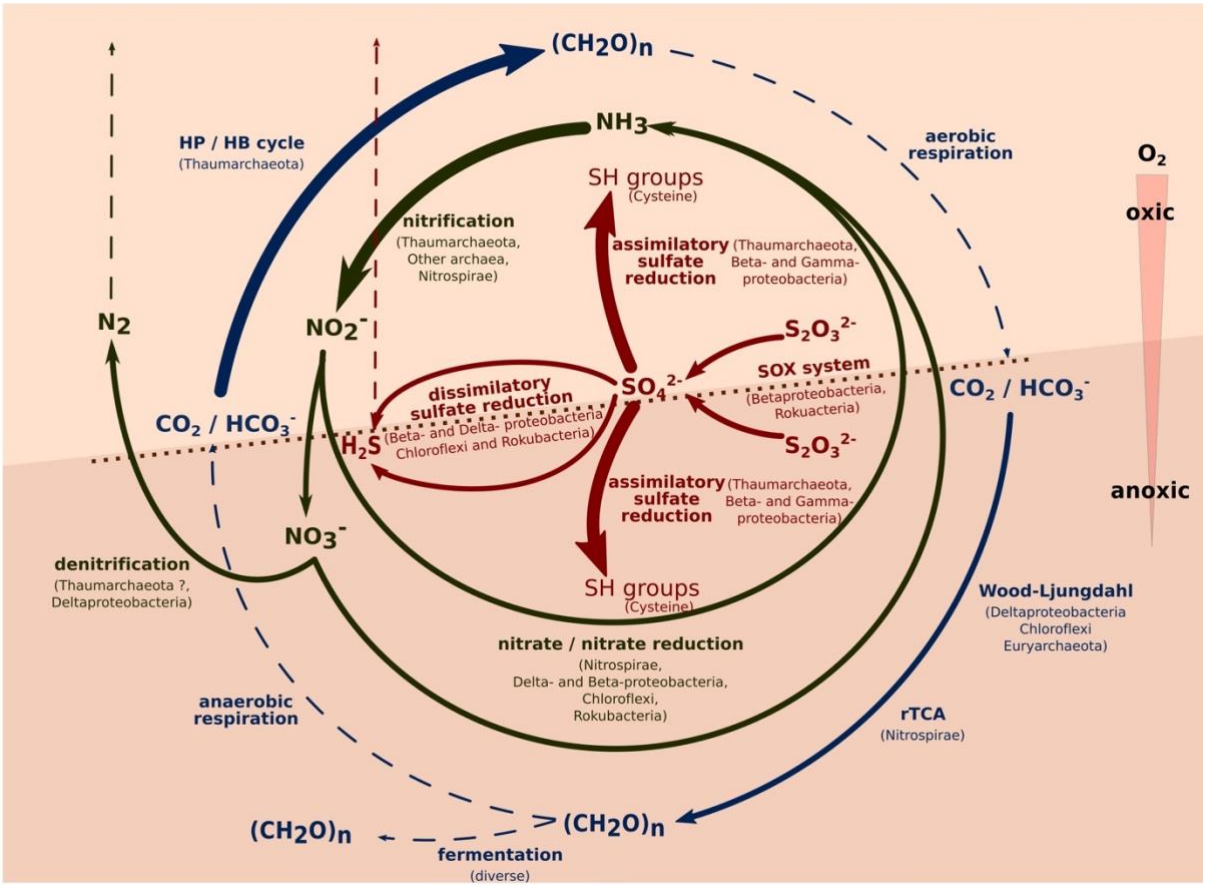
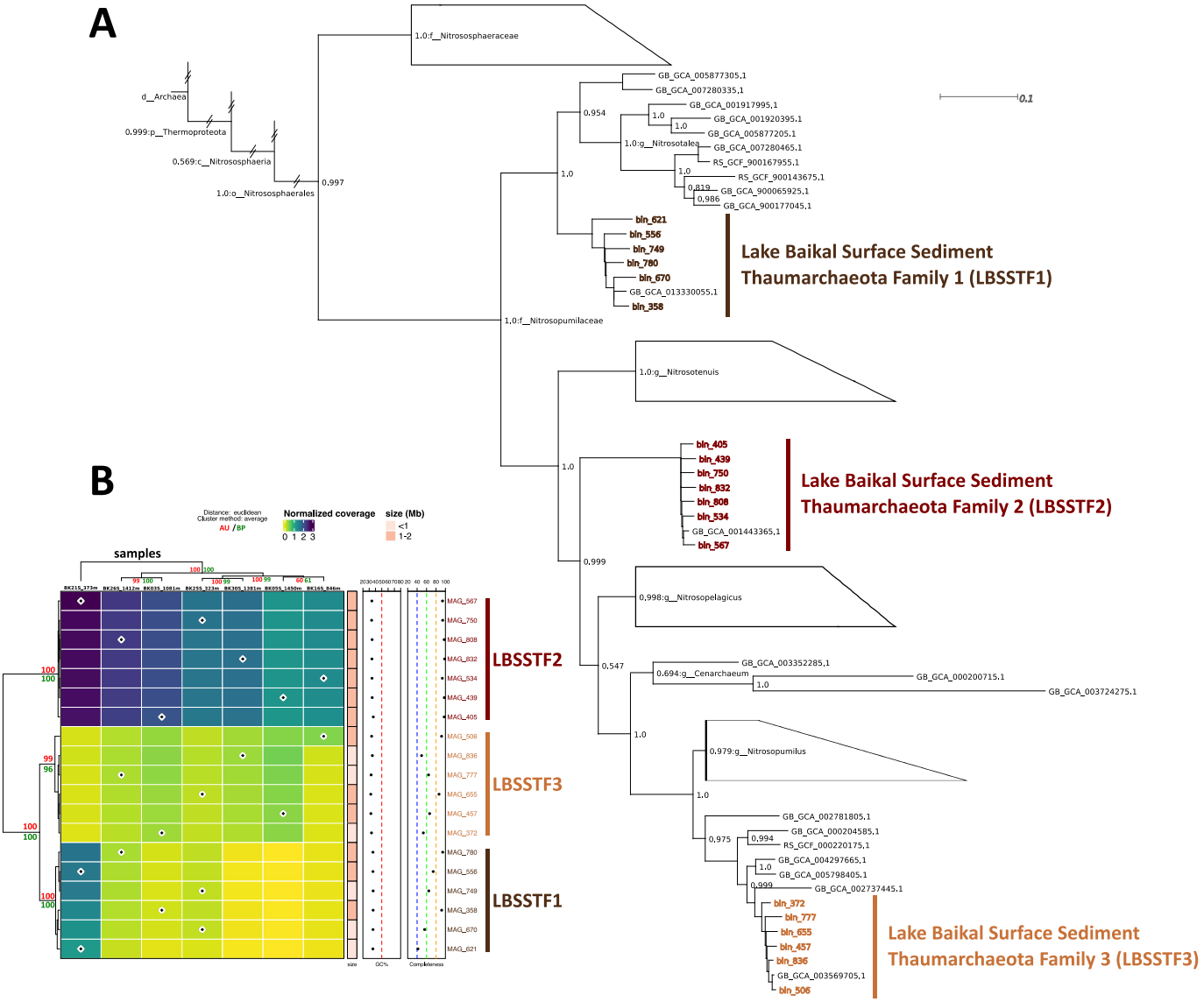
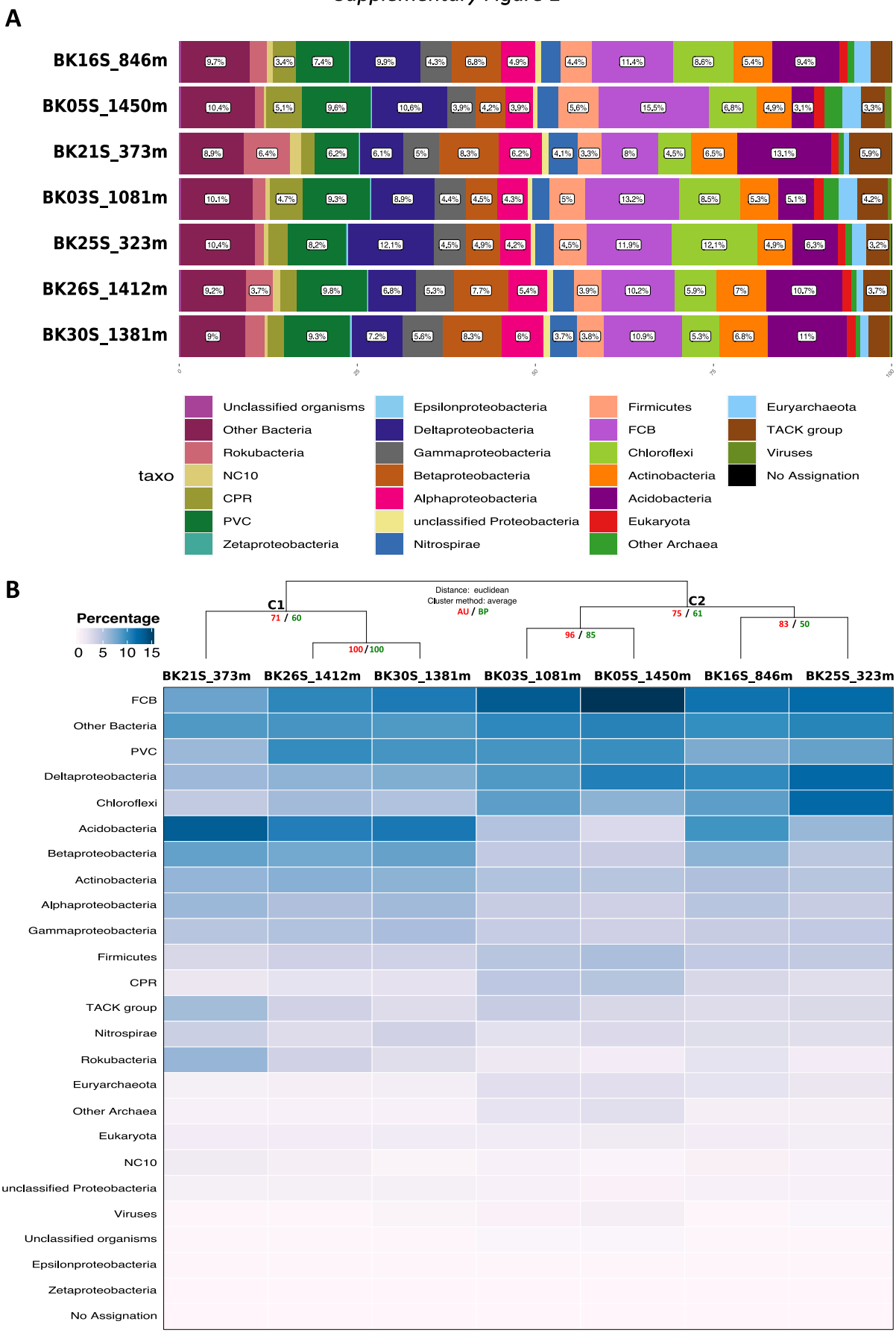


Figure 6

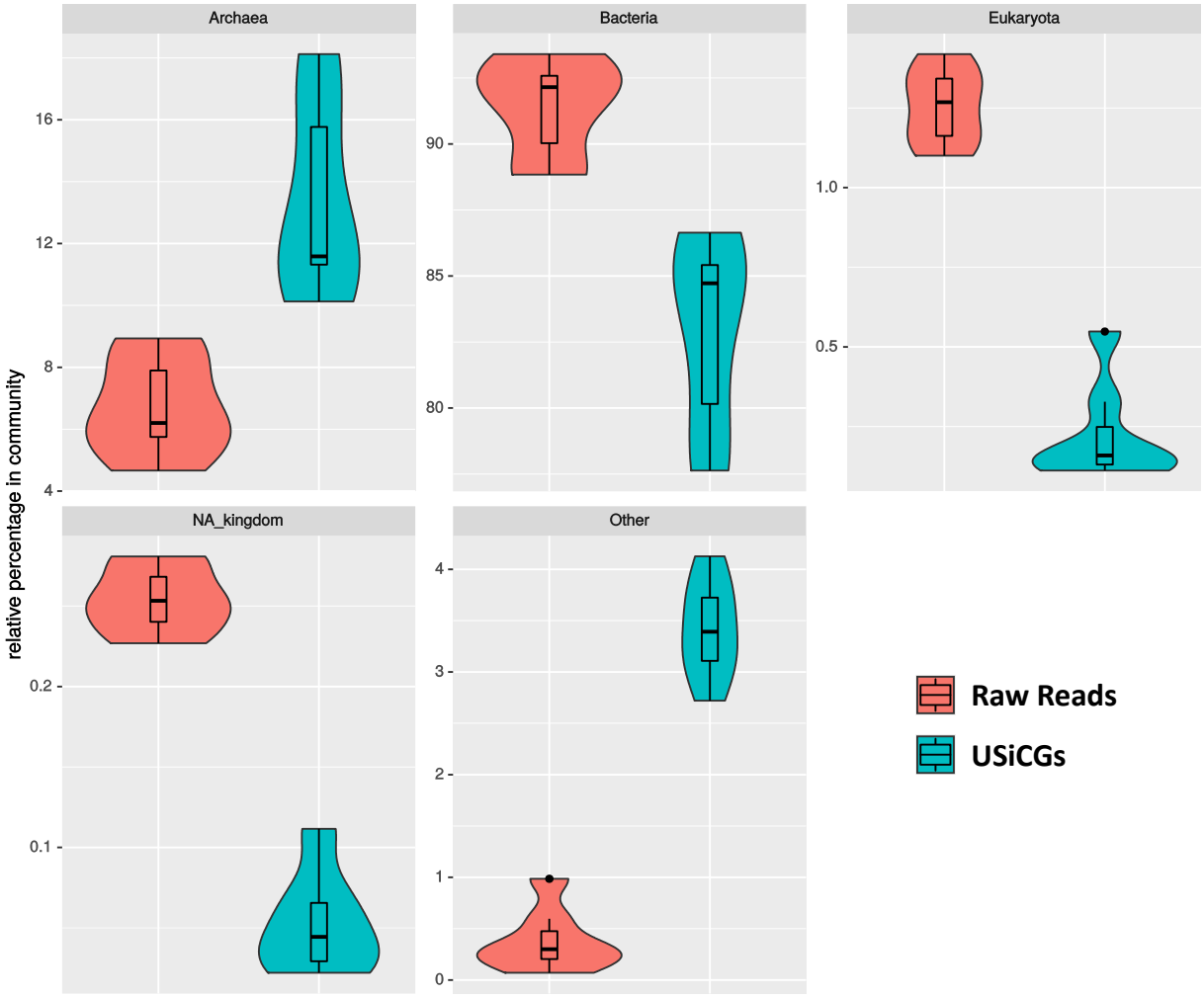

A
1.0:f__Nitrososphaeraceae
d__Archaea
0.999:p__Thermoproteota
0.569:c__Nitrososphaeria
1.0:o__Nitrososphaerales
0.997
GB_GCA_005877305.1
GB_GCA_007280335.1
GB_GCA_001917995.1
GB_GCA_001920395.1
GB_GCA_005877205.1
GB_GCA_007280465.1
RS_GCF_900167955.1
RS_GCF_900143675.1
GB_GCA_900065925.1
GB_GCA_900177045.1
0.954
1.0:g__Nitrosotalea
bin_621
bin_556
bin_749
bin_780
bin_670
GB_GCA_013330055.1
bin_358
Lake Baikal Surface Sediment
Thaumarchaeota Family 1 (LBSSTF1)
1.0:f__Nitrosopumilaceae
1.0:g__Nitrosotenuis
B
Distance: euclidean
Cluster method: average
AU / BP
Normalized coverage
0 1 2 3
size (Mb)
<1
1-2
samples
bin_405
bin_439
bin_750
bin_832
bin_808
bin_534
GB_GCA_001443365.1
bin_567
Lake Baikal Surface Sediment
Thaumarchaeota Family 2 (LBSSTF2)
0.998:g__Nitrosopelagicus
GB_GCA_003352285.1
0.694:g__Cenarchaeum
GB_GCA_000200715.1
GB_GCA_003724275.1
0.979:g__Nitrosopumilus
GB_GCA_002781805.1
GB_GCA_000204585.1
RS_GCF_000220175.1
GB_GCA_004297665.1
GB_GCA_005798405.1
GB_GCA_002737445.1
bin_372
bin_777
bin_655
bin_457
bin_836
GB_GCA_003569705.1
bin_506
Lake Baikal Surface Sediment
Thaumarchaeota Family 3 (LBSSTF3)
LBSSTF2
LBSSTF3
LBSSTF1
0.1

Supplementary information

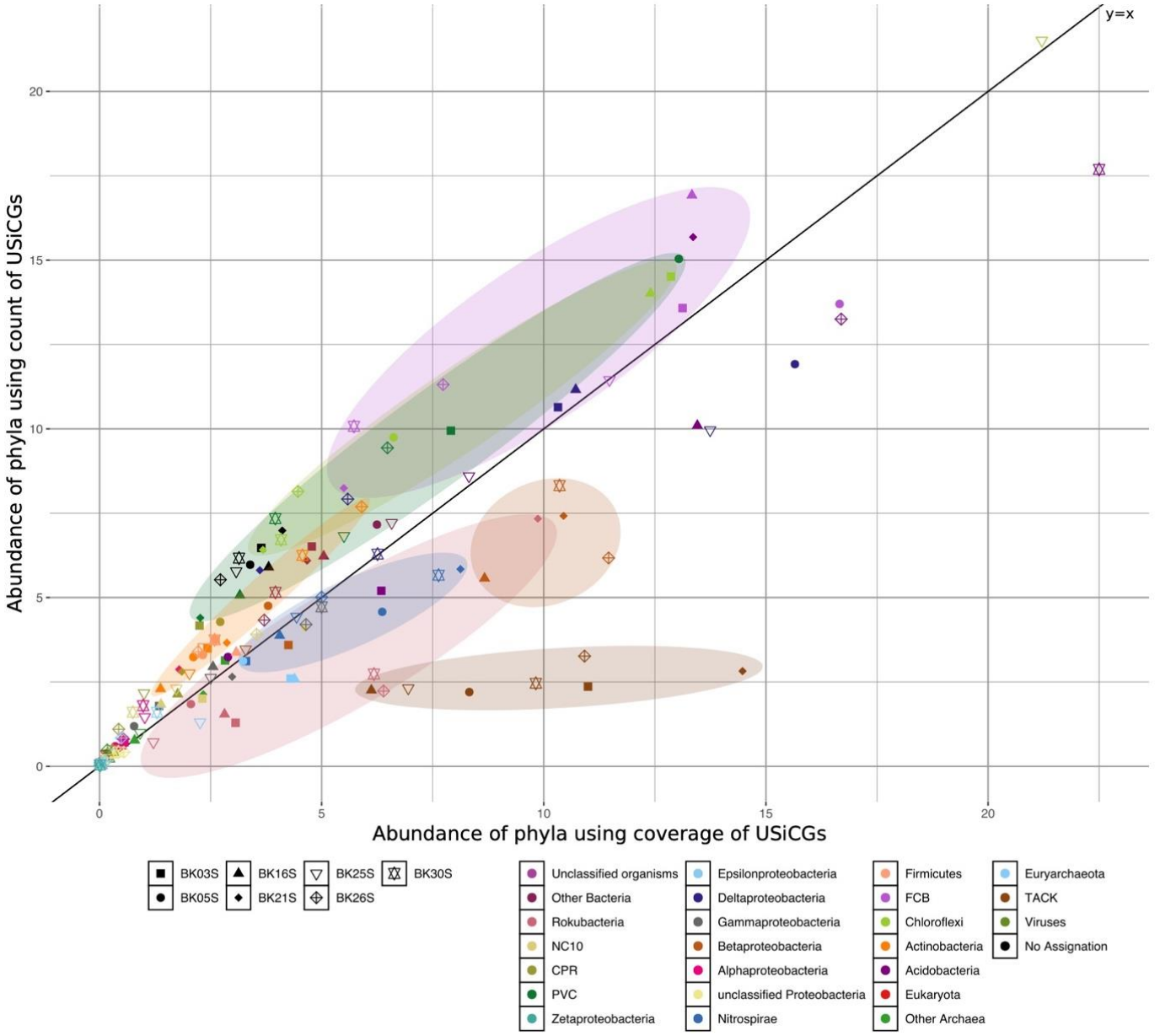
Supplementary Figure 1



817 *Supplementary Figure 2*
818
819

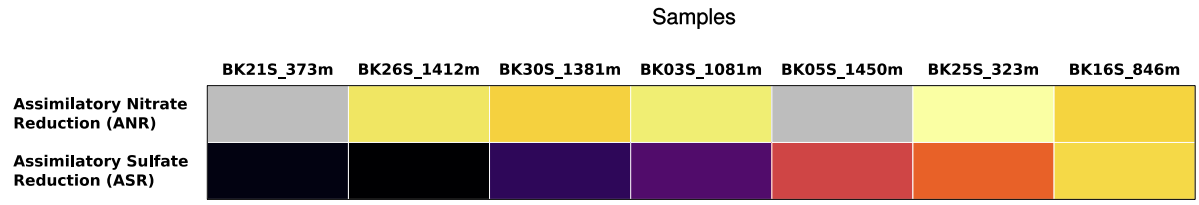


Supplementary Figure 3

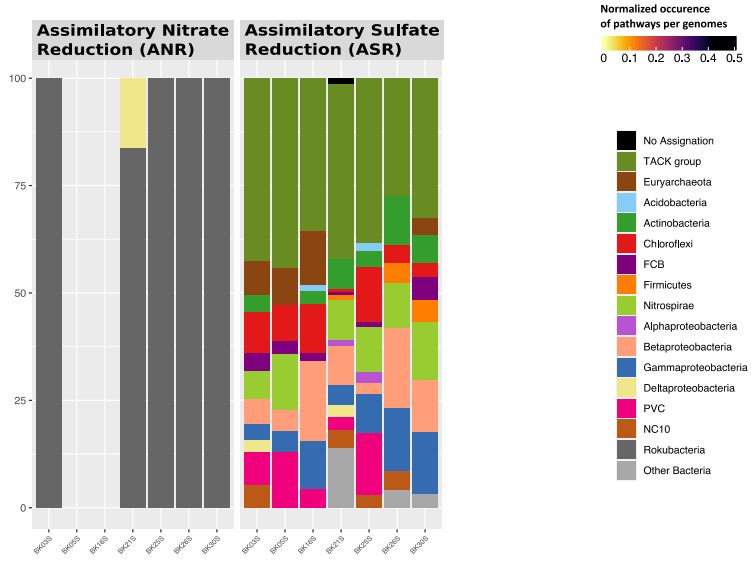


Supplementary Figure 4

A



B



Sample	Basin	Depth (m)	Depth (category)	Temperature (°C)	Sediment fraction (cmbsf)	Sampling date	Longitude	Latitude
BK03S	Central	1081	deep	4.8	0-2	02/07/2017	55.06.259 N	109.16.104 E
BK05S	Central	1450	deep	3.6	0-2	04/07/2017	53.30.175 N	107.52.633 E
BK16S	Northern	846	deep	4.2	0-3	29/06/2017	53.31.096 N	108.24.583 E
BK21S	Northern	373	medium	NA	0-3	29/06/2017	52.41.4014 N	106.44.208 E
BK25S	Central	323	medium	NA	0-3	04/07/2017	52.29.854 N	106.05.288 E
BK26S	Southern	1412	deep	3.9	0-3	05/07/2017	51.52.628 N	105.15.294 E
BK30S	Southern	1381	deep	NA	0-3	07/07/2017	51.47.817 N	104.46.449 E

Supplementary table S1*: Information about the sampling area and sampling sites of this study's samples.*

Sample	Initial reads	Paired-end reads remaining (PERR)	Fraction of PERR	Number of contigs	Number of contigs >1500bp (for binning)	Longest contig	GC%	N50	L50	Number of bin	Number of selected MAGs
BK03S	61 924 447	51 036 008	82.42%	2 467 567	46 152	132 953	51.28	1113	78 618	77	45
BK05S	41 923 334	34 548 190	82.41%	1 743 679	32 212	101 546	47.72	1100	56 543	58	29
BK16S	53 949 344	43 613 619	80.84%	2 048 138	49 016	285 219	55.99	1196	72 827	76	43
BK21S	58 124 584	45 913 131	78.99%	2 255 194	56 016	276 481	60.41	1177	86 528	72	41
BK25S	108 255 255	91 660 391	84.67%	4 349 648	122 863	97 715	54.64	1255	166 237	145	94
BK26S	41 804 886	34 852 184	83.37%	1 486 253	30 107	87 128	57.53	1116	51 291	40	29
BK30S	40 110 778	33 230 557	82.85%	1 346 810	26 281	70 929	58.14	1104	44 960	32	22

Supplementary table S2*: Statistics and characteristics of the metagenome datasets of lake Baikal upper layer sediments analyzed in this study.*

Initial taxonomy	Modified taxonomy
# remove "Terrabacteria group" as phylum level	
;Terrabacteria group;	;
;Terrabacteria group	;unclassified Terrabacteria
;unclassified Terrabacteria group;	;
# change the name of PVC and FCB phyla	
;FCB group;	;FCB;
;PVC group;	;PVC;
# remove Proteobacteria as Phylum and all classes are now phyla	
;Proteobacteria;delta/epsilon subdivisions;Epsilonproteobacteria	;Epsilonproteobacteria
;Proteobacteria;delta/epsilon subdivisions;Deltaproteobacteria	;Deltaproteobacteria
;Proteobacteria;Gammaproteobacteria	;Gammaproteobacteria
;Proteobacteria;Zetaproteobacteria	;Zetaproteobacteria
;Proteobacteria;Alphaproteobacteria	;Alphaproteobacteria
;Proteobacteria;Betaproteobacteria	;Betaproteobacteria
;Proteobacteria;unclassified Proteobacteria	;unclassified Proteobacteria
;Proteobacteria	;unclassified Proteobacteria
# remove "unclassified Bacteria;Bacteria candidate phyla" as phyum and class level remove "candidate" or "Candidatus" of new phyla names change "Patescibacteria group" to "CPR" remove "group" for Microgenomates and Parcubacteria as class of CPR	
unclassified Bacteria;Bacteria candidate phyla;candidate division BRC1	BRC1
unclassified Bacteria;Bacteria candidate phyla;candidate division CPR1	CPR;CPR1
unclassified Bacteria;Bacteria candidate phyla;candidate division CPR2	CPR;CPR2
unclassified Bacteria;Bacteria candidate phyla;candidate division CPR3	CPR;CPR3
unclassified Bacteria;Bacteria candidate phyla;candidate division Hyd24-12	FCB;Hyd24-12
unclassified Bacteria;Bacteria candidate phyla;candidate division Kazan-3B-28	CPR;Kazan
unclassified Bacteria;Bacteria candidate phyla;candidate division KD3-62	FCB;KD3-62
unclassified Bacteria;Bacteria candidate phyla;candidate division KSB1	FCB;KSB1
unclassified Bacteria;Bacteria candidate phyla;candidate division NC10	NC10
unclassified Bacteria;Bacteria candidate phyla;candidate division SR1	CPR;Absconditabacteria
unclassified Bacteria;Bacteria candidate phyla;candidate division TA06	FCB;TA06
unclassified Bacteria;Bacteria candidate phyla;candidate division WOR-1	WOR-1
unclassified Bacteria;Bacteria candidate phyla;candidate division WOR-3	FCB;WOR-3
unclassified Bacteria;Bacteria candidate phyla;candidate division WWE3	CPR;Katanobacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Abawacabacteria	CPR;Abawacabacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Acetothermia	Acetothermia
unclassified Bacteria;Bacteria candidate phyla;Candidatus Aminicenantes	Aminicemantes
unclassified Bacteria;Bacteria candidate phyla;Candidatus Atribacteria	Atribacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Berkelbacteria	CPR;Berkelbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Coatesbacteria	Coatesbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Dadabacteria	Dadabacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Delongbacteria	FCB;Delongbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Dependentiae	Dependentiae
unclassified Bacteria;Bacteria candidate phyla;Candidatus Desantisbacteria	PVC;Desantisbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Dojkabacteria	CPR;Dojkabacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Doudnabacteria	CPR;Parcubacteria;Doudnabacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Edwardsbacteria	FCB;Edwardsbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Eisenbacteria	FCB;Eisenbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Firestonebacteria	Firestonebacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Fischerbacteria	Fisherbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Fraserbacteria	Fraserbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Glassbacteria	FCB;Glassbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Handelsmanbacteria	FCB;Handelsmanbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Lindowbacteria	Lindowbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Margulisbacteria	Margulisbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Peregrinibacteria	CPR;Peregrinibacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Poribacteria	Poribacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Raymondbacteria	FCB;Raymondbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Riflebacteria	Riflebacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Rokubacteria	Rokubacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Saccharibacteria	CPR;Saccharibacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Schekmanbacteria	Schekmanbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Wallbacteria	Wallbacteria
unclassified Bacteria;Bacteria candidate phyla;Candidatus Wirthbacteria	Wirthbacteria
unclassified Bacteria;Bacteria candidate phyla;NA_order	Unclassified bacteria

unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Candidatus Gracilibacteria;unclassified Candidatus Gracilibacteria	CPR;Gracilibacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Amesbacteria	CPR;Microgenomates;Amesbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Beckwithbacteria	CPR;Microgenomates;Beckwithbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Blackburnbacteria	CPR;Microgenomates;Blackburnbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Chisholmbacteria	CPR;Microgenomates;Chisholmbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Collierbacteria	CPR;Microgenomates;Collierbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Curtissbacteria	CPR;Microgenomates;Curtissbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Daviesbacteria	CPR;Microgenomates;Daviesbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Gottesmanbacteria	CPR;Microgenomates;Gottesmanbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Levybacteria	CPR;Microgenomates;Levybacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Microgenomates	CPR;Microgenomates;unclassified Microgenomates
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Pacebacteria	CPR;Microgenomates;Pacebacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Roizmanbacteria	CPR;Microgenomates;Roizmanbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Shapirobacteria	CPR;Microgenomates;Shapirobacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Woesebacteria	CPR;Microgenomates;Woesebacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;Candidatus Woykebacteria	CPR;Microgenomates;Woykebacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;NA_genus	CPR;Microgenomates;unclassified Microgenomates
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Microgenomates group;unclassified Microgenomates group	CPR;Microgenomates;unclassified Microgenomates
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;NA_family;NA_genus	CPR;unclassified CPR;unclassified CPR
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Adlerbacteria	CPR;Parcubacteria;Adlerbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Andersenbacteria	CPR;Parcubacteria;Andersenbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Azambacteria	CPR;Parcubacteria;Azambacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Brennerbacteria	CPR;Parcubacteria;Brennerbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Buchananbacteria	CPR;Parcubacteria;Buchananbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Campbellbacteria	CPR;Parcubacteria;Campbellbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Colwellbacteria	CPR;Parcubacteria;Colwellbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Falkowbacteria	CPR;Parcubacteria;Falkowbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Giovannonibacteria	CPR;Parcubacteria;Giovannonibacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Harrisonbacteria	CPR;Parcubacteria;Harrisonbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Jacksonbacteria	CPR;Parcubacteria;Jacksonbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Jorgensenbacteria	CPR;Parcubacteria;Jorgensenbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Kaiserbacteria	CPR;Parcubacteria;Kaiserbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Kerfeldbacteria	CPR;Parcubacteria;Kerfeldbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Komeilibacteria	CPR;Parcubacteria;Komeilibacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Kuenenbacteria	CPR;Parcubacteria;Kuenenbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Liptonbacteria	CPR;Parcubacteria;Liptonbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Lloydbacteria	CPR;Parcubacteria;Lloydbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Magasanikbacteria	CPR;Parcubacteria;Magasanikbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Moranbacteria	CPR;Parcubacteria;Moranbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Nealsonbacteria	CPR;Parcubacteria;Nealsonbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Niyogibacteria	CPR;Parcubacteria;Niyogibacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Nomurabacteria	CPR;Parcubacteria;Nomurabacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Parcubacteria	CPR;Parcubacteria;unclassified Parcubacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Portnoybacteria	CPR;Parcubacteria;Portnoybacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Ryanbacteria	CPR;Parcubacteria;Ryanbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Spechtbacteria	CPR;Parcubacteria;Spechtbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Staskawiczbacteria	CPR;Parcubacteria;Staskawiczbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Sungbacteria	CPR;Parcubacteria;Sungbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Tagabacteria	CPR;Parcubacteria;Tagabacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Taylorbacteria	CPR;Parcubacteria;Taylorbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Terrybacteria	CPR;Parcubacteria;Terrybacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Uhrbacteria	CPR;Parcubacteria;Uhrbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Veblenbacteria	CPR;Parcubacteria;Veblenbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Vogelbacteria	CPR;Parcubacteria;Vogelbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Wildermuthbacteria	CPR;Parcubacteria;Wildermuthbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Wolfebacteria	CPR;Parcubacteria;Wolfebacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Yanofskybacteria	CPR;Parcubacteria;Yanofskybacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Yonathbacteria	CPR;Parcubacteria;Yonathbacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;Candidatus Zambryskibacteria	CPR;Parcubacteria;Zambryskibacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;NA_genus	CPR;Parcubacteria;unclassified Parcubacteria
unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;Parcubacteria group;unclassified Parcubacteria group	CPR;Parcubacteria;unclassified Parcubacteria

Supplementary table S3*: List of the manual changes into the Kaiju affiliated taxonomy in order to remove several groups/superphyla ("Bacteria Candidate Phyla", "Protebacteria" (because*

we used only the more precise affiliation), "Terrabacteria group") and rename others like ("FCB group", "PVC group", "Patescibacteria group").

KEGG Ortholog	Metabolic pathway	KEGG module
K00360	Assimilatory.Nitrate.Reduction	M00531
K00366	Assimilatory.Nitrate.Reduction	M00531
K00367	Assimilatory.Nitrate.Reduction	M00531
K00372	Assimilatory.Nitrate.Reduction	M00531
K10534	Assimilatory.Nitrate.Reduction	M00531
K17877	Assimilatory.Nitrate.Reduction	M00531
K00380	Assimilatory.Sulfate.Reduction	M00176
K00381	Assimilatory.Sulfate.Reduction	M00176
K00390	Assimilatory.Sulfate.Reduction	M00176
K00392	Assimilatory.Sulfate.Reduction	M00176
K00855	Calvin.Cycle	M00165
K01100	Calvin.Cycle	M00165
K01601	Calvin.Cycle	M00165
K01602	Calvin.Cycle	M00165
K05298	Calvin.Cycle	M00165
K14467	DC-HB.cycles	M00374
K00368	Denitrification	M00529
K00376	Denitrification	M00529
K02305	Denitrification	M00529
K04561	Denitrification	M00529
K15864	Denitrification	M00529
K00362	Dissimilatory.Nitrate.Reduction	M00530
K00363	Dissimilatory.Nitrate.Reduction	M00530
K03385	Dissimilatory.Nitrate.Reduction	M00530
K15876	Dissimilatory.Nitrate.Reduction	M00530
K00394	Dissimilatory.Sulfate.Reduction	M00596
K00395	Dissimilatory.Sulfate.Reduction	M00596
K11180	Dissimilatory.Sulfate.Reduction	M00596
K11181	Dissimilatory.Sulfate.Reduction	M00596
K09709	HP-bicycle	M00376
K14468	HP-bicycle	M00376
K14469	HP-bicycle	M00376
K14470	HP-bicycle	M00376
K14471	HP-bicycle	M00376
K14472	HP-bicycle	M00376
K15052	HP-bicycle	M00376
K14466	HP-HB.cycles	M00375
K15018	HP-HB.cycles	M00375
K15019	HP-HB.cycles	M00375
K15020	HP-HB.cycles	M00375
K15039	HP-HB.cycles	M00375

K00399	methanogenesis.all	M00567,M00356,M00357,M00563
K00401	methanogenesis.all	M00567,M00356,M00357,M00563
K00402	methanogenesis.all	M00567,M00356,M00357,M00563
K08265	methanogenesis.all	M00567,M00356,M00357,M00563
K00204	methanogenesis.CO2	M00567
K00319	methanogenesis.CO2	M00567
K13942	methanogenesis.CO2	M00567
K00531	N.Fixation	M00175
K02586	N.Fixation	M00175
K02588	N.Fixation	M00175
K02591	N.Fixation	M00175
K22896	N.Fixation	M00175
K22897	N.Fixation	M00175
K22898	N.Fixation	M00175
K22899	N.Fixation	M00175
K00436	NiFe.Hydrogenases	/
K00437	NiFe.Hydrogenases	/
K05586	NiFe.Hydrogenases	/
K05587	NiFe.Hydrogenases	/
K05588	NiFe.Hydrogenases	/
K18005	NiFe.Hydrogenases	/
K18006	NiFe.Hydrogenases	/
K18007	NiFe.Hydrogenases	/
K18008	NiFe.Hydrogenases	/
K10535	Nitrification	M00528 and M00804
K10944	Nitrification	M00528 and M00804
K10945	Nitrification	M00528 and M00804
K10946	Nitrification	M00528 and M00804
K01958	rTCA.Arnon.Buchanan.Cycle	M00173
K01959	rTCA.Arnon.Buchanan.Cycle	M00173
K01960	rTCA.Arnon.Buchanan.Cycle	M00173
K15230	rTCA.Arnon.Buchanan.Cycle	M00173
K15231	rTCA.Arnon.Buchanan.Cycle	M00173
K15232	rTCA.Arnon.Buchanan.Cycle	M00173
K15233	rTCA.Arnon.Buchanan.Cycle	M00173
K15234	rTCA.Arnon.Buchanan.Cycle	M00173
K18209	rTCA.Arnon.Buchanan.Cycle	M00173
K18210	rTCA.Arnon.Buchanan.Cycle	M00173
K18556	rTCA.Arnon.Buchanan.Cycle	M00173
K18557	rTCA.Arnon.Buchanan.Cycle	M00173
K18558	rTCA.Arnon.Buchanan.Cycle	M00173
K18559	rTCA.Arnon.Buchanan.Cycle	M00173

K18560	rTCA.Arnon.Buchanan.Cycle	M00173
K17222	SOX.system	M00595
K17223	SOX.system	M00595
K17224	SOX.system	M00595
K17225	SOX.system	M00595
K17226	SOX.system	M00595
K17227	SOX.system	M00595
K22622	SOX.system	M00595
K00198	Wood-Ljungdahl	M00377
K05299	Wood-Ljungdahl	M00377
K14138	Wood-Ljungdahl	M00377
K15022	Wood-Ljungdahl	M00377
K15023	Wood-Ljungdahl	M00377
Hereafter Universal Single Copy Genes (USiCGs) list from MUSiCC software (Manor 2015)		
K00133	USiCGs	
K00789	USiCGs	
K00927	USiCGs	
K00939	USiCGs	
K01689	USiCGs	
K01803	USiCGs	
K01866	USiCGs	
K01867	USiCGs	
K01868	USiCGs	
K01869	USiCGs	
K01870	USiCGs	
K01872	USiCGs	
K01873	USiCGs	
K01874	USiCGs	
K01875	USiCGs	
K01876	USiCGs	
K01881	USiCGs	
K01883	USiCGs	
K01887	USiCGs	
K01889	USiCGs	
K01890	USiCGs	
K01892	USiCGs	
K01937	USiCGs	
K02357	USiCGs	
K02519	USiCGs	
K02528	USiCGs	
K02600	USiCGs	

K02601	USiCGs	
K02835	USiCGs	
K02838	USiCGs	
K02863	USiCGs	
K02864	USiCGs	
K02867	USiCGs	
K02871	USiCGs	
K02874	USiCGs	
K02876	USiCGs	
K02878	USiCGs	
K02879	USiCGs	
K02881	USiCGs	
K02884	USiCGs	
K02886	USiCGs	
K02887	USiCGs	
K02890	USiCGs	
K02892	USiCGs	
K02895	USiCGs	
K02904	USiCGs	
K02906	USiCGs	
K02926	USiCGs	
K02931	USiCGs	
K02933	USiCGs	
K02946	USiCGs	
K02948	USiCGs	
K02950	USiCGs	
K02952	USiCGs	
K02956	USiCGs	
K02961	USiCGs	
K02965	USiCGs	
K02967	USiCGs	
K02982	USiCGs	
K02986	USiCGs	
K02988	USiCGs	
K02992	USiCGs	
K02994	USiCGs	
K02996	USiCGs	
K03040	USiCGs	
K03076	USiCGs	
K03106	USiCGs	
K03110	USiCGs	
K03438	USiCGs	

K03470	USiCGs	
K03664	USiCGs	
K03687	USiCGs	
K03702	USiCGs	
K06942	USiCGs	
K09903	USiCGs	
K10773	USiCGs	

Supplementary table S4: *List of the KEGG orthologs (KOs) used for metabolic inferences and their respective metabolic pathway and KEGG module.*

Carbon fixation pathways

	BK03S	BK05S	BK16S	BK21S	BK25S	BK26S	BK30S	sum	avg
HP-HB.cycles	0.22638	0.14008	0.12387	0.30808	0.12366	0.24008	0.22370	***1.38584***	***0.19798***
Wood-Ljungdahl	0.10472	0.25384	0.11161	0.00502	0.18873	0.06406	0.03826	***0.76623***	***0.10946***
Calvin.Cycle	0.05436	0.02410	0.09112	0.08590	0.03181	0.15442	0.05133	***0.49305***	***0.07044***
rTCA.Arnon.Buchanan.Cycle	0.03402	0.05988	0.02621	0.02999	0.00470	0.02592	0.02352	***0.20423***	***0.02918***
	0.41947	***0.47790***	***0.35281***	***0.42899***	***0.34890***	***0.48447***	***0.33680***		

Energy metabolism pathways

	BK03S	BK05S	BK16S	BK21S	BK25S	BK26S	BK30S	sum	avg
Nitrification	0.21509	0.19247	0.09651	0.33798	0.16927	0.22733	0.20049	***1.43913***	***0.20559***
Dissimilatory.Nitrate.Reduction	0.14130	0.09962	0.07343	0.08846	0.09192	0.05666	0.07317	***0.62455***	***0.08922***
SOX.system	0.06459	0.03583	0.07107	0.13898	0.03017	0.14067	0.13341	***0.61472***	***0.08782***
Denitrification	0.11567	0.10463	0.12505	0.04421	0.09619	0.05091	0.06117	***0.59783***	***0.08540***
Dissimilatory.Sulfate.Reduction	0.05189	0.13753	0.08861	0.03715	0.09204	0.05983	0.07803	***0.54508***	***0.07787***
NiFe.Hydrogenases	0.02391	0.07664	0.01414	NA	0.05047	NA	0.01475	***0.17991***	***0.03598***
	0.61245	***0.64672***	***0.46882***	***0.64678***	***0.53005***	***0.53540***	***0.56101***		

Assimilatory pathways

	BK03S	BK05S	BK16S	BK21S	BK25S	BK26S	BK30S	sum	avg
Assimilatory.Nitrate.Reduction	0.03054	NA	NA	0.03768	0.01143	0.05391	0.05470	***0.18826***	***0.03765***
Assimilatory.Sulfate.Reduction	0.32040	0.40638	0.19176	0.42050	0.15465	0.04986	0.35613	***1.89969***	***0.27138***
	0.35095	***0.40638***	***0.19176***	***0.45818***	***0.16608***	***0.10376***	***0.41083***		

Supplementary table S5: relative percentages of organisms using the corresponding pathways (normalized by USiCGs abundances intra and inter-samples using MUSiCC (Manor and Borenstein, 2015) *software)*

4.3 Metagenomes Assembled Genomes or MAGs

In the previous section I mentioned that we recovered 304 MAGs[1] (details of these can be found in Figure 4.2) from our datasets. However, only the Thaumarchaeotal ones were included in the study at this stage. An overview of the other MAGs is available on Figure 4.1, where I highlight the 55 high-quality MAGs recovered in the lake Baikal upper layer sediment samples. These high-quality MAGs were predicted with a completeness above 90% and a contamination below 10% by the CheckM tool (Parks et al., 2015).

Many MAG families[2] can be seen on this preliminary figure, especially a Thaumarchaeotal, Betaproteobacterial, Acidobacterial and Euryarchaeal ones. Thaumarchaeota are the only lineages with a GC% under 50% and Acidobacteria were predicted with large genome sizes, in accordance with the literature (Stieglmeier et al., 2014; Eichorst et al., 2018). Chloroflexi MAGs were also found in this set of high-quality MAGs. These, however, did not group into a MAG family, suggesting that their abundance did not allow for consistent high-quality reconstruction across sites or that, more intriguingly, Chloroflexi taxa might be more diverse across different sites. This latter possibility would be in line with the idea of a stability of the microbial communities along geographical and depth gradients across the lake at the phyla-level, but within those phyla, the composition of the communities on lower phylogenetic levels might show differences. Rokubacteria were represented by 8 MAGs binned with completeness ranging from 40 to 80%, thus none of which were retained as high-quality MAGs.

In conclusion, this project is still ongoing. At this stage, the datasets have been computed bioinformatically, producing contigs, MAGs, protein predictions, taxonomic predictions for genes and contigs. Many of these need to be explored futher to gain a deeper insight into the microbial ecology of these unique samples.

[1] MAGs were selected with loose parameters, completeness>30% and contamination<10%

[2] MAG families in the draft manuscript have been defined as group of MAGs with same affiliated taxonomy and

same coverage profil along the samples but recovered from different samples

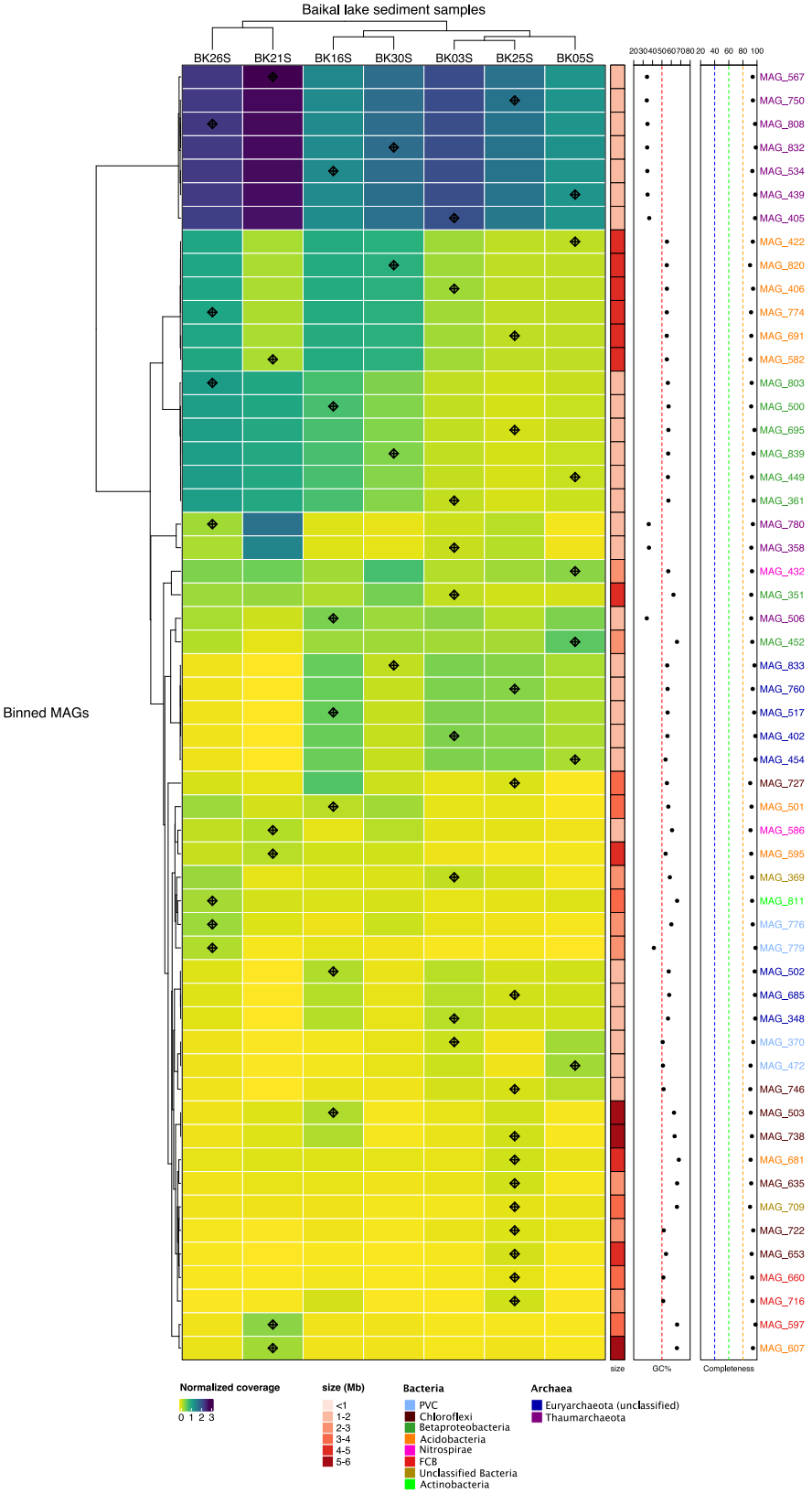


Figure 4.1 – The fraction of high-quality MAGs with completeness above 90% and contamination less than 10% recovered in lake Baikal upper layer sediments. MAGs names have been colored with their affiliated taxonomy predicted by consensus of hits from Kaiju (Menzel et al., 2016). Symbols in rectangles indicates the sample of origin.

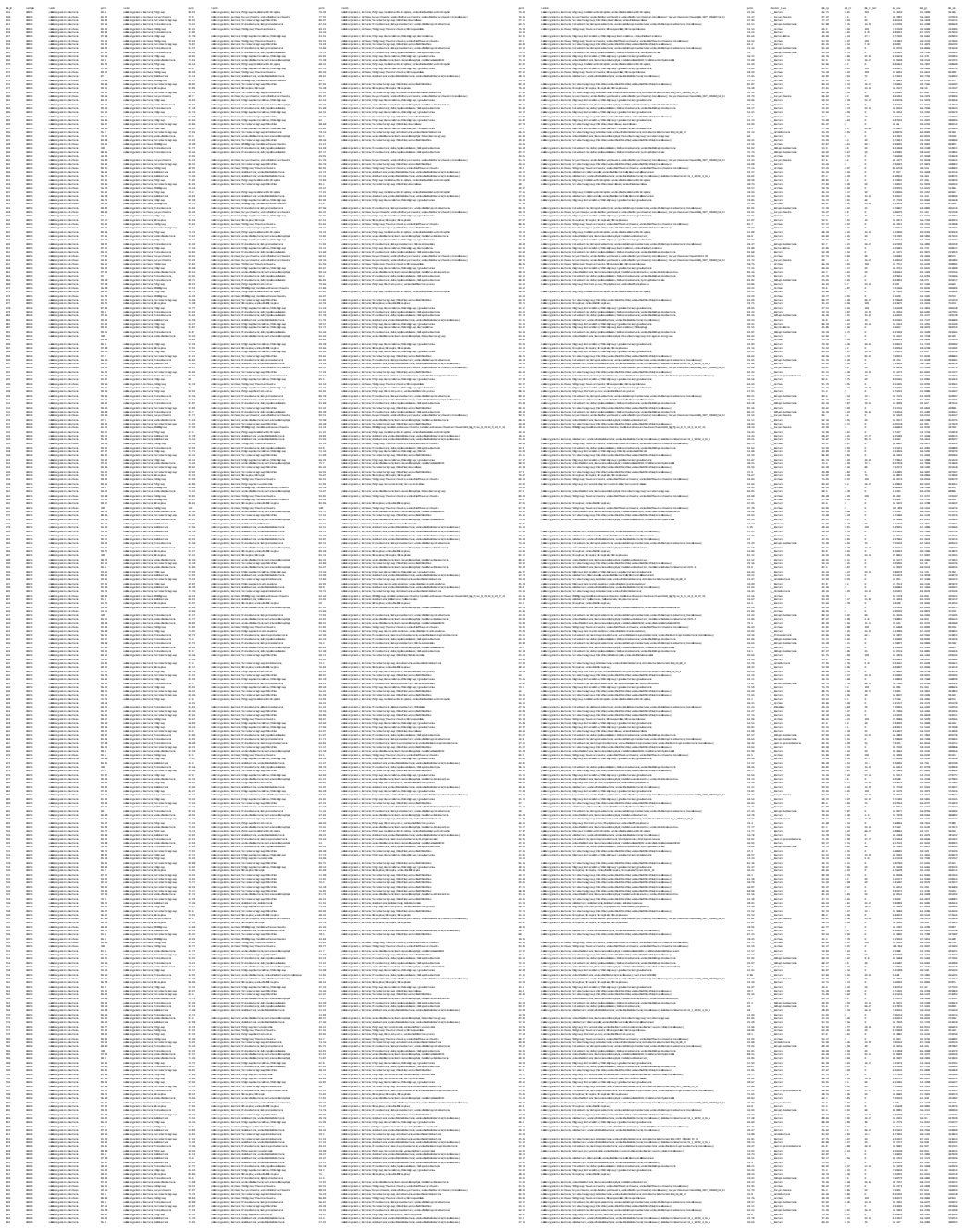


Figure 4.2 – All the MAGs with completeness above 30% and contamination less than 10% recovered in lake Baikal upper layer sediments. rankX: cpctX: percntage

CHAPTER

5

DISCUSSION AND PERSPECTIVES

The overarching work of this thesis involved to a large extent the employment of bioinformatics methods, including the development of a metabarcoding pipeline. I have applied the metabarcoding approach to describe the communities in two unique environments, and I have also employed metagenomics to better describe the diversity of microbial upper sediment layers of lake Baikal. This work also included the investigation of the major metabolic pathways and the taxa bearing them in the sediments of lake Baikal, and I have also reconstructed MAGs for a closer look at the phylogeny and gene content of some key players.

This discussion section contains four parts. The first section discusses the pipeline I have developed. I will first describe the limitations of the current version of the pipeline as well as possible improvements that could be introduced. The second section discusses at length the limitations and particularities of the bioinformatic approaches that have been applied throughout this thesis, discussing where-ever possible the implications for the work described in this

manuscript. A third section discusses the main findings of the thesis with regards to investigating the microbial communities inhabiting the upper layers of deep sediments of lake Baikal. Finally, I discuss possible ways to address the limitations presented earlier in future work as well as perspectives and future directions to build on the biological findings of this thesis.

5.1 In-house metabarcoding pipeline

Since the development of the pipeline in the early years of this thesis, I have had the opportunity to apply it in a number of studies, in my own work described in this manuscript and as part of collaborations which can be found in the Appendix (see Chapters 2 and 3). Throughout this time, I have been made aware of a number of ways in which the pipeline could be improved.

5.1.1 Reference databases

The lack of consensus in the taxonomic reference databases is a limitation of metabarcoding in general (see Section 1.2.2.2). Indeed, comparing information with an outdated or erroneous reference database can lead to inaccurate scientific conclusions. While ideally, a reference database would be a solid foundation for a biological study, the multitude of reference databases can cast doubt and confusion in the minds of impressionable young biologists.

The current version of the metabarcoding pipeline uses an in-house database, in which two databases of reference sequences were manually adapted and merged in order to improve the taxonomic resolution. However, at the time of writing this manuscript, these two databases are quite outdated: PR2v4.5 (now 4.12; Guillou et al. (2013)) and SILVAv128 (now 132; Quast et al. (2013)). The extensive manual adaptations needed to merge these too versions are not compatible with automatic updates following the database releases. As new data is collected and phylogenetic studies are released, these databases are frequently updating and/or changing their taxonomic groups. While most of the changes do not affect the relations on a high level (e.g. phyla), relying on outdated versions to classify the OTUs is a major drawback for the current version of the pipeline.

In order to address this issue, three approaches are possible, listed here in order of ease of implementation. First, one could choose a single reference database and update it automatically when a new release becomes available. This has the potential issue of making it harder to reproduce results produced just before an update. Second, one can give the user a choice between multiple external databases, with new releases being automatically included and the old releases can be removed when outdated (e.g. at *x* month old or at the *xTH* release before the current one) or manually. A possible drawback of this approach would be the space required to keep all the versions of the available databases on disk. Third, one could automate the merging process of the two databases as they are updated. However, it is possible for unforeseen errors to occur during an automated merge.

5.1.2 ASV

So far, the metabarcoding pipeline is available with operational taxonomic units (OTUs) as the only available output. Recent approaches including DADA2 (Callahan et al., 2016) explore the microbial diversity at the strain level and therefore produce amplicon sequence variants (ASV) instead. For more in-depth analysis on the species/strain level, the pipeline would require modifications. In particular, one would either need to program in the ASV approach and DADA2 into the pipeline, or possibly use an integrative approach involving Qiime2.

5.1.3 Metadata for online data submission

One of the initial motivations for creating an in-house pipeline was the integration with an effective local SQL database. This database was initially engineered to store the data resulting from the pipeline but also the metadata upstream of the pipeline, containing information on the origin of the data outputted by the sequencer. Examples of categories included in the metadata are sampling place, sampling date, sampling coordinates, sample nature, replicates Y/N, MID, PCR as well as sequencer/sequencing information such as the platform, and strategy. I followed the standards introduced by the Genomic Standards Consortium (GSC; Field et al. (2011)), who

published their list of specification about genomes and metagenomes (Field et al., 2008), calling them the minimum information about a genome sequence (MIGS) and minimum information about a metagenome sequence (MIMS), respectively.

The current version of the pipeline requires the information in the MIxS format[1]. In order to upload the data on any of the databases in the international nucleotide sequence database collaboration (INSDC[2]), additional information can also be inputted and it is certainly good practice to do so. These metadata help other scientist to discriminate between the plethora of data available online. As more and more studies incorporate previously published datasets in their analysis, the need for comprehensive detailed metadata cannot be understated. When uploading the data to the aforementioned databases, the metabarcoding and metagenomic data themselves are usually uploaded as SRA (sequence read archive) *via* a submission portal available on every INSDC website. This allows other scientists to incorporate the data in their own pipeline and treat the sequences with updated or different tools to produce other scientific outputs, in addition to following the same protocols to replicate the same results as the authors of the submitted data.

This uploading step can be very time-consuming for scientists, cross-checking information for every entry to the form. A possible improvement to the pipeline which could save time would be a module to make this metadata information available in the form of a MIMS table, which can then be directly uploaded on the INSDC website. Indeed, the local SQL database already stores all the MIMS information precisely with the aim to help submit data prior to manuscript review or publishing, and the web interface was also designed to reflect this aim, and it therefore requires users to fill in all the required metadata. At its current state, all this metadata information is available only through database command lines in PostgreSQL, but the creation of such a module would be helpful when one is submitting SRA data for metabarcoding-based scientific communications.

[1]Generic format name for all *minimum information about a **x** sequence*. MIxS include multiple checklists, including MIGS, MIMS and other standards, discussed in detail here: `https://gensc.org/mixs/mixs-compliance-and-implementation/`

[2]INSDC is a collaboration involving NCBI/GenBank (USA; `www.ncbi.nlm.nih.gov/`), EBI-ENA (Europe; `www.ebi.ac.uk/`), and the DDBJ (Japan; `www.ddbj.nig.ac.jp/`)

5.1.4 Web interface

In its current state, the web interface is not completely bug-free and the bugs can be of either informatics or biological origin.

One possible improvement that needs implementation is the addition of further verification steps to check that all the inputs are coherent and in the expected format. One may, for example, include a verification that the minimum overlapping parameter is smaller than the maximum overlapping parameter. Moreover, while creating a new *run*, it would be helpful to verify that the files R1 and R2 are in a format compatible with the pipeline, that the provided filepaths are different (in case of a typo where the user inputted the same file twice) and contain the same number of sequences.

Another possible improvement would be to allow for degenerated nucleotides (standing for any of the A, T, C or G nucleotide) when filling in the primers used in PCR. Given this lack of functionality, PCR primers involving degenerated nucleotides currently need to be added to the SQL database, bypassing the web interface.

Another area for improvement would be to extend the information displayed on the 'Dashboard' main page. Presently, the *run* information is not visible while browsing the *samples*. Along the same vein, the trimming length and the merging parameters are not displayed while browsing the *analyses*.

While the resolution of these issues could improve the user experience, they are not critical to the pipeline function and do not therefore affect the scientific results.

5.1.5 Sequence formats

Another possible improvement to the pipeline would be to make it compatible with extensions besides FASTQ and FASTQ.gz, which are the only extensions currently supported. Other formats such as bz2 and tar.gz offer more compression, and the integration of those to the pipeline would be helpful to easily run it over sequences sent by a company or a collaborator in a compressed format.

5.1.6 Validation

In terms of performance, the pipeline has not been directly tested to date and it could be benchmarked by characterizing mock communities.

In addition, it would also be interesting to compare the pipeline performance relative to Qiime2 (Bolyen et al., 2019). However, given the ASV focus of Qiime2 and its lack of implementation of Swarm and other OTU-based clustering tools, a direct comparison of the results is not possible. While a performance comparison is thus unattainable at this stage, in terms of the user experience, the pipeline has the major advantages of benefiting a web interface magnifying the interaction with a database. On this front, Qiime2 is currently developing a graphical interface named 'q2studio' which is currently at the prototype phase. `https://docs.qiime2.org/2020.8/interfaces/q2studio/`.

5.2 Limitations of metabarcoding and metagenomics

In addition to the limitations already mentioned in the introduction of metabarcoding (Section 1.2.2.2) and metagenomics (Section 1.2.2.3), this section discusses at length the possible sources of bias with these approaches. The resulting methodological considerations are discussed, with a focus on sediments, and the potential solutions to those problems are presented where applicable.

5.2.1 Pre-sequencing sources of bias

In a molecular ecology study, the first possible introduction of bias is with the act of sampling itself. Indeed, it is possible that simply due to chance, the extracted sample contains a microhabitat not representative of the community targeted by the study. The use of biological replicates can prevent such a mishap, although this increases the costs of the study (see Section 1.2.2.2). Moreover, the type of debris present in the sample and in particular, the presence of some specific (in)organic molecules such as calcium ions or polysaccharides (which can depend on the type of sample, e.g. soil, human microbiome, planktonic) can inhibit some steps in the

molecular biology protocols used for DNA extraction and purification (Pollock et al., 2018).

Another source of bias are the sources of DNA in the sample. Indeed, common protocols of DNA extraction and purification do not discriminate between intracellular (iDNA) or extracellular (exDNA) DNA or DNA from active or dormant cells (Knight et al., 2018). This means that a proportion of the sequenced DNA does not reflect active members of the community. This issue is discussed in detail in the next section Section 5.2.2.

Furthermore, cell lysis during the extraction part of the protocol can be an additional source of bias. The cell wall composition of a given organism determines its sturdiness over the cell lysis protocol, which may result in the under-representation of some organisms (see Section 1.2.2.2). While this cell wall bias is known for some microbial taxa and can therefore be corrected for in some cases, this potential is limited to these known taxonomic groups. Such corrections can therefore not be automatically applied to analyses aiming to explore a community's diversity and highlight potential new taxonomic groups. To mitigate this bias, one may opt for technical replicates with multiple lysis protocols.

5.2.2 Extracellular or Cell-free DNA

As mentioned previously Section 5.2.1, the total extracted environmental DNA contains both intracellular as well as extracellular DNA (iDNA and exDNA, respectively). More than just a by-product of cell death, exDNA plays roles of ecological importance in its environment (Nagler et al., 2018), including supplying benthic heterotrophic communities with nutrients and in particular phosphorous (Dell'Anno, Danovaro, 2005), serving as a substrate of horizontal gene transfer (Collins, Deming, 2011), contributing to the structural stability of biofilms[3]

There are multiple sources of exDNA in sediments such as the upper layer sediments of lake Baikal explored in this manuscript. First, exDNA can be released locally by dying cells in the sediments themselves. Second, exDNA may be transported from the upper layers of the lake

[3]This role of exDNA can be very relevant to the floating mats in Movile Cave. In such cases, the majority of the exDNA can be secreted by a handful of species (Dominiak et al., 2011), which may possibly introduce bias in approaches sequencing all the DNA (such as metagenomics). However, the 16S and 18S rRNA genes amplified for metabarcoding in the study presented in Chapter 2 are unlikely to be secreted.

through sinking debris. Third, in some cases exDNA has been shown to be actively secreted (Nagler et al., 2018), although known examples concentrate on secretion by multicellular organisms or by microorganisms in biofilm formation. Fourth, exDNA may be excreted by multicellular organisms who failed to completely digest the microbial DNA of their prey[4]. Only exDNA from the second scenario can be considered a contaminant representing non-local communities from upper layers, biasing the abundances in favour of upper layer taxa, as exDNA of local origins would represents the community just as iDNA.

What is the proportion of exDNA in the total DNA sequenced? In sediments, the amount of exDNA is three orders of magnitude higher than the one in plankton (Torti et al., 2015), reaching up to 85–98% of the total DNA in some marine sediments (Alawi et al., 2014). Salinity has been shown to help preserve DNA Borin et al. (2008), so the expected amount of exDNA in freshwater sediments is expected to be lower. A lower bound of the possible amount of exDNA in lake Baikal is the reported value (40–50%) in the upper layer sediments of the freshwater lake Towuti (Vuillemin et al., 2016), but the high temperature in this tropical lake year-round (28°C) is incompatible with exDNA preservation, so the amount of exDNA is expected to be higher in lake Baikal, where the temperature is 4°C year-round.

What part of this substantial amount of exDNA representative of dead cells from upper layers, biasing ecological results? Torti et al. (2018) examined the issue in depth, describing the metabarcoding-inferred phylogeny of marine sediments using exDNA and iDNA separately. The main findings were that the overall phylogenetic profiles are preserved and high-level representation was very similar between the exDNA and iDNA metabarcoding results, and the nearest BLAST hits of the vast majority of exDNA sequences were produced *in situ* ($\sim$80% of exDNA of sediment origin, compared to $\sim$85% for iDNA, with the rest labelled as 'symbiotic', 'water' or other categories, with the exDNA and iDNA profiles highly similar), suggesting that despite the potentially large proportions of exDNA in our samples, the vast majority of exDNA would be representative of the sampling site, with only a minor possible effect of exDNA sequenced from dead organisms of non-local origin.

[4]Indeed, micrometazoa have been documented in upper layer of the deep sediments of the South Basin (Yi et al., 2017), making this a possible scenario.

Nevertheless, there are possible biases to take into account. The exDNA might have led us to overestimate the richness and diversity of the sampled communities (non-significant trend in Lennon et al. (2018) and statistically confirmed by Torti et al. (2018)). This may be due to temporal dynamics (seasonal changes or other disturbances) as including both exDNA and iDNA, we effectively sample both from the past and the present of a dynamic community in addition to (minor) proportion of exDNA sunk from upper layers. Moreover, it is of note that exDNA seems to represent a higher proportion of archaeal DNA – this was the case in the marine upper layer sediments in Torti et al. (2018) as well as the freshwater sediments in Vuillemin et al. (2016) (although it was not statistically tested). Future studies can investigate if this is a significant trend and if it may be caused by the higher turnover or over-representation of specific archaeal taxa in the exDNA.

To avoid any possible bias due to the inclusion of genes from dead organisms, there are molecular approaches to separate out exDNA from iDNA. One alternative is Viability PCR, in which prior to PCR the samples are first incubated in propidium monoazide (PMA) (Cangelosi, Meschke, 2014). This molecule binds to cell-free DNA in response to light, interfering with its amplification by PCR, but DNA in intact cells is protected as the charged PMA cannot penetrate the membrane. PMA was introduced when its predecessor, EMA, was shown to partially penetrate the membranes of certain bacterial species, biasing results against those taxa (Nocker et al., 2006) and to date, no such bias has been reported for PMA. An alternative approach to investigate viable cells only is metatranscriptomics, studying RNA which degrades rapidly outside the cell, an advantage that, at the same time, makes it more difficult to study (Laroche et al., 2017). It is possible that such an approach may lead to a bias towards organisms growing or adaptation over other members of the community (Blazewicz et al., 2013). In addition to these approaches to discard exDNA altogether, protocols have been developed to separate out exDNA from iDNA and potentially analyse both separately (Alawi et al., 2014; Lever et al., 2015); indeed, this was the strategy that allowed Torti et al. (2018) to compare the exDNA and iDNA content as described above.

5.2.3 Metabarcoding

While the metabarcoding approach consists of several well established steps, the possible protocols performing these steps are diverse, each with its own biases (D'Amore et al., 2016). This non-uniformity makes it difficult to directly compare the results of different studies. In addition to the possible biases in sampling, DNA extraction and purification and exDNA content described above, PCR amplification and sequencing are a major source of bias (Knight et al., 2018; Pollock et al., 2018; Zinger et al., 2019).

5.2.3.1 PCR

PCR amplification is the pillar of metabarcoding studies as the goal is to target specific marker gene sequences (16S or 18S rRNA genes in this thesis) and amplify them to accumulate enough DNA content to proceed for sequencing. The PCR process in metabarcoding is therefore inherently linked with the sequencing platform and the PCR primers need to be in the appropriate range of the platform of choice. In addition, the PCR primers should target DNA regions relatively conserved across taxa, flanking an informative (*i.e.* variable) region of the marker gene sequence.

It is possible to test the performance of a different PFC primer/setup combinations through the use of mock communities – artificial microbial communities (for a ground truth dataset) which allow benchmarking and thus direct comparisons of protocols and tools[5]. With this approach, Fouhy et al. (2016) showed that the best taxonomic resolution was achieved with Illumina MiSeq platform (over Ion PGM) with PCR primers targeting the hypervariable V4–V5 region of the 16S rRNA genes. Following these recommendations, the metabarcoding approach in this thesis involved primers targeting the V4–V5 region of the 16S rRNA genes and the resulting amplicons were sequenced with Illumina MiSeq. However, inherent biases in primer selection make the direct comparison of results between studies problematic.

In addition to possible biases related to PCR primer selection, imperfections in the PCR pro-

[5]While mock communities are useful in benchmarking protocols and tools, note that good performance on characterizing mock communities is not guaranteed to translate to good performance on communities involving previously unknown taxa, which may be poorly matched by the primers.

cess can represent additional challenges. In theory, each amplicon is replicated at each step of the PCR cycle. However, in some cycles this replication would fail, bringing about two issues. First, the stochastic nature of these errors can result in bias especially when a sequence is present in low copy numbers: if a relatively rare variant failed to replicate at a very early cycle, its abundance after amplification would be much lower than another equally rare variant that replicated successfully. In the work presented in this thesis, we addressed this by performing multiple (5) PCR amplification replicates and pooling them for sequencing, reducing the bias accumulated in any one of the independently performed PCRs. Second, aborted polymerizations also introduce chimeras in the data. This happens because the short unfinished sequence can serve as a primer in the next PCR cycle, amplifying another amplicon (Haas et al., 2011). Downstream bioinformatic analyses can detect this and correct for it in some cases, because chimeras are constituted of fragments from the true amplicons. Indeed, I implemented a step of identifying and discarding chimeras in the pipeline.

With the development of new sequencing technologies such as SMS, it is possible to perform PCR-free metabarcoding, putting an end to primer design and limitations altogether. Indeed, SMS-based metabarcoding approaches have been shown to be suitable for precise taxonomic affiliation (Mosher et al., 2014; Benítez-Páez et al., 2016). This area is rapidly growing and since 2016, numerous protocols have been developed to treat such sequences (Singer et al., 2016; Earl et al., 2018; Callahan et al., 2019).

5.2.3.2 Copy number

As introduced in Section 1.2.2.2, and discussed in depth in the excellent review by Pérez-Cobas et al. (2020), the marker gene approach can be biased by the gene copy number as the latter may differ between species.

Ideally, the number of copies for a marker genes should be 1 per organism as it is the case with universal single copy genes (USiCG). However, metabarcoding takes advantage of the hypervariable regions flanked by highly conserved regions as is the case in the ubiquitous 16S or 18S rRNA genes, which are not single copy. According to the latest release (v5.6 from Octo-

ber 2019) of the ribosomal RNA operon database (rrnDB, (Klappenbach et al., 2001; Stoddard et al., 2015)), the current estimate of the number of 16S rRNA genes in bacterial species is 5 on average (median 5, range 1–21, estimated over 15486 genomes for 4568 species) and in archaea that number is 1.7 on average (median=1, range 1–5, estimated over 343 genomes for 261 species). Thus, the resultant OTU counts may not reflect the true abundance of a given species, but be biased towards the species with higher copy numbers. The relative abundances can therefore be off by a factor, sometimes reaching orders of magnitude. In eukaryotes, the copy number of the 18S rRNA marker gene has recently been investigated in planktonic ecosystems and ranged from 2 to 166 (over 7 species) and from 16 to 109 among different strains of the same species (Gong, Marchetti, 2019). This major bias puts into question the numerous previous characterizations of protist communities.

When the copy number is known, assuming each copy is effectively amplified by PCR to the same extent (ideally true, but may not be the case in practice), one may attempt to correct for this bias. Indeed, different tools have been developed to extrapolate the copy number from the copy numbers of known genomes from phylogenetically close taxa. However, a recent study detected a low accuracy of extrapolation methods, especially for novel organisms with large distance to previously sequenced genomes (Louca et al., 2018). As such inaccurate predictions can introduces additional noise in the data, the authors therefore recommended against corrections in metabarcoding analyses of microbial communities (Louca et al., 2018). Unfortunately, to date there is no solution to effectively overcome this limitation while still employing metabarcoding. This is one of the good reasons to interpret metabarcoding data cautiously as semi-quantitative results.

5.2.4 Metagenomics

Unlike metabarcoding, metagenomics has the advantage of being PCR-free. However, recently McLaren et al. (2019); Browne et al. (2020) have demonstrated that it is not bias-free. Apart from the methodological issues of biological nature described above, the main sources of bias in metagenomics are linked with the bioinformatics protocols and the workflow used to treat the

sequences, of which there are many (see Figure 1.11b from Pérez-Cobas et al. (2020)). For example, with the exception of low complexity environments, protist genomes are generally not well recovered from metagenomic datasets and specific tools were recently developed to overcome this limitation (West et al., 2018; Saary et al., 2020). While there are efforts to benchmark tools to assess the performance of different metagenomics protocols (Sczyrba et al., 2017), it may still be worthwhile to apply several approaches to one's dataset and compare the results to make an informed choice, especially for complex environments.

5.2.5 Metabolic inference

Automated bioinformatics methods are powerful tools to aid analysis. However, laborious manual checking is sometimes essential to an analysis. This was the case in the choice of genes for the analysis regarding metabolic inferences from metagenomic data, with the goal of retrieving the key players in the nutrient and carbon cycles, described in Chapter 3 of this manuscript. I used a pre-selected list of KEGG Orthologs (KOs) from Gutiérrez-Preciado et al. (2018) which I manually adapted using the KEGG online database. For each pathway, I selected the exclusive KOs, *i.e.* the ones that were not present in any other KEGG module, to create a list of exclusive KOs. This step could not have been automatized, because in some cases a single pathway was represented by multiple KEGG modules (for example, a gene may appear in 'nitrification' as well as 'complete nitrification'). Indeed, many standard markers for functions such as *AmoA* were detected in multiple KEGG modules and would therefore have been discarded were it not for manual verification.

5.3 Lake Baikal sediments

This section discusses the main findings of the studies presented in Chapters 3 and 4 of this manuscript. First, I focus on the topics explored by both the metabarcoding and metagenomic approaches, including the stability of microbial communities in different sites of the lake, as well as the community composition of prokaryotes and eukaryotes. Next, I discuss the insights the

metagenomics approach has provided into the ecosystem of the sediments upper layer. This is followed by a discussion of the potential of comparative studies into the communities residing in the sediments of different bodies of water and important considerations for meta-analyses.

5.3.1 Stability

The stability of the microbial diversity in surface layers of sediments at different sampling sites was one of the major results in our studies of lake Baikal (see Chapters 3 and 4). Importantly, this finding was reproduced in both the metabarcoding and the metagenomics approach, making it unlikely that particular limitations of either technique (e.g. PCR step in metabarcoding or the particular protocol used in metagenomics) were the source of this result.

First, we expected to find an effect of geographical basin. The basins are geographically distinct, with different incoming waters from rivers of presumably distinct ecosystems. Although possible subtle variations could have been missed due to lack of statistical power (n=4 or 5 sampling sites per basin), there were no striking differences in terms of microbial composition in the three basins, suggesting a relatively stable composition of the microbial communities across the lake. Such geographic stability has also been found in spatially dispersed deep ocean sediments (Hewson et al., 2007), although the authors reported a depth gradient.

We also expected to find a depth gradient of the diversity of the microbial communities residing in the studied sediments. Our sampling sites ranged from equivalent to shallow ponds to depths rarely achieved in freshwater lakes (our deepest sampling site, 1450m, is deeper that the maximal depth of all but two freshwater lakes). Indeed, the effect of the depth on sediments has been documented in freshwater lakes (samples from 0 to 93.5m depth) (Wu et al., 2019) as well as marine ecosystems (shallow samples 20-600m vs deep samples $\sim$3000m) (Fernanda Sánchez-Soto Jiménez et al., 2018). In general, in these cases depth can be considered as a proxy for multiple variables, including physiochemical variables. In particular, in the latter study, shallow and deep sediments clustered separately on a NMDS analysis, in which the depth fit mirrored the fits for other variables, such as the redox potential, sulfur concentration, and the percentage of clay.

Despite having samples from geographical clearly defined basins with different incoming waters or a wide range of depth from sea-like deep to shallow pounds we did not detect significant differences in the microbial communities. This stability is explained by the fact that local physicochemical features are very similar despite depth (unlike in many other oceanic waters) because, essentially, temperature is the same everywhere in the lake sediments due to the weak stratification and the strong down-welling event occurring in the lake after the freezing period and of 4°C. Temperature is a major determinant and so whichever the other differences, temperature overrides the rest. It is possible that nevertheless, the communities may be strongly affected by the physiochemical variables traditionally associated with depth but not strictly following a depth gradient in our samples. Unfortunately, those were not measured in our sampling campaign, leaving such questions open for future studies.

5.3.2 Prokaryotes

Overall, both approaches recovered a similar phylogenetic profile in our samples, with a large fraction of Archaea, up to ∼20%. The more relatively dominant groups of our sediment samples were the bacterial phyla PVC (Planctomycetes, Verrucomicrobia and Omnitrophica order equally distributed within), FCB (Latescibacteria, Ignavibacteria and Bacteroidetes), Chloroflexi and Acidobacteria as well the archaeal phyla TACK (mainly represented by Thaumarchaeota) and Euryarchaeota (Thermoplasmatales), recovered at similar percentages in both studies. With regards to archaea, our results are in agreement with previous studies showing the importance of these two archaeal phyla among the microbial communities in lake Baikal sediments associated with gas discharge (Bukin et al., 2016).

Intriguingly, we also recovered many DPANN OTUs using the metabarcoding approach but only few were recovered in the metagenomic dataset. One possible origin for this difference is the lack of representative DPANN lineages and, because they are known to be fast-evolving, possibly most of the fraction of unknown archaea should have been affiliated with DPANN. Nonetheless even speculating that all the share of unknown archaea is DPANN, the difference is still consequent. Another possible source of this discrepancy and metabarcoding and metage-

nomics results in general is the gene copy number as discussed in Section 5.2.3.2. However, most DPANN phyla are not represented in the rrnDB (Klappenbach et al., 2001; Stoddard et al., 2015), with the exception of Micraracheota for which this was not the case (in the rrnDB, there is 1 genome with a single copy), making it impossible to prove or disprove this hypothesis. Similarly, it is possible that the PCR primers used could be biased due to a higher affinity DPANN sequences, especially Woesearchaeal and Pacearcheal ones. This hypothesis is most likely to be false as PCR primers are known to have the opposite effect for some newly discovered lineages such as DPANN or Lokiarchaeota(Bahram et al., 2019). An alternative hypothesis for the discrepancy could be the small genome size of DPANN, which may make it less likely to be recovered by the metagenomic approach. However, Thaumarchaeota also have small genomes and they were nevertheless successfully recovered, which argues against this hypothesis. Finally, Dombrowski et al. (2019) recently reported that DPANN genomes have only around $\sim$70% of the completeness estimated by the standard tool checkM (Parks et al., 2015), which means that these genomes may lack $\sim$30% of the USiCGs typically used. Such bias but be resulting in us underestimating the DPANN abundance in the metagenomic dataset. It is impossible to conclude at this stage whether DPANN are overestimated by the metabarcoding approach or underestimated by metagenomics.

Moreover, the five upper centimeters of oxic seafloor sediments are typically rich environments with a high concentration of cells where recombination, uptake of exDNA, viral infection and protists feeding are likely to take place, with the main representatives of these communities members of Thaumarchaeota, Proteobacteria and Chloroflexi (Orsi, 2018). Our results paint a similar picture of the upper layer sediments of lake Baikal.

5.3.3 Protists

Unfortunately, the metagenomic approach recovered very little Eukaryotes, which reflects known limitations of current protocols (see Section 5.2.4) and their lower abundance than bacteria. New tools targeting eukaryotes specifically have been developed since and would be helpful in revealing more about these relatively neglected communities (West et al., 2018; Saary et al.,

2020).

In lake Baikal surface layer sediments, our metabarcoding approach revealed a dominance by the Stramenopiles and Alveolates. Surprisingly, the majority of Stramenopiles amplicons in our samples were affiliated to the Ochrophyta division, which is mostly known to represent golden-brown algae – photosynthetic organisms. While sinking materials could be a field contaminant to unveil the proper diversity, them accounting for such a major group might be unlikely as the vast majority of the exDNA in sediments should be of local origin (Section 5.2.2). However, many Ochrophyta lineages have undergone plastid reduction. Indeed, we recovered numerous Spumella- and Paraphysomonas- affiliated OTUs which are groups known to be exclusively non-photosynthetic (Dorrell et al., 2019). With regards to Alveolates, we found Dinoflagellates and closely related Syndiniales, of which most are parasites and known to accumulate in sediments as dinocysts or dinospores (López-García et al., 2007; Sauvadet et al., 2010). The other Alveolates were Ciliophora, which have long been known to be benthic marine colonizers; recently, Schoenle et al. (2017) reviewed their wide diversity.

5.3.4 Ecosystem functioning

In Chapter 4 of this manuscript, we inferred the main nutrient cycles in the lake Baikal sediments and the associated key players. In oxic marine sediments Thaumarchaeota are known to be key primary producers also oxidizing ammonia, and therefore depend on the ammonia regeneration by Proteobacteria and Chloroflexi (Orsi, 2018). In lake Baikal upper layer sediments, we also recovered those players as well as others like Rokubacteria and Nitrospirae. Rokubacteria (or previously “Candidate phylum Spring Alpine Meadow” - SPAM) have already been found in oxic sediments and recently, Becraft et al. (2017) highlighted their surprisingly large size for an oligotrophic inhabitant microbial species (6-8 Mbps). Still, relatively little is known about this phylum and comparative study of genomes/MAGs from databases along with novel studies could be welcome.

The taxa we found as the key players for nitrification and nitrite reduction are in line with the known ecosystem of the nitrogen cycle in marine environments (Pajares, Ramos, 2019).

5.3.5 A sea or a lake?

One may consider pressure, oligotrophy, pH and water temperature as important drivers for the composition of microbial communities. Indeed, these are important factors, but nevertheless, salinity appears to have a stronger impact on these communities according to studies on lake Baikal (Cabello-Yeves et al. (2017); Lomakina et al. (2020) and Chapters 3 and 4 of this manuscript). Indeed, Lake Baikal (see Section 1.3.2), also called the 'Sacred sea' has all the typical marine characteristics except salinity of its oligotrophic waters. Indeed, the lake is 1.6km deep at its maximum, contains gas and oil discharge zone and its own seal population (the only freshwater seals on Earth), features previously thought to be only linked to marine environments. Despite these similarities, very recent microbial ecology studies on lake Baikal waters and sediments show communities that differ considerably from the marine communities (Cabello-Yeves et al., 2017; Lomakina et al., 2020). Even when marine related taxa have been found, they were of relatively low abundance and never among the dominating taxa (Annenkova et al. (2020) and Chapter 3 of this manuscript).

As discussed previously, limitations in the metabarcoding and metagenomics approaches (such sequencing technology, PCR primers, metagenomics protocol) make comparisons difficult. Nevertheless, I attempted to compare the results I obtained to other freshwater, brackish or saline lakes as well as open ocean samples. To do so, I targeted studies with relatively similar approaches in their metabarcoding workflow as the one I used in my pipeline. One major limitation is that most studies do not provide tables with clear raw count / relative abundances results for different phyla or classes. Indeed, providing extensive supplementary tables with such results would have a great impact on the reproducibility and meta-analyses in the scientific community. It will allow other scientists to directly compare their results without having to redo the analyses or reach out to authors to email numbers.

The metagenomics analysis in Chapter 4 of this manuscript was initially performed as a part of a comparative study with data available for other environments. I have performed the comparisons with the relevant datasets available. However, surprisingly, there are very few metagenomic datasets of sediments available to download and use, as most of the available datasets are ei-

ther planktonic or sometimes, sediment cores. Due to the very low number of sampling sites per category which could be compared, I have chosen to not include these comparisons in this manuscript.

5.4 Perspectives

5.4.1 Investigating lake Baikal species adaptation

The environment in lake Baikal sediments is unique and it would be interesting to investigate the adaptations of microorganisms to this particular environment. In order to do so, the use of the recovered MAGs is fundamental. The work of this thesis has produced a wide diversity of recovered medium-to-high quality MAGs, awaiting in-depth analyses of groups of MAGs. The located strains could then be compared to other strains from different environments using comparative genomics. The outcome of such analyses would be an identification of all metabolic pathways and system functioning occurring in the recovered MAGs of lake Baikal and potentially provide scientific knowledge on their taxonomic groups and contribute to future ecological or evolutionary investigations.

The new technology of single cell sequencing makes an alternative approach possible. One may perform cell sorting and single-cell genomics approaches, which would lead to identification of the important traits and potentially result in the complete genome.

5.4.2 Gaining insight into freshwater-marine transition

In order to gain insight into the freshwater-marine transition topic through lake Baikal, the experimental protocol of Wang et al. (2012) is unsurpassed. Indeed, Wang et al. (2012) performed, to my knowledge, the only study in which the authors themselves sampled the freshwater, brackish and saline sites using the same equipment under the same conditions. Then, with a comparable sampling and molecular biology, the potential differences due to unequal bias as introduced by the methodology, is greatly reduced and conclusions can be more easily drawn.

As most species comprising a community are each present in numbers, metabarcoding still offers the best solution to unveil the entire diversity and being able to infer phylogenetic trees of the related species between the samples or with the OTUs compiled from multiple origins. Nonetheless, recent metagenomic studies (Cabello-Yeves et al., 2017; Cabello-Yeves, Rodriguez-Valera, 2019; Cabello-Yeves et al., 2019) have managed to successfully identified and recovered such low abundance marine taxa in lake Baikal, showing that metagenomic approach might still hold promise even in investigating this topic.

Eukaryotes are typically excluded from most metagenomics approaches, but the newly developed tools (West et al., 2018; Saary et al., 2020) for metagenomic data hold promise to recover protist genomes. Recovering marine-like protist genomes through this approach or, alternatively, by the use of single-cell genomics, would be of great interest.

5.4.3 Assessing the diversity of active microbes

As we discussed earlier, neither metabarcoding nor metagenomic approaches can discriminate exDNA or iDNA in a total DNA pool as was the case in our studies as performed classical protocols using the power soil kit from Qiagen. The fact that exDNA can represent a subsequent share of the observed diversity, a good control for future studies would be to use the protocol described by Lever et al. (2015). In the one hand, this would allow us to overcome the bias induced by exDNA and observe the real fraction of cells actively thriving or in dormance in the sediments. On the other hand, one could in addition explore the composition of exDNA representing dead cells. Both the proportion and the content of exDNA may be very different in the oligotrophic environment of deep freshwater sediments in lake Baikal than the shallow sea sediments described by Torti et al. (2018).

Another solution would be the use of metatranscriptomic approach. Such an approach would allow the identification of the active cells in the sediment and the possibility to confirm or find new players in the nutrient cycle happening. One drawback here despite the relative stability over season suspected in lake Baikal upper layer sediments would be that metatranscriptomic do recover only the active player at the specific time of sampling, possibly over-representing or-

ganisms in growth, which can represent a bias in a real thriving ecosystem. Therefore, temporal series are needed to have an objective point of view on seasonal and always active players, but this increase significantly the sampling and molecular biology costs.

5.4.4 Other interesting points

5.4.4.1 Core sediment

Recent history on sediment analyses using metagenomics revealed numerous new taxonomic groups at different levels on the tree of life (Biddle et al., 2008; Rinke et al., 2013; Baker et al., 2015, 2016; Spang et al., 2015; Solden et al., 2016; Seitz et al., 2019). A significant part of these studies were based on marine sediments from below the sea floor. As Baikal is the only freshwater lake on earth with gas hydrate at high-depth similar to marine environments, sampling core sediments to investigate potential phylogenetic and metabolic novelties seem promising. Ideally, multiple cores could be taken, from which biological replicates could be sampled, building a robust dataset to accomplish these goals.

5.4.4.2 Viral communities

Using currently available and already processed data from our metagenomics datasets used in the Chapter 4 of this thesis, or by sequencing additional ones, the viral fraction of lake Baikal sediments is a topic of interest. Recent work by Coutinho et al. (2020) explores the viral communities in the water column of lake Baikal, and a comparison with sediments would be interesting.

APPENDIX

A

FIRST APPENDIX

ARTICLES
https://doi.org/10.1038/s41559-019-1005-0
nature ecology & evolution

Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area

Jodie Belilla[1], David Moreira[1], Ludwig Jardillier[1], Guillaume Reboul[1], Karim Benzerara[2], José M. López-García[3], Paola Bertolino[1], Ana I. López-Archilla[4] and Purificación López-García[1]*

Microbial life has adapted to various individual extreme conditions; yet, organisms simultaneously adapted to very low pH, high salt and high temperature are unknown. We combined environmental 16S/18S ribosomal RNA gene metabarcoding, cultural approaches, fluorescence-activated cell sorting, scanning electron microscopy and chemical analyses to study samples along such unique polyextreme gradients in the Dallol–Danakil area in Ethiopia. We identified two physicochemical barriers to life in the presence of surface liquid water defined by (1) high chaotropicity–low water activity in Mg^{2+}/Ca^{2+}-dominated brines and (2) hyperacidity–salt combinations (pH ~0/NaCl-dominated salt saturation). When detected, life was dominated by highly diverse ultrasmall archaea that were widely distributed across phyla with and without previously known halophilic members. We hypothesize that a high cytoplasmic K^+-level was an original archaeal adaptation to hyperthermophily, subsequently exapted during several transitions to extreme halophily. We detect active silica encrustment/fossilization of cells but also abiotic biomorphs of varied chemistry. Our work helps circumscribing habitability and calls for cautionary interpretations of morphological biosignatures on Earth and beyond.

Microbial life has adapted to so-called extreme values of temperature, pH or salinity, but also to several polyextreme, for example hot acidic or salty alkaline, ecosystems[1,2]. Various microbial lineages have been identified in acidic brines in the pH range 1.5–4.5, for example in Western Australia[3,4] and Chile[3]. However, although some acidophilic archaea thrive at pH ~0 (*Picrophilus oshimae* grows at optimal pH 0.7)[5] and many halophilic archaea live in hypersaline systems (>30% weight/volume; NaCl-saturation conditions), organisms that adapted simultaneously to very low pH (<1) and high salt, and eventually also high temperatures, are not known among cultured prokaryotic species[1]. Are molecular adaptations to these combinations incompatible or are (hot) hyperacidic hypersaline environments simply rare and unexplored? The Dallol geothermal dome and its surroundings (Danakil Depression, Afar, Ethiopia) allow this question to be addressed by offering unique polyextreme gradients combining high salt content (33 to >50%; either Mg^{2+}/Ca^{2+} or Na^+(/$Fe^{2+/3+}$)-rich), high temperature (25 to 110 °C) and low pH (≤−1.5 to 6.0).

Dallol is an uplifted (~40 m) dome structure located in the north of the Danakil Depression (~120 m below sea level). The Danakil Depression is a 200-km-long basin within the Afar Rift at the junction between the Nubian, Somalian and Arabian Plates[6]. Lying only 30 km north of the hypersaline, hydrothermally influenced Lake Assale (Karum) and the Erta Ale volcanic range, Dallol does not display volcanic outcrops but intense degassing and hydrothermalism. These activities are observed on the salt dome and the adjacent Black Mountain and Yellow Lake (Gaet'Ale) areas[6,7] (Fig. 1a,b). Gas and fluid isotopic measurements indicate that meteoritic waters, notably infiltrating from the high Ethiopian plateau (>2,500 m), interact with an underlying geothermal reservoir (280–370 °C)[7,8]. Further interaction of those fluids with the 1-km thick marine evaporites filling the Danakil Depression results in unique combinations of polyextreme conditions and salt chemistries[6,7,9,10], which have led some authors to consider Dallol as a Mars analogue[11].

Here, we use environmental 16S/18S ribosomal RNA gene metabarcoding, cultural approaches, fluorescence-activated cell sorting (FACS) and scanning electron microscopy (SEM) combined with chemical analyses to explore microbial occurrence, diversity and potential fossilization along Dallol–Danakil polyextreme gradients[12–15].

Results and discussion

To investigate the distribution and, eventually, type of microbial life along those polyextreme gradients, we analysed a large variety of brine and mineral samples collected mainly from two field expeditions (January 2016 and 2017; a few additional samples were collected in 2018) in four major zones (Fig. 1 and Extended Data Figs. 1–3). The first zone corresponded to the hypersaline (37–42%) hyperacidic (pH between ~0 and −1; values down to pH −1.6 were measured on highly concentrated and oxidized brines on site) and sometimes hot (up to 108 °C) colourful ponds at the top of the Dallol dome (Fig. 1c and Extended Data Figs. 1a, 2a–h and 3). The second zone consisted of the salt canyons located at the southwestern extremity of the Dallol dome and the Black Mountain area, which includes the Black Lake (Fig. 1b,d and Extended Data Figs. 1b,c and 2l–q). Brine samples collected in a cave reservoir (Gt samples) and in ephemeral pools with varying degrees of geothermal influence at the dome base (PS/PS3 samples) were hypersaline (~35%), with moderate temperature (~30 °C) and acidity (pH ~4–6). By contrast, pools located near the small (~15 m diameter), extremely hypersaline (>70%), hot (~70 °C) and acidic (pH ~3) Black Lake were slightly more acidic (pH ~3), warmer (40 °C) and hypersaline (35–60%) than the dome-base pools (PSBL samples; Extended Data Fig. 3). The third zone corresponded to the Yellow Lake and neighbouring ponds (Fig. 1e and Extended Data Figs. 1d and 2i–k), which

[1]Ecologie Systématique Evolution, CNRS, Université Paris-Sud, Université Paris-Saclay, AgroParisTech, Orsay, France. [2]Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, CNRS, Sorbonne Université, Muséum National d'Histoire Naturelle, Paris, France. [3]Instituto Geológico y Minero de España, Palma de Mallorca, Spain. [4]Departamento de Ecología, Universidad Autónoma de Madrid, Madrid, Spain. *e-mail: puri.lopez@u-psud.fr

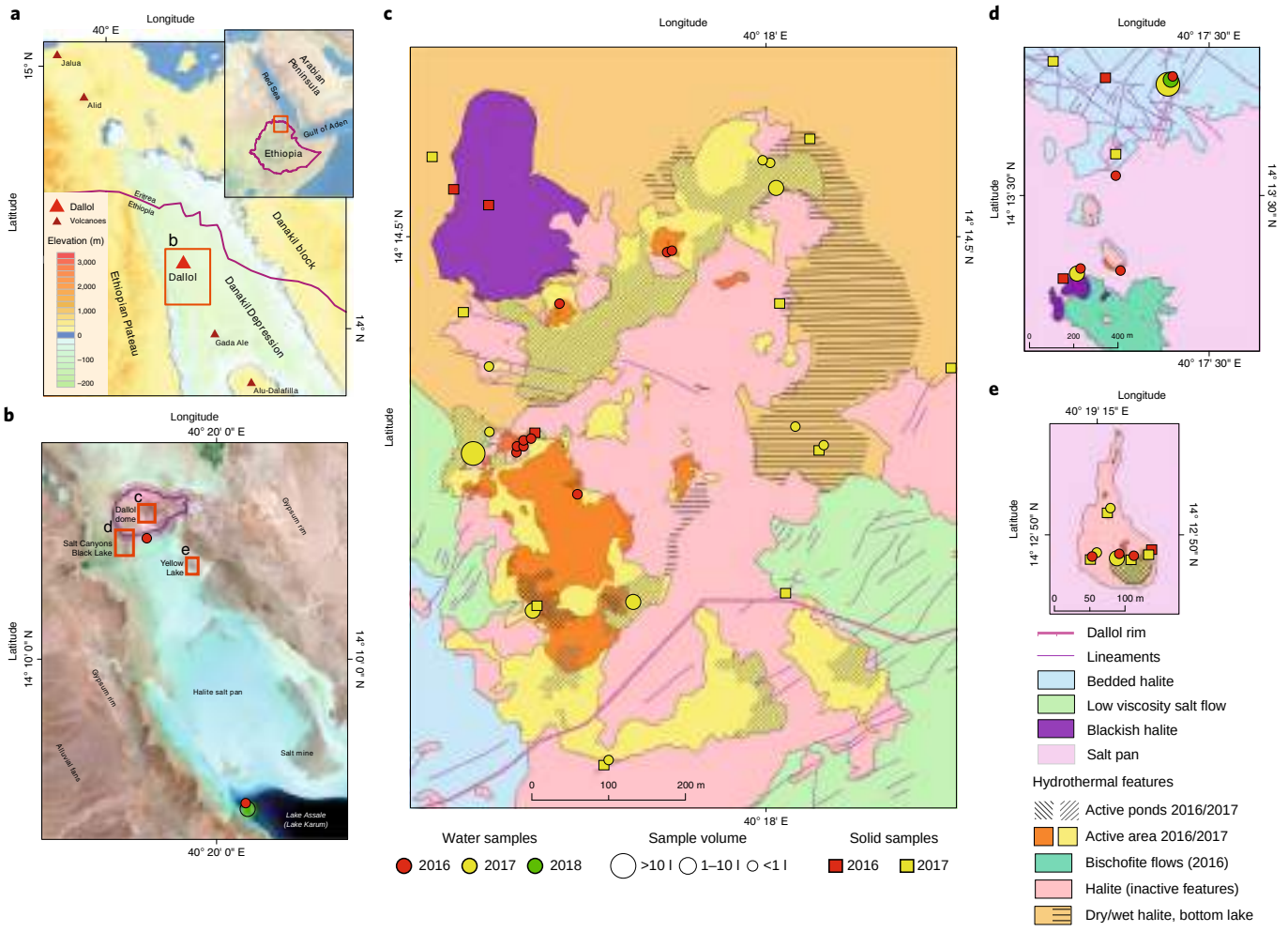


Fig. 1 | Overview of sampling sites at the polyextreme geothermal field of Dallol and its surroundings in the Danakil Depression, Ethiopia. a, Location of the Dallol dome area in the Danakil Depression following the alignment of the Erta Ale volcanic range (Gada Ale, Alu-Dalafilla), Northern Ethiopia. **b**, Closer view of the sampling zones in the Dallol area and Lake Assale or Karum (satellite image from Copernicus Sentinel 1; 19 January 2017). **c**–**e**, Geological maps showing the sampling sites at the Dallol dome summit (**c**), west salt canyons and Black Mountain, including the Black Lake (**d**) and Yellow Lake (Gaet'Ale) zone (**e**). Squares (solid samples) and circles (liquid samples) indicate the nature of the collected samples. The colour indicates the collection date (red, 2016; yellow, 2017; green, 2018). The size of circle is proportional to the collected brine volume for analyses. Specific sample names are indicated in the aerial view shown in Extended Data Fig. 1.

were acidic (pH ~1.8), warm (~40 °C) and extremely hypersaline[16] (≥50%). The Yellow Lake actively bubbles and emits toxic gases for animals, as illustrated by the presence of numerous dead birds. The gas phase includes light hydrocarbons[8]. The fourth zone consisted of the hypersaline (36%), almost neutral (pH ~6.5), Lake Assale (Fig. 1b and Extended Data Fig. 2r), which we used as a milder, yet extreme Danakil system for comparison. In contrast to a continuous degassing activity, the hydrothermal manifestations were highly dynamic, particularly on the dome and the Black Mountain area. The area affected by hydrothermal activity in January 2017 was much more extensive than the previous year (Fig. 1 and Extended Data Fig. 1). Dallol chimneys and hyperacidic ponds can appear and desiccate in a matter of days or weeks, generating a variety of evaporitic crystalline structures observable in situ[17]. Similarly, very active and occasionally explosive (salt 'bombs') hydrothermal activity that was characterized by hot (110 °C), slightly acidic (pH ~4.4) black hypersaline fluids was detected in the Black Mountain area in 2016 ('Little Dallol'; sample BL6–01; Extended Data Figs. 1b and 2l) but not in the following years. Active bischofite flows[6,7,18] (116 °C) were also observed in the Black Mountain area in 2016 but not in 2017.

To assess potential correlations between microbial life and local chemistry, we analysed the chemical composition of representative samples used in parallel for microbial diversity analyses (see Methods). Our results revealed three major types of solution chemistry depending on the dominant elements (Fig. 2f and Extended Data Fig. 4a). In agreement with recent observations, Dallol ponds were characterized by NaCl-supersaturated brines that were highly enriched in iron with different oxidation states, which explained the colour variation[17]. Potassium and sulfur were also abundant (Supplementary Table 1). By contrast, samples from the salt canyons and plain near Dallol and Lake Assale were NaCl-dominated with a much lower iron content, and the Yellow and Black lakes and associated ponds had very high Mg^{2+} and Ca^{2+} concentrations (Supplementary Table 1). Many aromatic compounds were identified, particularly in Dallol and Yellow Lake fluids (Supplementary Table 2). High chaotropicity associated with Mg_2Cl-rich brines, high ionic strength and low water activity (a_w) is thought to be a limiting factor for life[12,13,19,20]. We therefore determined these parameters in representative samples (Extended Data Fig. 5). Based on our experimental measures and theoretical calculations from dominant

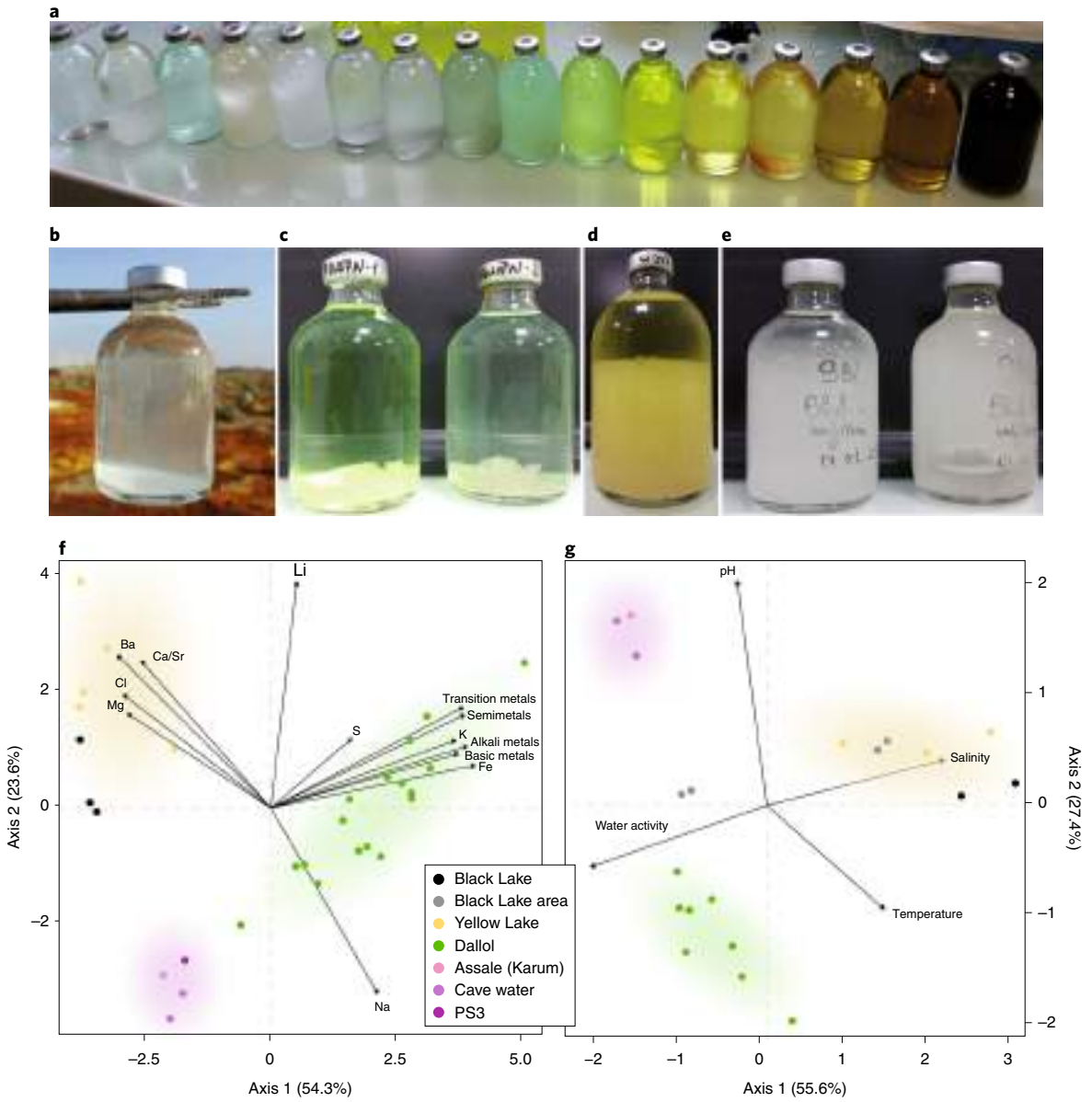


Fig. 2 | Physicochemical features of liquid samples from the Dallol area. a, Examples of colours displayed by the different samples analysed in this study, reflecting different chemistries and oxidation states. **b–e**, Examples of salt-oversaturated samples. Immediate (seconds) precipitation of halite crystals as water from a hot spring (108 °C) cools down upon collection (**b**), salt precipitates forming after storage at ~8 °C in water collected from Dallol hyperacidic ponds (**c**), Yellow Lake (**d**) and Black Lake (**e**). **f**, PCA of 29 samples according to their chemical composition (see Supplementary Table 2). Transition metals, Cr, Mo, Mn, Sc, Zn, V, U, Ce, La, Cu; semimetals, As, B, Sb, Si; alkali metals, Rb, Cs; basic metals, Tl, Al, Ga, Sh. Some elements are highlighted out of these groups owing to their high relative abundance or to their distant placement. A PCA showing all individual metal variables can be seen in Supplementary Fig. 3a. **g**, PCA of 21 samples and key potentially life-limiting physicochemical parameters in the Dallol area (temperature, pH, salinity, water activity). Water activity and salinity-related parameters are provided in Extended Data Fig. 5. Coloured zones in PCA analyses highlight the groups of samples corresponding to the three major chemical zones identified in this study.

salts, only samples in the Yellow and Black Lake areas displayed life-limiting chaotropicity and a_w values according to established limits[12,13,19,20]. A principal component analysis (PCA) showed that the sampled environments were distributed in three major groups depending on solution chemistry, pH and temperature: Black and Yellow Lake samples, anticorrelating with a_w; Dallol dome samples, mostly correlating with a_w but anticorrelating with pH; and Dallol canyon cave reservoir (Gt samples) and Lake Assale, correlating with a_w and pH (Fig. 2g). These results are consistent with those obtained with analysis of variance and subsequent post-hoc analysis, which show significant differences between the three major chemical zones (coloured areas in Fig. 2f,g) among them for the variables tested (Supplementary Table 4).

To ascertain the occurrence and diversity of microbial life along these physicochemical gradients, we purified DNA from a broad selection of brine samples (0.2–30 µm size fraction) and solid samples (gypsum and halite-rich salt crusts, compacted sediment and soil-like samples; Extended Data Fig. 3). We carried out 16S/18S rRNA gene-based diversity studies by high-throughput short-amplicon sequencing (metabarcoding approach) but also sequenced almost-full length genes from clone libraries, providing local reference sequences for more accurate phylogenetic

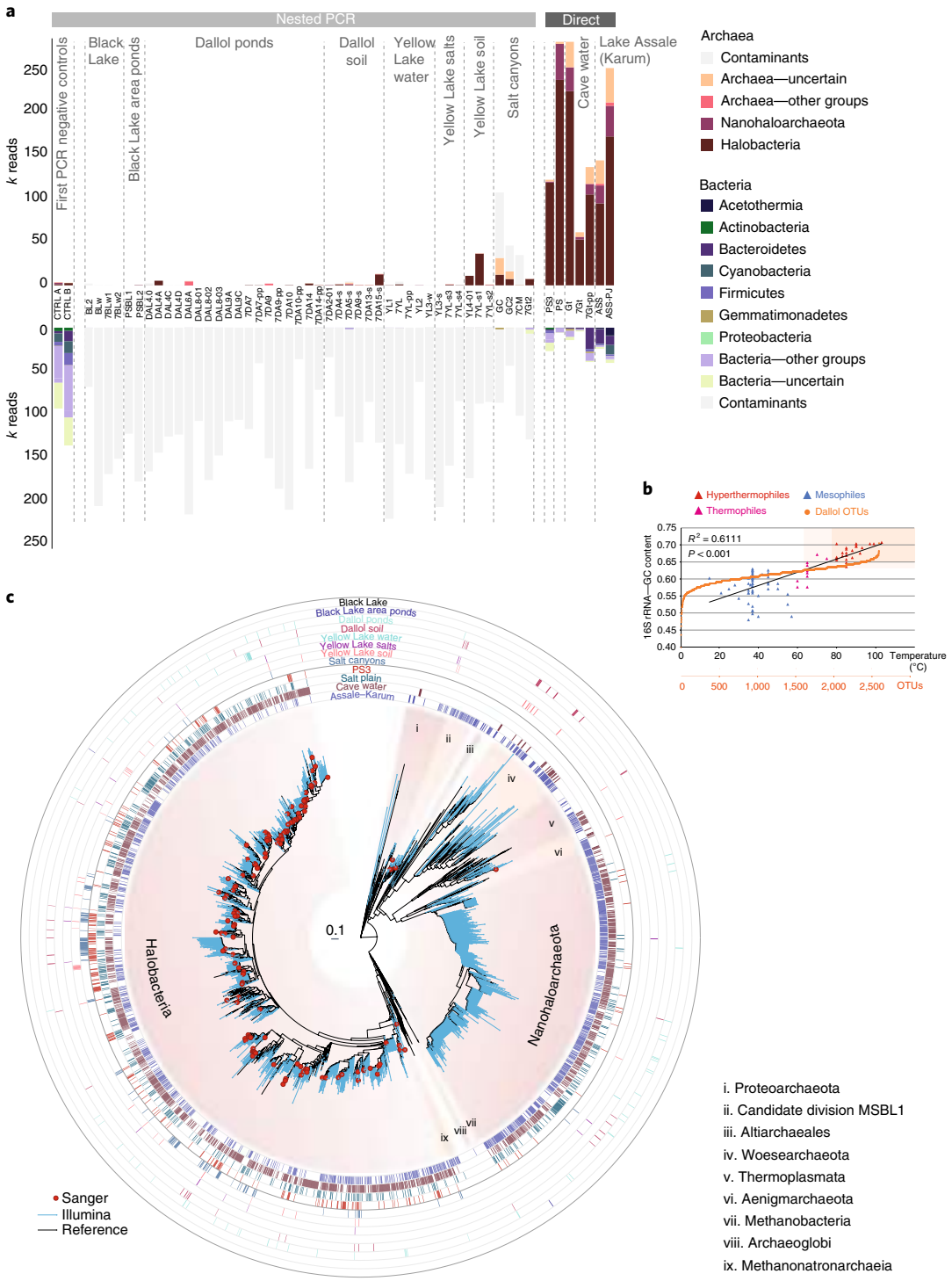


Fig. 3 | Distribution and diversity of prokaryotes in samples from the Dallol dome and surrounding areas based on 16S rRNA gene metabarcoding data. a, Histograms showing the presence/absence and abundance of amplicon reads of archaea (upper) and bacteria (lower) obtained with universal prokaryotic primers. Samples yielding amplicons directly (negative PCR controls were negative) are shown on the right (direct). Samples for which amplicons were only obtained after nested PCR, all of which also yielded amplicons in 'negative' controls, are displayed on the left (nested PCR). Sequences identified in the 'negative' controls, considered as contaminants, are shaded in light grey in the corresponding Dallol samples. The phylogenetic affiliation of dominant archaeal and bacterial groups is colour coded. For details, see Supplementary Table 5. k reads, thousand reads. The names of the different samples are provided on the *x* axis. **b**, GC content of archaeal OTUs plotted on a graph showing the positive correlation of GC content (for the same 16S rRNA region) and growth temperature of diverse described archaeal species. **c**, Phylogenetic tree of archaeal 16S rRNA gene sequences showing the phylogenetic placement of archaeal OTUs identified in the different environmental samples (full tree provided in Supplementary Data 1). Sequences derived from metabarcoding studies are represented with blue branches (Illumina sequences); those derived from cloning and Sanger sequencing of environmental samples, cultures and FACS-sorted cells are labelled with a red dot. Reference sequences are in black. Concentric circles around the tree indicate the presence/absence of the corresponding OTUs in different groups of samples (groups shown in (**a**)).

analyses (see Methods). Despite intensive PCR efforts and extensive sampling in Dallol polyextreme ponds, including pools that were active in two consecutive years (Extended Data Fig. 1) to minimize ephemeral system-derived effects, we only amplified 16S/18S rRNA genes from Dallol canyon cave water, the dome-base geothermally influenced salt plain and Lake Assale, but never from the Dallol dome or Black/Yellow lakes (Fig. 3a). To check whether this resulted from excessively low DNA amounts in those samples (although relatively large volumes were filtered), we carried out seminested PCR reactions using, as templates, potential amplicons produced during the first PCR–amplification reaction, including the first PCR negative controls. Almost all samples produced amplicons in seminested PCR reactions, including the first PCR blanks (Fig. 3a). Metabarcoding analysis revealed that amplicons from direct PCR reactions (PS/PS3, Gt, Assale) were largely dominated by archaeal sequences (>85%) grouping in diverse and abundant operational taxonomic units (OTUs) (Extended Data Fig. 6). By contrast, amplicons derived from Dallol ponds, Black and Yellow lakes and also first PCR 'negative' controls were dominated by bacterial sequences. Most of them were related to well-known molecular biology kit and laboratory contaminants[21,22], whilst others were human-related bacteria probably introduced during intensive afar and tourist daily visits to the site. A few archaeal sequences might also result from aerosol cross-contamination, despite extensive laboratory precautions (see Methods). After the removal of contaminant sequences (grey bars in Fig. 3a, and Supplementary Table 5), only rare OTUs encompassing few reads (mostly archaeal) could be associated with Dallol dome or Yellow Lake brines, which we interpret as dispersal forms (dusty wind is frequent in the area). Slightly higher abundances of archaeal OTUs were identified in 'soil' samples, that is samples retrieved from salty consolidated mud or crusts where dust brought by the wind from the surrounding plateaux accumulates and starts constituting a proto-soil (with incipient microbial communities; for example, Extended Data Fig. 2i). Therefore, although we cannot exclude the presence of active life in these 'soil' samples, our results strongly suggest that active microbial life is absent from polyextreme Dallol ponds and the Black and Yellow lakes.

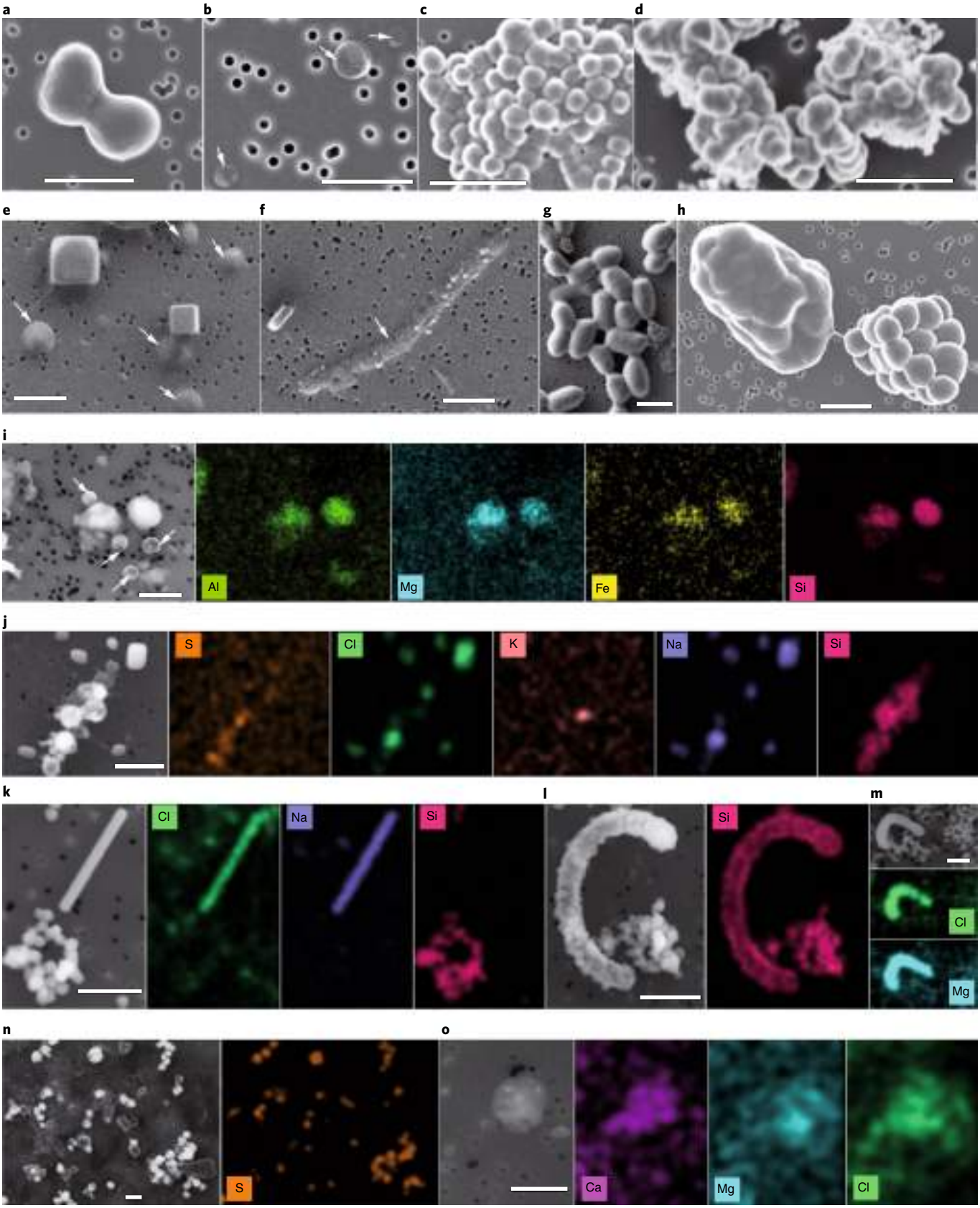
By contrast, PS/PS3, Gt and Assale samples harboured extremely diverse archaea (2,653 OTU conservatively determined at 95% identity, that is genus level) that virtually spanned the known archaeal diversity (Fig. 3, Extended Data Fig. 6 and Supplementary Table 5). Around half of that diversity belonged to Halobacteria, and an additional quarter to the Nanohaloarchaeota[23]. The rest of archaea distributed in lineages typically present in hypersaline environments, for example, the Methanonatronoarchaeia[24,25] and Candidate Division MSBL1, which is thought to encompass methanogens[26] and/or sugar-fermentors[27]. However, they also included other archaeal groups not specifically associated with salty systems (although they can sometimes be detected in hypersaline settings, for example some Thermoplasmata or Woesearchaeota). These included Thermoplasmata and Archaeoglobi within Euryarchaeota, Woesearchaeota and other lineages (Aenigmarchaeota, Altiarchaeales) usually grouped as DPANN[28–30], and Thaumarchaeota and Crenarchaeota (Sulfolobales) within the TACK/Proteoarchaeota[31] (Fig. 3a and Supplementary Table 5). In addition, because rRNA GC content correlates with growth temperature, around 27% and 6% of archaeal OTUs were inferred to correspond to thermophilic and hyperthermophilic organisms, respectively (see Methods; Fig. 3b). As previously observed[23,28,29], common archaeal primers for near-full 16S rRNA genes (Fig. 3c, red dots) failed to amplify Nanohaloarchaeota and other divergent DPANN lineages. These probably encompass ectosymbionts or parasites[28–30,32]. Given their relative abundance and co-occurrence in these and other ecosystems, it is tempting to suggest that Nanohaloarchaeota are (ecto)symbionts of Halobacteria, and Woesearchaeota could potentially be associated with *Thermoplasma*-like archaea. Although much less abundant, bacteria belonging to diverse phyla, including CPR (Candidate Phyla Radiation) lineages, were also present in these samples (710 OTUs; Extended Data Figs. 6 and 7 and Supplementary Table 5). In addition to typical extreme halophilic genera (for example, *Salinibacter*, Bacteroidetes), one Deltaproteobacteria group and two divergent bacterial clades were overrepresented in Dallol canyon Gt samples. Eukaryotes, which were less abundant and diverse, were present in Lake Assale and occasionally in the salt plain and Gt. They were dominated by halophilic *Dunaliella* algae (Extended Data Fig. 8 and Supplementary Table 6).

Consistent with metabarcoding results, and despite the use of various culture media and growth conditions mimicking local environments (see Methods), cultural approaches did not yield enrichments for the Dallol dome, Black Lake and Yellow Lake samples. We obtained enrichments from the canyon cave (Gt/7Gt) and salt plain (PS/PS3) samples in most culture media (except in benzoate/hexadecane) and tested conditions (except at 70 °C in the dark). However, all attempts to isolate microorganisms at pH <3 from these enrichments failed. The most acidophilic isolate obtained from serial dilutions (PS3-A1) only grew at 37 °C and optimal pH 5.5 (range pH 3–7). Its 16S rRNA gene was 98.5% identical to that of *Halarchaeum rubridurum* MH1-16-3 (NR_112764), an acidophilic haloarchaeon growing at pH 4.0–6.5 (ref. [33]).

In agreement with metabarcoding and culture-derived observations, multiparametric fluorescence analysis showed no DNA fluorescence above background for Dallol and Yellow Lake samples (Extended Data Fig. 9). Because optical and SEM observations suggested that indigenous cells were unusually small, we applied FACS to samples from the different Dallol environments (Extended Data Fig. 3), followed by systematic SEM analysis of sorted events. Despite some samples showed no difference in fluorescence after incubation with DNA dyes, we sorted all events above background limit (as defined in Extended Data Fig. 9a). We only detected cells in Dallol cave water and salt plain samples, but not in Dallol dome ponds or Yellow Lake samples (Extended Data Fig. 9). Consistent with this, after DNA purification of FACS-sorted particles, 16S rRNA gene amplicons could only be obtained from different cave

Fig. 4 | SEM pictures and chemical maps of cells and abiotic biomorphs identified in samples from the Dallol region. a–h, SEM pictures of cells (**a**–**c**,**e**–**h**) and abiotic biomorphs (**d**). FACS-sorted dividing cells from sample PS (hydrated salt pan between the Dallol dome base and the Black Lake) (**a**); FACS-sorted ultrasmall cells from 7Gt samples (cave water reservoir, Dallol canyons) (**b**); FACS-sorted colony of ultrasmall cells from sample PS (note cytoplasmic bridges between cells) (**c**); FACS-sorted abiotic silica biomorphs from the Dallol pond 7DA9 (note the similar shape and morphology compared to cells in **c**) (**d**); cocci and halite crystals in 8Gt samples (cave water) (**e**); long rod in 8Gt (**f**); FACS-sorted cells from Gt samples (**g**) and FACS-sorted colonies from sample PS (note the bridge between one naked colony and one colony covered by an exopolymeric-like matrix) (**h**). **i**–**o**, SEM images and associated chemical maps of cells and biomorphs. Colour intensity provides semiquantitative information of the mapped elements. Small cocci and amorphous Al–Mg–Fe-rich silicate minerals from Gt (**i**); NaCl crystals and S–Si rich abiotic biomorphs from Dallol pond sample 7DA7 (**j**); NaCl crystal and Si biomorphs (**k**); Si-encrusted cell and Si biomorphs in sample 8Ass (Lake Assale) (**l**); Mg–Cl biomorph in sample BLPS_04 (Black Lake area pond) (**m**); S-rich biomorphs in Dallol pond 7DA9 (**n**) and Ca–Mg–Cl biomorph in YL-w2 (Yellow Lake pond) (**o**). SEM photographs were taken using In Lens or AsB detectors. For additional images and SEM details, see Supplementary Figs. 1 and 2. White arrows indicate cells difficult to recognize due to their small size and/or flattened aspect, which may result from sample preparation and/or high vacuum conditions within the SEM. Scale bars, 1 µm.

and salt plain samples but not from Dallol dome or Yellow Lake samples. Cell counts estimated from FACS for the cave and salt plain samples were low (average 3.1×10^4 cells ml^{-1} and 5.3×10^4 cells ml^{-1} for the cave and PS samples, respectively). Sorted cells were usually small to ultrasmall (down to 0.25–0.3 µm diameter; Fig. 4). In PS samples, some of these small cells formed colonies (Extended Data Fig. 9 and Fig. 4c), which were sometimes surrounded by an exopolymeric matrix cover (Fig. 4h). The presence of cytoplasmic



bridges and/or potential cell fusions (Extended Data Fig. 9 and Fig. 4c) suggest that they might be archaeal colonies[34].

FACS-sorted fluorescent particles in Dallol pond samples appeared to correspond exclusively to salt crystals or cell-sized amorphous minerals morphologically resembling cells, that is biomorphs[35,36] (for example, Fig. 4d compared with Fig. 4c). This prompted us to carry out a more systematic search for abiotic biomorphs in our samples. SEM observations coupled with chemical mapping by energy-dispersive X-ray spectrometry (EDXS) showed a variety of cocci-like biomorph structures of diverse elemental compositions. Many of them were Si biomorphs (Dallol ponds, Yellow Lake and Assale Lake), but we also detected Fe–Al silicates (Gt), S or S-rich biomorphs (Dallol ponds), and Ca or Mg chlorides (Yellow Lake, BLPS samples) (Fig. 4, Extended Data Fig. 10 and Supplementary Figs. 1 and 2). We also observed Si-encrusted rod-shaped cells in Lake Assale samples (Fig. 4l). Therefore, silica-rounded precipitates represent ultrasmall cell-like biomorphs in samples with no detectable life but they contribute to cell encrustment and potential fossilization when life is present.

Our work has three major implications. First, by studying the microbial distribution along gradients of polyextreme conditions in the geothermal area of Dallol and its surroundings in the Danakil Depression, we identify two major physicochemical barriers that prevent life from thriving in the presence of liquid water on Earth and, potentially, elsewhere[14], despite the presence of liquid water at the surface of a planet being a widely accepted criterion for habitability. In line with previous studies[12,13,19,20], one barrier is imposed by high chaotropicity and low a_w, which are associated with high Mg^{2+}-brines in the Black Lake and Yellow Lake areas. The second barrier seems to be imposed by the hyperacid–hypersaline combinations found in the Dallol dome ponds (pH ~0; salt >35%), regardless of temperature. This suggests that molecular adaptations to simultaneous very low pH and high salt extremes are incompatible beyond those limits. In principle, more acidic proteins, intracellular K^+ accumulation ('salt-in' strategy) or internal positive membrane potential generated by cations or H^+/cation antiporters serve both acidophilic and halophilic adaptations[37–39]. However, membrane stability/function problems and/or high external Cl^- concentrations that induce H^+ and cation (K^+/Na^+) import, potentially disrupting membrane bioenergetics[38], might be deleterious under these conditions. We cannot exclude other explanations linked to the presence of several stressors, such as high metal content or an increased susceptibility to the presence of local chaotropic salts in the Dallol hyperacidic ponds even if measured chaotropicity values are relatively low (−31 to +19 kJ kg^{-1}) compared to the established limit for life (87.3 kJ kg^{-1})[12,13,20] (Extended Data Fig. 6). Future studies should help to identify the molecular barriers limiting the adaptation of life to this combination of extremes. Second, although extreme environments are usually low-diversity systems, we identify exceptionally diverse and abundant archaea spanning known major taxa in hypersaline, mildly acidic systems near life-limiting conditions. A wide archaeal (and to a lesser extent, bacterial) diversity seems consistent with suggestions that NaCl-dominated brines are not as extreme as previously thought[40] and, with recent observations that the mixing of meteoric and geothermal fluids leads to hyperdiverse communities[41]. Nonetheless, life under high salt conditions requires extensive molecular adaptations[12,13,19,40], which might seem at odds with several independent adaptations to extreme halophily across archaea. Among those adaptations, the intracellular accumulation of K^+ ('salt-in' strategy), together with the corresponding adaptation of intracellular proteins to function under those conditions, has been crucial. Based on the observation that the deepest archaeal branches correspond to (hyper)thermophilic lineages[42] and that nonhalophilic hyperthermophilic archaea accumulate high intracellular K^+ (1.1–3 M) for protein thermoprotection[43,44] (thermoacidophiles also need K^+ for pH homeostasis[38]), we hypothesize that intracellular K^+ accumulation is an ancestral archaeal trait linked to thermophilic adaptation that has been independently exapted in different taxa for adaptation to hypersaline habitats. Finally, the extensive occurrence of abiotic, mostly Si-rich, biomorphs mimicking the simple shape and size of ultrasmall cells in the hydrothermally influenced Dallol settings reinforces the equivocal nature of morphological 'microfossils'[35] and calls for the combination of several biosignatures before claiming the presence of life on the early Earth and beyond.

Methods

Sampling and measurement of physicochemical parameters on site. Samples were collected during two field trips in January 2016 and January 2017 (when air temperature rarely exceeded 40–45 °C); a few additional samples were collected in January 2018 (Fig. 1 and Extended Data Figs. 1 and 3). All sampling points and mapping data were georeferenced using a Trimble handheld global positioning system (Juno SB series) equipped with Environmental Systems Research Institute software ArcPad 10. Cartography of hydrogeothermal activity areas was generated using Environmental Systems Research Institute GIS ArcMap mapping software ArcGis 10.1 over georeferenced Phantom-4 drone images taken by O. Grunewald during field campaigns, which was compared with and updated previous local geological cartography[7]. Samples were collected from three major areas at the Dallol dome and its vicinity (Fig. 1b): (1) the top of the Dallol dome, consisting of various hydrothermal pools with diverse degrees of oxidation (Fig. 1c); (2) the Black Mountain area (Fig. 1d), including the Black Lake and surrounding bischofite flows and the southwestern salt canyons, which contain water reservoirs often influenced by the geothermal activity; and (3) the Yellow Lake (Gaet'Ale) area (Fig. 1e). We also collected samples from the hypersaline Lake Assale (Karum), located a few kilometres to the south in the Danakil Depression (Fig. 1b). Physicochemical parameters (Fig. 3) were measured in situ with a YSI Professional Series Plus multiparameter probe (pH, temperature, dissolved oxygen, redox potential) up to 70 °C and a Hanna HI93530 temperature probe (working range −200/1,000 °C) and a Hanna HI991001 pH probe (working pH range −2.00/16.00) at higher temperatures. Salinity was measured in situ with a refractometer on 1/10 dilutions in MilliQ water. Brine samples for chemical analyses were collected in 50-ml glass bottles after prefiltration through 0.22-µm pore-diameter filters; bottles were filled to the top and sealed with rubber stoppers to prevent the (further) oxidation of reduced fluids. Solid and water samples for microbial diversity analyses and culturing assays were collected under aseptic conditions to prevent contamination (gloves, sterile forceps and containers). Samples for culture assays were kept at room temperature. Salts and mineral fragments for DNA-based analyses were conditioned in Falcon tubes and fixed with absolute ethanol. Water samples (volumes for each sample are indicated in Supplementary Table 1) were filtered through 30-µm pore-diameter filters to remove large particles and sequentially filtered either through 0.22-µm pore-diameter filters (Whatman) or 0.2-µm pore-size cell-trap units (MEM-TEQ Ventures). Filters or cell-trap concentrates retaining 0.2–30 µm particles were fixed in 2-ml cryotubes with absolute ethanol (>80% final concentration). Back in the laboratory, ethanol-fixed samples were stored at −20 °C until use.

Chemical analyses, salinity, chaotropicity, ionic strength and water activity. The chemical composition of solid and 0.2-µm prefiltered liquid samples was analysed at SIDI Service (Servicio Interdepartamental de Investigación, Universidad Autónoma de Madrid). Major and trace elements in liquid samples were analysed by total reflection X-ray fluorescence with a TXRF-8030c FEI spectrometer and inductively coupled plasma–mass spectrometry using a Perkin–Elmer NexION 300XX instrument. Ions were analysed using a Dionex DX-600 ion chromatography system. Organic molecules were characterized using a Varian HPLC–diode array detector/FL/LS liquid chromatograph. Crystalline phases in solid samples were characterized by X-ray diffraction using a X'Pert PRO Theta/Theta diffractometer (Panalytical) and identified by comparison with the International Centre for Diffraction Data PDF-4+ database using the High Score Plus software (Malvern Panalytical https://www.malvernpanalytical.com/es/products/category/software/x-ray-diffraction-software/highscore-with-plus-option). Inorganic data are provided in Supplementary Table 1 and organic and ionic chemistry data in Supplementary Tables 2 and 3, respectively. Salinity (weight/volume, expressed in percentage throughout the manuscript) was measured in triplicates (and up to six times) by weighing the total solids after heat-drying 1-ml aliquots in ceramic crucibles at 120 °C for at least 24 h. Chaotropicity was measured according to the temperature of gelation of ultrapure gelatin (for Ca-rich samples) and agar (rest of samples) and determined using the spectrometric assay developed by Cray et al.[45] (Extended Data Fig. 5). Chaotropicity was also calculated according to Cray and coworkers[46] based on the abundance of dominant Na, K, Mg, Ca and Fe cations and, on the ground that Cl is the dominant anion, assuming they essentially form chlorine salts (NaCl, KCl, $MgCl_2$, $CaCl_2$ and $FeCl_2$). Ionic strength was calculated according to Fox-Powell et al.[47]. Water activity was measured on 10-ml unfiltered aliquots at room temperature (25 °C) using a HC2-AW probe and HP23-AW-A indicator (Rotronic AG) calibrated at 23 °C using the

AwQuick acquisition mode (error per measure 0.0027). From a strict biological perspective, these water activity measurements are not sufficiently accurate and need to be considered as indicative because cells can be sensitive to a 0.001 water activity change[48]. However, the measurements follow the same trend as shown by the other related parameters measured experimentally (salinity, chaotropicity). We used R-software[49] packages FactoMineR[50] and factoextra[51] to carry out a PCA of samples, chemical and physicochemical parameters (Fig. 2 and Extended Data Fig. 4). Differences between the groups of samples belonging to the same physicochemical zone that segregated in the PCA were tested using the one-way analysis of variance module of IBM SPSS Statistics 24 software. The significance of differences among groups and with the measured parameters were checked by a post-hoc comparison using the Bonferroni test.

DNA purification and 16S/18S rRNA gene metabarcoding. DNA from filters, cell-trap concentrates and grinded solid samples was purified using the Power Soil DNA Isolation Kit (MoBio) under an ultraviolet-irradiated Erlab CaptairBio DNA/RNA PCR Workstation. Before DNA purification, filters were cut into small pieces with a sterile scalpel and the ethanol remaining in cryotubes was filtered through 0.2 µm pore-diameter filters and processed in the same way. Ethanol-fixed cell-trap concentrates were centrifuged for 10 min at 13,000 r.p.m. and the pellet resuspended in the first kit buffer. Samples were rehydrated for at least 2 h at 4 °C in the kit resuspension buffer. We used the Arcturus PicoPure DNA Isolation kit (Applied Biosystems; samples labelled pp) for a selection of cell-trap concentrates, FACS-sorted cells and for monitoring potential culture enrichments. DNA was resuspended in 10 mM Tris–HCl buffer, pH 8.0 and stored at −20 °C. Bacterial and archaeal 16S rRNA gene fragments of approximatively 290 bp encompassing the V4 hypervariable region were amplified with PCR using U515F (5′-GTGCCAGCMGCCGCGGTAA) and U806R (5′-GGACTACVSGGGTATCTAAT) primers. PCR reactions were conducted in 25 µl, using 1.5 mM $MgCl_2$, 0.2 mM of each dNTP (PCR Nucleotide Mix, Promega), 0.1 µM of each primer, 1–5 µl of purified 'DNA' and 1 unit of the hot-start Taq Platinum polymerase (Invitrogen). GoTaq (Promega) was also used when amplicons were not detected, but did not yield better results. Amplification reactions were performed for 35 cycles (94 °C for 15 s, 50–55 °C for 30 s and 72 °C for 90 s), after a 2 min-denaturation step at 94 °C and before a final extension at 72 °C for 10 min. Amplicons were visualized after gel electrophoresis and staining with ultrasensitive GelRed nucleic acid gel (Biotium) on an ultraviolet-light transilluminator. When direct PCR reactions failed to yield amplicons after several assays, PCR conditions and using increasing amounts of input potential DNA, we carried out seminested reactions. For seminested reactions, we used those same primers for PCR amplification but we used as input potential DNA 1 µl of PCR products, from a first amplification reaction performed with universal prokaryotic primers U340F (5′-CCTACGGGRBGCASCAG) and U806R, including the negative controls from the first PCR reaction. Eukaryotic 18S rRNA gene fragments that included the V4 hypervariable region were amplified using primers EK-565F (5′-GCAGTTAAAAAGCTCGTAGT) and 18S-EUK-1134-R-UNonMet (5′-TTTAAGTTTCAGCCTTGCG). Primers were tagged with different molecular identifiers (MID) to allow multiplexing and subsequent sequence sorting. Amplicons from at least five independent PCR products for each sample were pooled together and then purified using the QIAquick PCR purification kit (Qiagen). Whenever seminested PCR reactions yielded amplicons, seminested reactions using first PCR negative controls as the input also yielded amplicons (second PCR controls did not yield amplicons). Products of these positive 'negative' controls were pooled in two control sets (1 and 2) and sequenced along with the rest of amplicons. DNA concentrations were measured using Qubit dsDNA HS assays (Invitrogen). Equivalent amplicon amounts obtained for 54 samples (including controls) were multiplexed and sequenced using paired-end (2 × 300 bp) MiSeq Illumina technology (Eurofins Genomics). In parallel, we tried to amplify near-complete 16S/18S rRNA gene fragments (~1,400–1,500 bp) using combinations of forward archaea-specific primers (21F, 5′-TTCCGGTTGATCCTGCCGGA; Ar109F, 5′-ACKGCTGCTCAGTAACACGT) and bacteria-specific primers (27F, 5′-AGAGTTTGATCCTGGCTCAG) with the prokaryotic reverse primer 1492R (5′-GGTTACCTTGTTACGACTT) and eukaryotic primers 82F (5′-GAAACTGCGAATGGCTC) and 1520R (5′-CYGCAGGTTCACCTAC). When amplified, DNA fragments were cloned using TopoTA cloning (Invitrogen) and clone inserts were Sanger-sequenced to yield longer reference sequences. Forward and reverse Sanger sequences were quality controlled and merged using Codon Code Aligner (http://www.codoncode.com/aligner/).

Sequence treatment and phylogenetic analyses. Paired-end reads were merged and treated using a combination of existing software to check quality, eliminate primers and MIDs, and to remove potential chimeras. Sequence statistics are given in Extended Data Fig. 6. Briefly, read merging was determined with FLASH[52]; primers and MIDs trimmed with cutadapt[53]; and clean merged reads dereplicated using vsearch[54] with the uchime_denovo option to eliminate potential chimeras. The resulting dereplicated clean merged reads were used to define OTUs at 95% identity cut-off using CD-HIT-EST[55]. This cut-off offered (1) a reasonable operational approximation to the genus-level diversity while producing a manageable number of OTUs to be included in phylogenetic trees (see below) and (2) a conservative identification of potential contaminants in our seminested PCR-derived datasets. Diversity (Simpson), richness (Chao1) and evenness indices were determined using R-package 'vegan' (Supplementary Table 5). OTUs were assigned to known taxonomic groups based on similarity with sequences of a local database, including sequences from cultured organisms and environmental surveys retrieved from SILVAv128 (ref. [56]) and PR2v4 (ref. [57]). The taxonomic assignation of bacteria and archaea was refined by phylogenetic placement of OTU representative sequences in reference phylogenetic trees. To build these trees, we used Mafft-linsi v.7.38 (ref. [58]) to produce alignments of near full-length archaeal and bacterial 16S rRNA gene sequences comprising Sanger sequences from our gene libraries (144 archaeal and 91 bacterial) and selected references for major identified taxa plus the closest blast-hits to our OTUs (702 archaea and 2,922 bacterial). Poorly aligned regions were removed using TrimAl[59]. Maximum likelihood phylogenetic trees were constructed with IQ-TREE[60] using the general time reversible (GTR) model of sequence evolution with a gamma law and taking into account invariable sites (GTR + G + I). Node support was estimated by ultrafast bootstrapping as implemented in IQ-TREE. Shorter OTU representative sequences (2,653 archaeal and 710 bacterial) were added to the reference alignment using MAFFT (accurate-linsi 'addfragments' option). This final alignment was split into two files (references and OTUs) before using the EPA-ng tool (https://github.com/Pbdas/epa-ng) to place OTUs in the reference trees reconstructed with IQ-TREE. The jplace files generated by EPA-ng were transformed into newick tree files with the genesis library (https://github.com/lczech/genesis). Tree visualization and ring addition were done with GraphLan[61]. To determine whether our OTUs might correspond to thermophilic species, we plotted the GC content of the 16S rRNA gene region used for metabarcoding analyses of a selection of 88 described archaeal species with optimal growth temperatures ranging from 15 to 103 °C. These included representatives of all Halobacteria genera because they are often characterized by high GC content. A regression analysis confirmed the occurrence of a positive correlation[62] between rRNA GC content and optimal growth temperature for this shorter 16S rRNA gene amplified region (Fig. 3b). We then plotted the GC content of our archaeal OTUs on the same graph. Dots corresponding to Halobacteria genera remain out of the dark shadowed area in Fig. 3b.

Cultures. Parallel culture attempts were carried out in two different laboratories (Orsay and Madrid). We used several culture media derived from a classical halophile base mineral growth medium[63] containing NaCl (234 g l^{-1}), KCl (6 g l^{-1}), NH_4Cl (0.5 g l^{-1}), K_2HPO_4 (0.5 g l^{-1}), $(NH_4)_2SO_4$ (1 g l^{-1}), $MgSO_4.7H_2O$ (30.5 g l^{-1}), $MnCl_2.7H_2O$ (19.5 g l^{-1}), $CaCl_2.6H_2O$ (1.1 g l^{-1}) and Na_2CO_3 (0.2 g l^{-1}). The pH was adjusted to 4 and 2 with 10 N H_2SO_4. The autoclaved medium was amended with filter-sterilized cyanocobalamin (1 µM final concentration) and 5 ml of an autoclaved $CaCl_2{\cdot}6H_2O$ 1 M stock solution. Our medium MDH2 contained yeast extract (1 g l^{-1}) and glucose (0.5 g l^{-1}). The MDSH1 medium had only two-thirds of each base medium salt concentration plus $FeCl_3$ (0.1 g l^{-1}) and 10 ml l^{-1} of Allen's trace solution. It was supplemented with three energy sources (prepared in 10 ml distilled water at pH 2 and sterilized by filtration): yeast extract (1 g l^{-1}) and glucose (0.5 g l^{-1}) (MDSH1–org medium); $Na_2S_2O_3$ (5 g l^{-1}) (MDSH1–thio medium) and $FeSO_4{\cdot}7H_2O$ (30 g l^{-1}) (MDSH1–Fe medium). Medium MDSH2 mimicked more closely some Dallol salts as it also contained $FeCl_3$ (0.1 g l^{-1}), $MnCl_2{\cdot}4H_2O$ (0.7 g l^{-1}), $CuSO_4$ (0.02 g l^{-1}), $ZnSO_4{\cdot}7H_2O$ (0.05 g l^{-1}) and LiCl (0.2 g l^{-1}). It also contained 10 ml l^{-1} of Allen's trace solution combined with the same energy sources used for MDSH1, yielding media MDSH2–org, MDSH2–thio and MDSH2–Fe. For enrichment cultures, we added 0.1 ml liquid samples to 5 ml medium at pH 2 and 4 and incubated at 37 °C, 50 °C and 70 °C in 10-ml sterile glass tubes depending on the original sample temperatures. Three additional variants of the base salt medium, which was supplemented with $FeCl_3$ and trace minerals, contained 0.2 g l^{-1} yeast extract (SALT–YE), 0.5 g l^{-1} thiosulfate (SALT–THIO) or 0.6 g l^{-1} benzoate and 5 mM hexadecane (SALT–BH). The pH of these media was adjusted with 34% HCl to pH 1.5 for Dallol and Black Lake samples, and to pH 3.5 for Yellow Lake, PS3 and PSBL samples. We added 1 ml of sample to 4 ml of medium and incubated it at 45 °C in light conditions and at 37 °C and 70 °C in dark conditions. We also tried cultures in anaerobic conditions. Potential growth was monitored by optical microscopy and, for some samples, SEM. In the rare cases where enrichments were obtained, we attempted isolation by serial dilutions.

Flow cytometry and FACS. The presence of cell/particle populations above background levels in Dallol samples was assessed with a flow-cytometer cell-sorter FACSAriaIII (Becton Dickinson). Several DNA dyes were tested for lowest background signal in forward scatter (FSC) red (695 ± 20 nm) and green (530 ± 15 nm) fluorescence (Extended Data Fig. 9a) using sterile SALT-YE medium as blank. DRAQ5 and SYTO13 (ThermoFisher) were retained and used at 5 µM final concentration to stain samples in the dark at room temperature for 1 h. Cell-trap concentrated samples were diluted at 20% with 0.1-µm filtered and autoclaved MilliQ water. The FACSAriaIII was set at purity sort mode triggering on the forward scatter (FSC). Fluorescent target cells/particles were gated based on the FSC and red or green fluorescence (Extended Data Fig. 9b) and flow-sorted at a rate of 1–1,000 particles s^{-1}. Sorting was conducted using the FACSDiva software (Becton Dickinson) and figures were produced using FCSExpress 6 software

(De Novo Software). Sorted cells/particles were subsequently observed by SEM for characterization. Minimum and maximum cell abundances were estimated based on the number of sorted particles, duration of sorting and minimal (10 µl min^{-1}) and maximal (80 µl min^{-1}) flow rates of the FACSAria (Becton Dickinson FACSAria manual).

SEM and elemental analysis. SEM analyses were carried out on natural samples, FACS-sorted cells/particles and a selection of culture attempts. Liquid samples were deposited onto 0.1 µm pore-diameter filters (Whatman) under a mild vacuum aspiration regime and briefly rinsed with 0.1-µm filtered and autoclaved MilliQ water under the same vacuum regime. Filters were allowed to dry and sputtered with carbon prior to SEM observations. A Zeiss ultra55 field emission gun SEM was used for the SEM analyses. Secondary electron images were acquired using an In Lens detector at an accelerating voltage of 2.0 kV and a working distance of ~7.5 mm. Backscattered electron images were acquired for chemical mapping using an angle selective backscattered detector at an accelerating voltage of 15 kV and a working distance of ~7.5 mm. Elemental maps were generated from hyperspectral images (HyperMap) by EDXS using an EDS QUANTAX detector. EDXS data were analysed using the ESPRIT software package (Bruker).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sanger sequences have been deposited in GenBank (National Center for Biotechnology Information) with accession numbers MK894601–MK894820 and Illumina sequences in GenBank Short Read Archive with BioProject number PRJNA541281.

Received: 30 May 2019; Accepted: 16 September 2019;
Published online: 28 October 2019

References

1. Harrison, J. P., Gheeraert, N., Tsigelnitskiy, D. & Cockell, C. S. The limits for life under multiple extremes. *Trends Microbiol.* **21**, 204–212 (2013).
2. Merino, N. et al. Living at the extremes: extremophiles and the limits of life in a planetary context. *Front. Microbiol.* **10**, 1785 (2019).
3. Johnson, S. S., Chevrette, M. G., Ehlmann, B. L. & Benison, K. C. Insights from the metagenome of an acid salt lake: the role of biology in an extreme depositional environment. *PLoS ONE* **10**, e0122869 (2015).
4. Zaikova, E., Benison, K. C., Mormile, M. R. & Johnson, S. S. Microbial communities and their predicted metabolic functions in a desiccating acid salt lake. *Extremophiles* **22**, 367–379 (2018).
5. Futterer, O. et al. Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. *Proc. Natl Acad. Sci. USA* **101**, 9091–9096 (2004).
6. Varet, J. in *Geology of Afar (East Africa). Regional Geology Reviews* (eds.Oberhänsli, R. et al.) Ch. 7 (Springer, 2018).
7. Franzson, H., Helgadóttir, H. M. & Óskarsson, F. Surface exploration and first conceptual model of the Dallol geothermal area, northern Afar, Ethiopia. In *Proc. World Geothermal Congress* (2015).
8. Darrah, T. H. et al. Gas chemistry of the Dallol region of the Danakil Depression in the Afar region of the northern-most East African Rift. *Chem. Geol.* **339**, 16–29 (2013).
9. Holwerda, J. G. & Hutchinson, R. W. Potash-bearing evaporites in the Danakil area, Ethiopia. *Econ. Geol.* **63**, 124–150 (1968).
10. Warren, J. K. *Danakhil Potash, Ethiopia: Beds of Kainite/Carnallite, Part 2 of 4* (SaltWork Consultants, 2015).
11. Cavalazzi, B. et al. The Dallol geothermal area, northern Afar (Ethiopia): an exceptional planetary field analog on Earth. *Astrobiology* **19**, 553–578 (2019).
12. Hallsworth, J. E. et al. Limits of life in $MgCl_2$-containing environments: chaotropicity defines the window. *Environ. Microbiol.* **9**, 801–813 (2007).
13. Stevenson, A. et al. Is there a common water-activity limit for the three domains of life? *ISME J.* **9**, 1333–1351 (2015).
14. McKay, C. P. Requirements and limits for life in the context of exoplanets. *Proc. Natl Acad. Sci. USA* **111**, 12628–12633 (2014).
15. Moissl-Eichinger, C., Cockell, C. & Rettberg, P. Venturing into new realms? Microorganisms in space. *FEMS Microbiol. Rev.* **40**, 722–737 (2016).
16. Pérez, E. & Chebude, Y. Chemical analysis of Gaet'Ale, a hypersaline pond in Danakil Depression (Ethiopia): new record for the most saline water body on Earth. *Aquat. Geochem* **23**, 109–117 (2017).
17. Kotopoulou, E. et al. A polyextreme hydrothermal system controlled by iron: the case of Dallol at the Afar triangle. *ACS Earth Space Chem.* **3**, 90–99 (2019).
18. Warren, J. K. *Danakhil Potash, Ethiopia: Is the Present Geology the Key? Part 1 of 4* (SaltWork Consultants, 2015).
19. Tosca, N. J., Knoll, A. H. & McLennan, S. M. Water activity and the challenge for life on early Mars. *Science* **320**, 1204–1207 (2008).
20. Stevenson, A. et al. *Aspergillus penicillioides* differentiation and cell division at 0.585 water activity. *Environ. Microbiol.* **19**, 687–697 (2017).
21. Sheik, C. S. et al. Identification and removal of contaminant sequences from ribosomal gene databases: lessons from the census of deep life. *Front. Microbiol.* **9**, 840 (2018).
22. Weyrich, L. S. et al. Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* **19**, 982–996 (2019).
23. Narasingarao, P. et al. De novo metagenomic assembly reveals abundant novel major lineage of archaea in hypersaline microbial communities. *ISME J.* **6**, 81–93 (2012).
24. Sorokin, D. Y. et al. Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat. Microbiol.* **2**, 17081 (2017).
25. Sorokin, D. Y. et al. *Methanonatronarchaeum thermophilum* gen. nov., sp. nov. and '*Candidatus* Methanohalarchaeum thermophilum', extremely halo(natrono)philic methyl-reducing methanogens from hypersaline lakes comprising a new euryarchaeal class *Methanonatronarchaeia* classis nov. *Int J. Syst. Evol. Microbiol.* **68**, 2199–2208 (2018).
26. Borin, S. et al. Sulfur cycling and methanogenesis primarily drive microbial colonization of the highly sulfidic Urania deep hypersaline basin. *Proc. Natl Acad. Sci. USA* **106**, 9151–9156 (2009).
27. Mwirichia, R. et al. Metabolic traits of an uncultured archaeal lineage–MSBL1–from brine pools of the red sea. *Sci. Rep.* **6**, 19181 (2016).
28. Castelle, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
29. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).
30. Dombrowski, N., Lee, J. H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366**, fnz008 (2019).
31. Petitjean, C., Deschamps, P., Lopez-Garcia, P. & Moreira, D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).
32. Golyshina, O. V. et al. 'ARMAN' archaea depend on association with euryarchaeal host in culture and in situ. *Nat. Commun.* **8**, 60 (2017).
33. Minegishi, H. et al. Acidophilic haloarchaeal strains are isolated from various solar salts. *Saline Syst.* **4**, 16 (2008).
34. Naor, A. & Gophna, U. Cell fusion and hybrids in archaea: prospects for genome shuffling and accelerated strain development for biotechnology. *Bioengineered* **4**, 126–129 (2013).
35. Garcia-Ruiz, J. M. et al. Self-assembled silica-carbonate structures and detection of ancient microfossils. *Science* **302**, 1194–1197 (2003).
36. Garcia-Ruiz, J. M., Melero-Garcia, E. & Hyde, S. T. Morphogenesis of self-assembled nanocrystalline materials of barium carbonate and silica. *Science* **323**, 362–365 (2009).
37. Slonczewski, J. L., Fujisawa, M., Dopson, M. & Krulwich, T. A. Cytoplasmic pH measurement and homeostasis in bacteria and Archaea. *Adv. Micro. Physiol.* **55**, 1–79 (2009).
38. Buetti-Dinh, A., Dethlefsen, O., Friedman, R. & Dopson, M. Transcriptomic analysis reveals how a lack of potassium ions increases *Sulfolobus acidocaldarius* sensitivity to pH changes. *Microbiology* **162**, 1422–1434 (2016).
39. Gunde-Cimerman, N., Plemenitas, A. & Oren, A. Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiol. Rev.* **42**, 353–375 (2018).
40. Lee, C. J. D. et al. NaCl-saturated brines are thermodynamically moderate, rather than extreme, microbial habitats. *FEMS Microbiol. Rev.* **42**, 672–693 (2018).
41. Colman, D. R., Lindsay, M. R. & Boyd, E. S. Mixing of meteoric and geothermal fluids supports hyperdiverse chemosynthetic hydrothermal communities. *Nat. Commun.* **10**, 681 (2019).
42. López-García, P., Zivanovic, Y., Deschamps, P. & Moreira, D. Bacterial gene import and mesophilic adaptation in Archaea. *Nat. Rev. Microbiol.* **13**, 447–456 (2015).
43. Hensel, R. & König, H. Thermoadaptation of methanogenic bacteria by intracellular ion concentration. *FEMS Microbiol. Lett.* **49**, 75–79 (1988).
44. Shima, S., Thauer, R. K. & Ermler, U. Hyperthermophilic and salt-dependent formyltransferase from *Methanopyrus kandleri*. *Biochem. Soc. Trans.* **32**, 269–272 (2004).
45. Cray, J. A., Russell, J. T., Timson, D. J., Singhal, R. S. & Hallsworth, J. E. A universal measure of chaotropicity and kosmotropicity. *Environ. Microbiol.* **15**, 287–296 (2013).
46. Cray, J. A. et al. Chaotropicity: a key factor in product tolerance of biofuel-producing microorganisms. *Curr. Opin. Biotechnol.* **33**, 228–259 (2015).
47. Fox-Powell, M. G., Hallsworth, J. E., Cousins, C. R. & Cockell, C. S. Ionic strength is a barrier to the habitability of Mars. *Astrobiology* **16**, 427–442 (2016).
48. Stevenson, A. et al. Multiplication of microbes below 0.690 water activity: implications for terrestrial and extraterrestrial life. *Environ. Microbiol.* **17**, 257–277 (2015).

49. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
50. Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
51. Kassambara, A. & Mundt, F. factoextra: extract and visualize the results of multivariate data analyses. https://CRAN.R-project.org/package=factoextra (2017).
52. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
53. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
54. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
55. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
56. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
57. Guillou, L. et al. The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013).
58. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
59. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
60. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
61. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).
62. Wang, H. C., Xia, X. & Hickey, D. Thermal adaptation of the small subunit ribosomal RNA gene: a comparative study. *J. Mol. Evol.* **63**, 120–126 (2006).
63. Rodriguez-Valera, F., Ruiz-Berraquero, F. & Ramos-Cormenzana, A. Behaviour of mixed populations of halophilic bacteria in continuous cultures. *Can. J. Microbiol.* **26**, 1259–1263 (1980).

Acknowledgements

We are grateful to O. Grunewald for co-organizing the Dallol expeditions, documenting field research and providing drone images, and also to Jean-Marie Hullot (in memoriam), Françoise Brenckmann and the Fondation Iris for funding the first field trip. We thank L. Cantamessa for the in situ logistics and discussions about local history. We acknowledge M. Tafari (Mekelle University), A. A. Aliyu and the Afar authorities for local assistance, as well as the Ethiopian army and the Afar police for providing security. We thank J. Barthélémy, E. Kotopoulou and J. Garcia-Ruiz for help and discussions during field trips. We thank H. Timpano and the UNICELL platform for cell sorting; A. Gutiérrez-Preciado for bioinformatic assistance; A. Kish and C. Faveau for allowing us to measure water activity of selected samples at the Muséum National d'Histoire Naturelle; E. Viollier for discussion on chemical analyses; C. Gille for help with cultures; G. Billo for script help to treat SEM pictures; and J. T. Díaz and P. T. Sanz for advice on statistical analyses. This research was funded by the French CNRS (National Center for Scientific Research) basic annual funding, the CNRS programme TELLUS INTERRVIE and the European Research Council (ERC) under the European Union's Seventh Framework Programme (ERC grant no. 322669 to P.L.-G.). We thank the European COST Action TD1308 Origins for funding a short stay of A.I.L.-A. in Orsay. J.B. was financed by the French Ministry of National Education, Research and Technology.

Author contributions

P.L.-G. and D.M. designed and supervised the research. P.L.-G. organized the scientific expeditions. J.B., P.L.-G., D.M., L.J. and J.M.L.-G. collected samples and took measurements in situ. J.B., P.L.-G. and P.B. carried out molecular biology analyses. J.B., A.I.L.-A. and D.M. performed culture, chemistry analyses and water–salt related measurements. A.I.L.-A. and J.B. performed statistical analyses. J.B., G.R. and D.M. analysed metabarcoding data. K.B. performed SEM and EDX analyses. J.M.L.-G. mapped geothermal activity and georeferenced all samples. L.J. and J.B. performed FACS-derived analyses. P.L.-G. and J.B. wrote the manuscript. All authors read and commented on the manuscript.

Competing interests

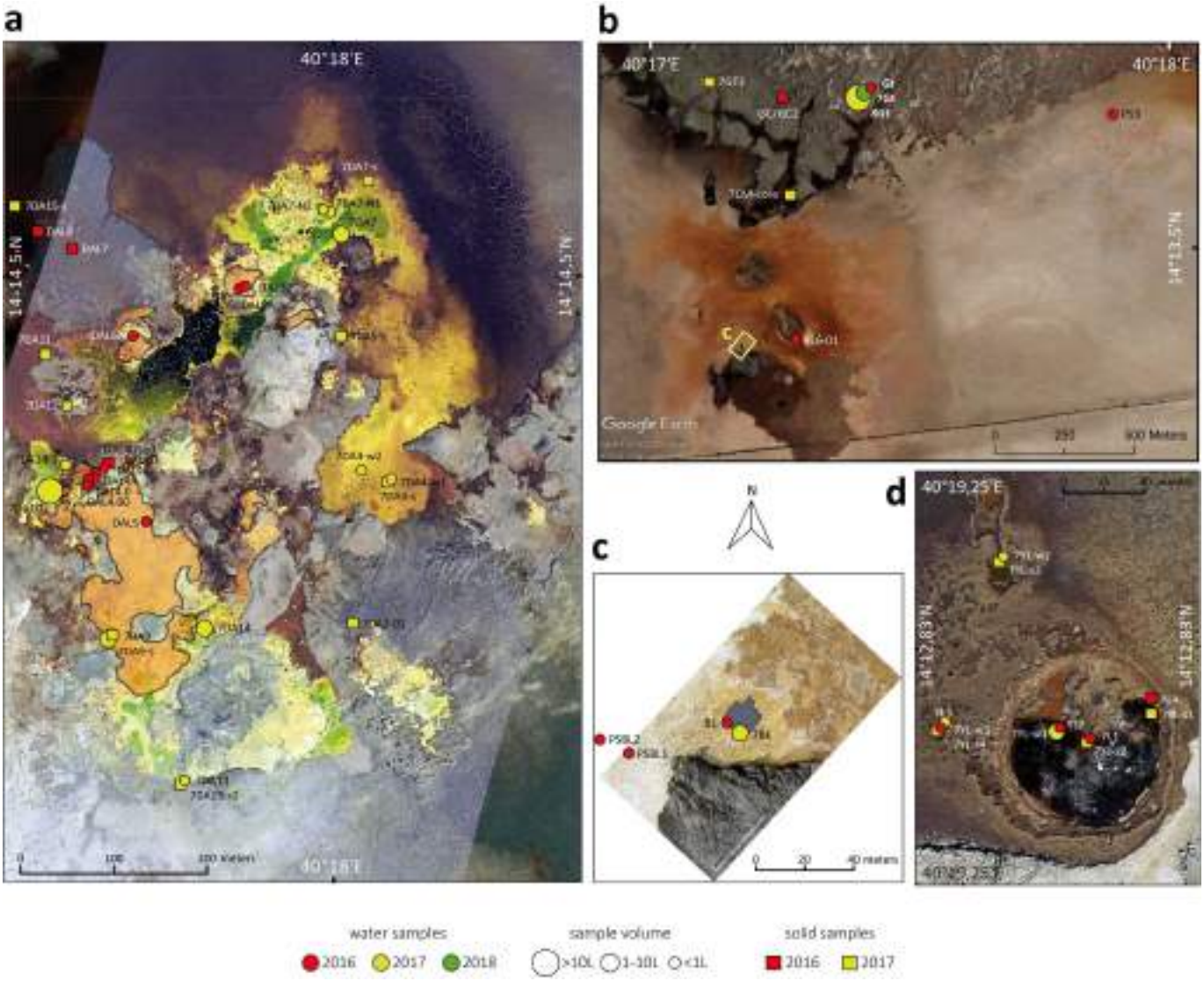
The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-019-1005-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-019-1005-0.

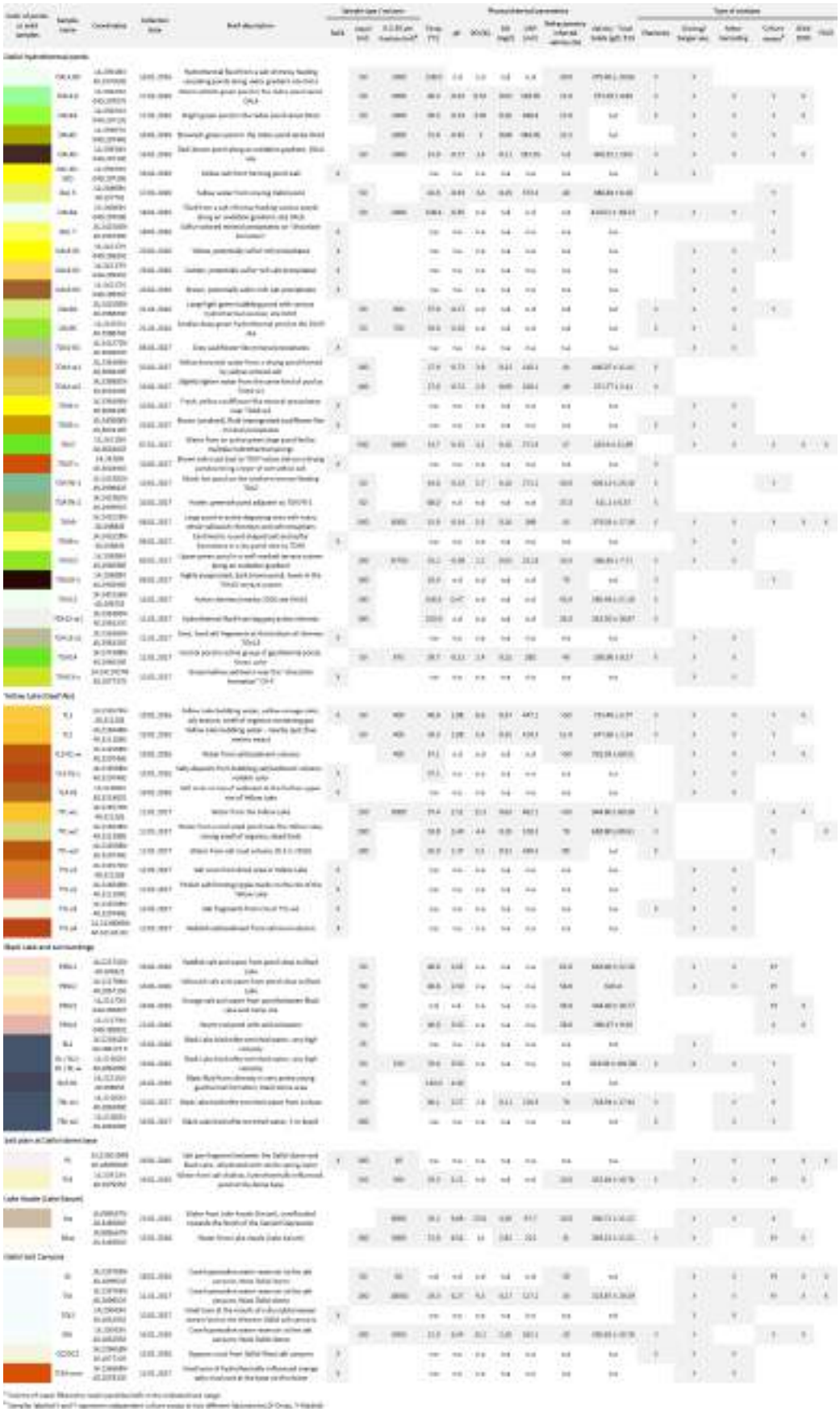
Correspondence and requests for materials should be addressed to P.L.-G.



Extended Data Fig. 1 | Aerial view of the main sampling sites in the Dallol area. a, Dallol dome summit showing the acidic green-yellow-brown coloured hydrothermal ponds and active degassing areas during our 2017 sampling trip; the orange-shaded area shows the active hydrothermal zone in January 2016. **b**, Dallol West salt canyons and Black Mountain area. **c**, Black Lake. **d**, Yellow Lake and surroundings. Names of samples and sampling sites are indicated. The size of circles is proportional to the water volume collected or filtered for subsequent analyses. Aerial photographs were taken from a drone by O. Grunewald, except b, which is a Google Earth aerial image (09/03/2016) obtained by the Sentinel satellite (ESA Copernicus program) provided by Image © 2019 CNES/Airbus.

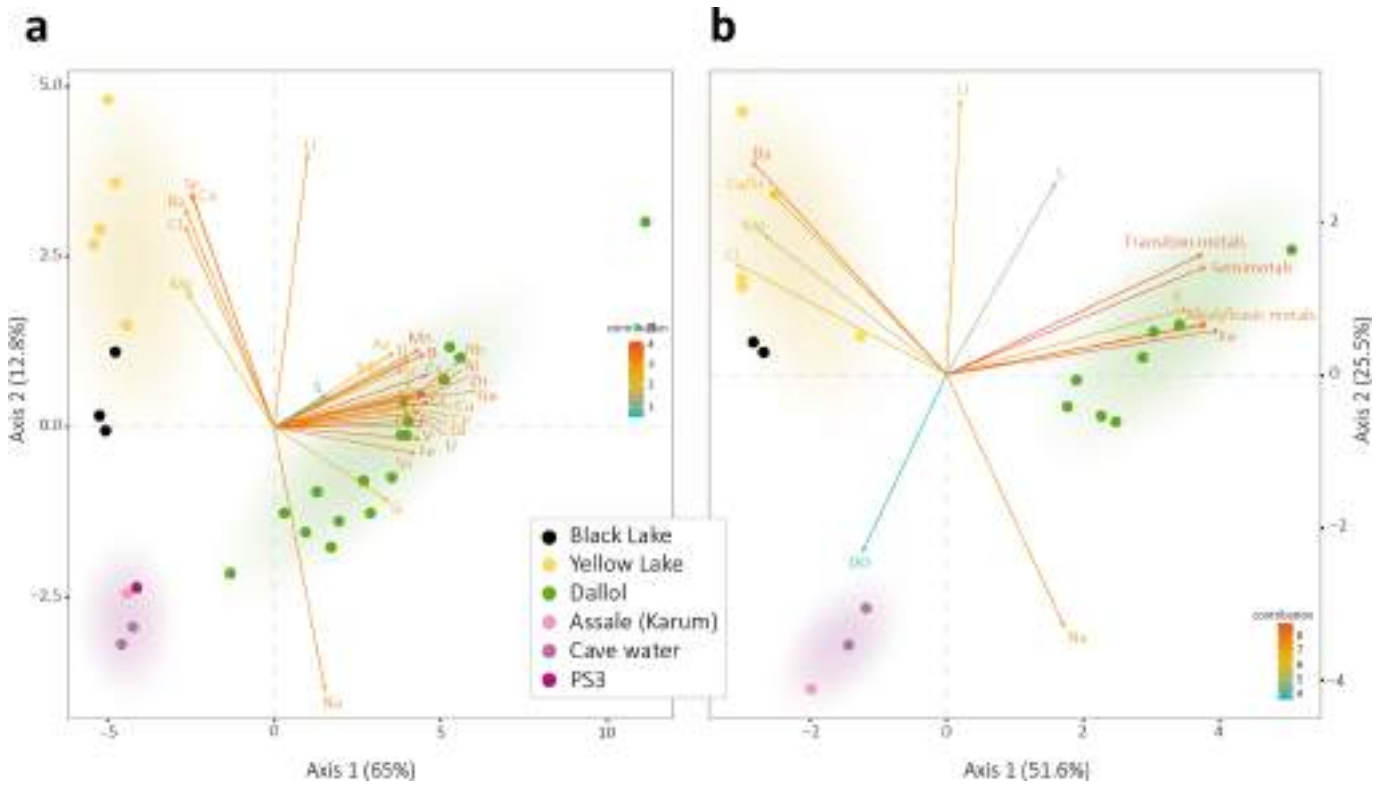
Extended Data Fig. 2 | see figure caption on next page.

Extended Data Fig. 2 | Views of different sampling sites in the Dallol dome and surroundings in the Danakil Depression. a, DAL4 sampling site ponds; **b**, DAL5 pond and active degassing area; **c**, active hydrothermal springs in DAL9 ponds; **d**, in situ cell-trap filtration at the 7DA7 sampling area; **e**, 7DA9 sampling site; **f**,7DA10 ponds showing increasingly darker and brownish colours along the oxidation gradient; **g**, water samples from the different 7DA10 ponds; **h**, DAL8 mineral precipitates; **i**,'proto-soil'-like salt crust (7YL-S1) near the Yellow Lake; **j**, Yellow Lake showing active degassing; **k**, YL3, salt-mud volcano in the Yellow Lake area; **l**, 'Little Dallol' hydrothermal very active area in 2016 on the way to the Black Mountain (in the distance; inlet, chimney emitting hydrocarbon-rich fluids at 110 °C); **m**, Black Lake; **n**, PSBL2 (Black Lake area ponds); **o**, wet salt plain, influenced by hydrothermal activity, corresponding to PS3 sample area; **p**, the cave in the salt canyons where Gt, 7Gt and 8Gt samples were collected; **q**, salt canyons; **r**, Assale (Karum) lake. Sample names starting by 7 indicate collection in 2017. Pictures from all other samples/sampling sites were taken during the 2016 expedition.



Extended Data Fig. 3 | see figure caption on next page.

Extended Data Fig. 3 | List and description of samples from the Dallol area analysed in this study and type of analyses performed. DO, dissolved oxygen; ORP, oxido-reduction potential; SEM–/EDXS, scanning electron microscopy/energy-dispersive x-ray spectrometry; FACS, fluorescence-activated cell sorting analysis; n.a., not applicable; n.d. not determined. Refractometry-derived salinity refers to the percentage (w/v) of local salt composition (see Supplementary Tables 1 and 3 for elementary and ionic analyses) measured in situ. Salinity was also directly measured by weighting the total solids (dry weight experimentally measured in triplicates; SD, standard deviation).



Extended Data Fig. 4 | Principal Component Analyses (PCA) of Dallol area sampling sites as a function of physicochemical parameters. PCA of 29 samples according to their chemical composition; only relatively abundant elements (see Supplementary Table 1) are included in the analysis. A summary of this analysis is shown in Fig. 2f. **b**, PCA including the same variables as Fig. 2f but additionally including dissolved oxygen (DO). Measured parameters on site can be found in Extended Data Fig. 3. Coloured zones in PCA analyses correspond to the three major chemical zones identified in this study.

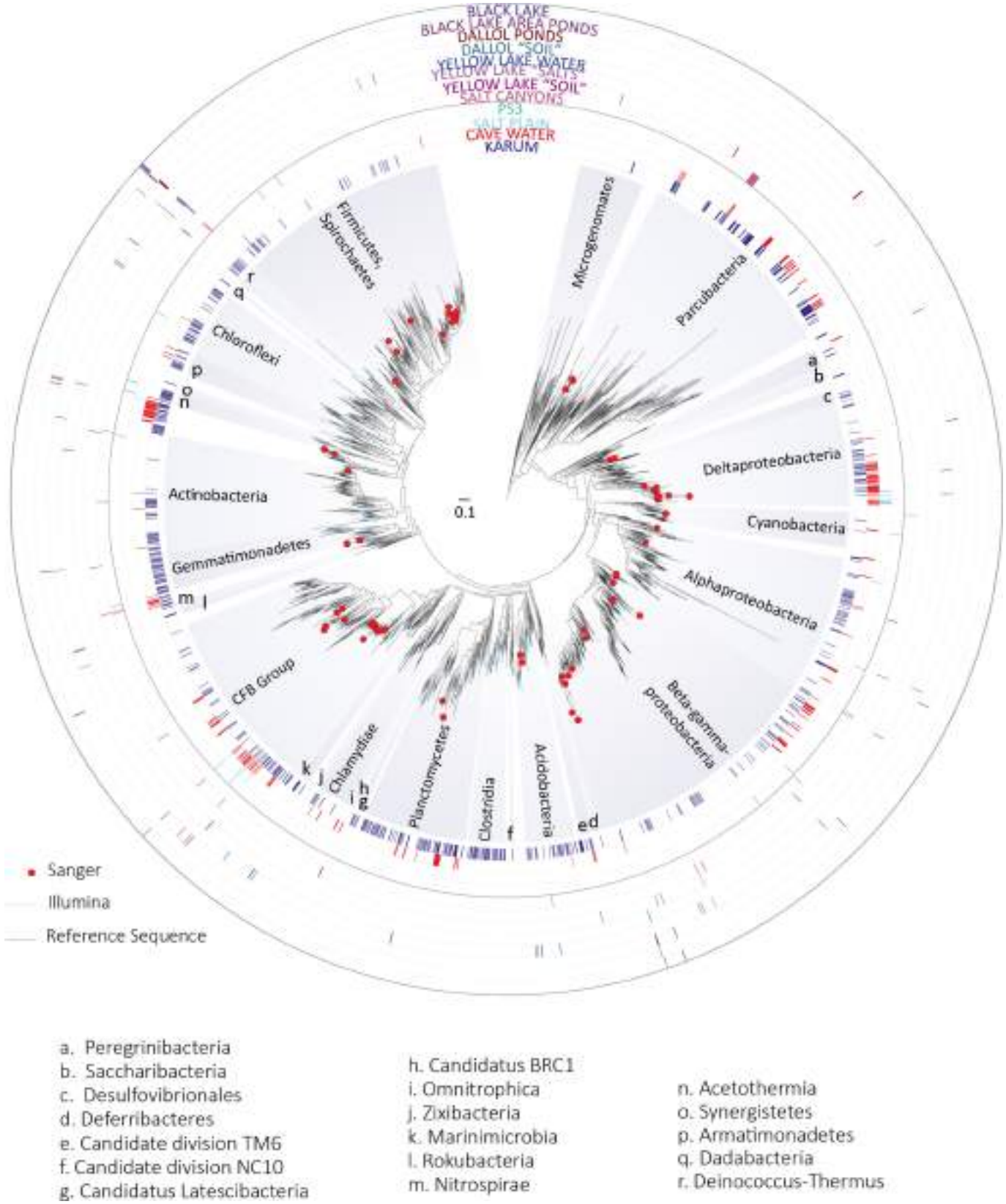
		Measured chaotropicity (kJ/kg)	Calculated chaotropicity (kJ/kg)	Ionic strength (mol/L)	Water activity (a_w)
Life threshold*		≤87.3		≤12.141	≥0.585
Cave water	Gt		n.d	n.d	0.728
	7Gt	-18.3	-23.80	4.751	0.729
	8Gt	-57.5	-56.65	6.873	0.731
Lake Assale	8Ass	n.d.	7.10	7.274	0.718
Geothermally influenced Salt Plain	PS3	n.d.	24.09	7.138	n.d
Dallol dome hydrothermal pools	DAL 4.00	-21.7	-17.87	6.104	0.719
	DAL 4.0	n.d.	-18.71	7.307	n.d
	DAL 4A	n.d.	-9.61	6.346	n.d
	DAL 4D	n.d.	2.14	7.104	n.d
	DAL 6A	n.d.	-23.97	7.203	n.d
	DAL 9A	n.d.	-7.77	7.529	n.d
	DAL 9C	n.d.	-16.15	8.349	n.d
	7DAL4-W1	19.3	40.44	6.314	0.667
	7DAL4-W2	8.3	14.28	5.383	0.698
	7DAL7	8.8	19.64	5.989	0.694
	7DAL-N1	9.2	20.84	6.472	0.694
	7DAL-N2	11.5	11.01	5.940	0.698
	7DAL9	-8.2	2.95	5.176	0.708
	7DAL10	2.1	-7.46	5.037	0.714
	7DAL10-1	n.d	n.d	n.d	0.580
	7DAL12	-31.2	-20.57	5.793	n.d
	7DAL13-W1	-24.8	-20.13	4.785	0.723
	7DAL14	-11.7	7.54	5.307	0.748
Black Lake area pools	PSBL1	108.3	n.d	n.d	0.334
	PSBL2	93.5	n.d	n.d	0.345
	PSBL3	63.4	n.d	n.d	0.722
	PSBL4	61.8	n.d	n.d	0.711
Black Lake	BL	288.3	354.19	19.155	0.319
	7BL-W1	198.5	259.41	14.206	0.322
	7BL-W2	201.3	268.89	14.721	n.d
Yellow Lake	YL1	n.d.	492.06	19.141	n.d
	YL2	n.d	574.04	22.085	n.d
	YL3	231.8	n.d	n.d	0.319
	7YL-W1	320.8	495.01	18.446	0.261
	7YL-W2	308.2	328.92	13.796	0.467
	7YL-W3	n.d.	466.64	17.609	n.d

* Data from Hallsworth et al (2007) and Stevenson et al (2015 and 2017)

Extended Data Fig. 5 | Chaotropicity, ionic strength and water activity for a selection of samples of the Dallol area. Chaotropicity was measured experimentally (see Methods) and also calculated, together with ionic strength values were from dominant Na, K, Mg, Ca, Fe chemistry data; water activity values were measured using a probe (see Methods). Known limits for life for each parameter are listed at the top of the table. Samples beyond that threshold for one or more of those parameters are shaded in grey.

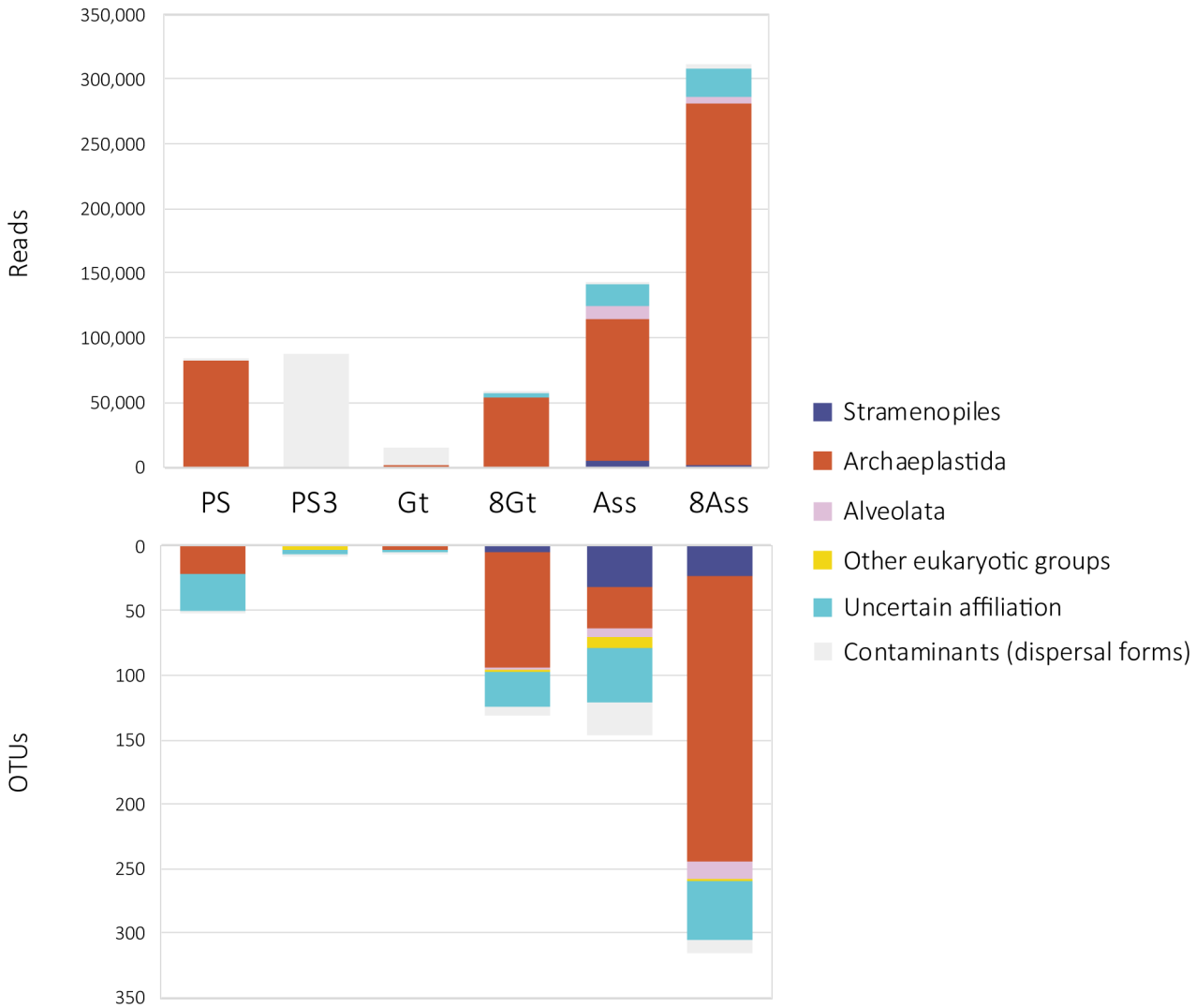
Sample name	Initial No. of merged reads	No. of high quality reads	No. of retained reads after chimera check	No. of retained reads after removing contaminants*	No. archaeal reads	No. OTUs	Diversity (Simpson index)	Evenness	Richness (Chao1) (s.e.)	No. of bacterial reads	No. OTUs	Diversity (Simpson index)	Evenness	Richness (Chao1) (s.e.)
Prokaryotic sequences							Archaea					Bacteria		
DAL4.0	169949	169848	165257	4	5	2	0.44	0.92	2 (0)	1	1	0.00	NA	1 (0)
DAL4A	152469	152298	149065	5023	5023	2	0.00	0.01	2 (0)	0	0	1.00	0.00	0 (NA)
DAL4C	126853	126665	123020	0	0	0	1.00	0.00	0 (NA)	0	0	1.00	0.00	0 (NA)
DAL4D	125034	124804	120619	0	0	0	1.00	0.00	0 (NA)	0	0	1.00	0.00	0 (NA)
DAL6A	234166	233935	224950	4314	4303	21	0.03	0.01	35 (11)	11	4	0.45	0.64	7 (4)
DAL8-01	113694	112383	108675	6	0	0	1.00	0.00	0 (NA)	6	2	0.28	0.65	2 (0)
DAL8-02	182460	170331	172030	2	0	0	1.00	0.00	0 (NA)	2	2	0.50	1.00	3 (2)
DAL8-03	154815	151700	148869	41	0	0	1.00	0.00	0 (NA)	41	4	0.34	0.49	4 (0)
DAL9A	132758	131862	126420	2	0	0	1.00	0.00	0 (NA)	2	2	0.50	1.00	3 (2)
DAL9C	107108	106589	104748	0	0	0	1.00	0.00	0 (NA)	0	0	1.00	0.00	0 (NA)
7DA7	133151	132970	128002	1	1	1	0.00	NA	1 (0)	0	0	1.00	0.00	0 (NA)
7DA7-pp	42089	42043	41253	2	2	2	0.50	1.00	3 (2)	0	0	1.00	0.00	0 (NA)
7DA9	158518	158248	152578	1821	1	1	0.00	NA	1 (0)	0	0	1.00	0.00	0 (NA)
7DA9-pp	217467	217096	192709	2	0	0	1.00	0.00	0 (NA)	2	1	0.00	NA	1 (0)
7DA10	213263	212784	205528	1	1	1	0.00	NA	1 (0)	0	0	1.00	0.00	0 (NA)
7DA10-pp	44566	44540	40224	62	0	0	1.00	0.00	0 (NA)	62	2	0.03	0.12	2 (0)
7DA14	168809	168530	162187	2096	2094	5	0.01	0.02	5 (0)	2	2	0.50	1.00	3 (2)
7DA14-pp	82249	82170	71068	471	345	30	0.94	0.88	32 (3)	126	7	0.69	0.72	7 (0)
7DA2-01	33880	33832	33711	45	31	5	0.68	0.82	5 (0)	14	7	0.83	0.94	9 (3)
7DA4-s	103418	103361	102476	1492	1480	13	0.81	0.68	15 (3)	2	2	0.50	1.00	3 (2)
7DA5-s	184910	184643	180701	5243	3208	22	0.91	0.82	32 (10)	2035	19	0.84	0.79	19 (0)
7DA9-s	130100	130259	129730	261	212	8	0.72	0.70	9 (1)	49	1	0.00	NA	1 (0)
7DA13-s	100425	100280	99510	298	298	7	0.66	0.67	7 (0)	0	0	1.00	0.00	0 (NA)
7DA15-s	143741	143552	142694	12589	12460	11	0.06	0.08	11 (0)	129	2	0.48	0.97	2 (0)
YL1	226774	226389	217468	42	1	1	0.00	NA	1 (0)	41	4	0.30	0.43	5 (2)
7YL	178284	177803	172465	1781	1302	13	0.57	0.47	13 (0)	468	7	0.74	0.76	7 (0)
7YL-III	65997	65556	62547	0	0	0	1.00	0.00	0 (NA)	0	0	1.00	0.00	0 (NA)
YL2	153107	152958	148028	2	2	1	0.00	NA	1 (0)	0	0	1.00	0.00	0 (NA)
YL3-01w	188511	188312	172260	126	0	0	1.00	0.00	0 (NA)	126	5	0.03	0.08	4 (2)
YL3-01s	232822	232611	210691	3	0	0	1.00	0.00	0 (NA)	3	2	0.44	0.92	2 (0)
7YL-s3	158645	158488	157145	979	898	15	0.86	0.83	15 (0)	81	4	0.07	0.14	7 (4)
7YL-s4	86408	86366	85180	207	157	4	0.60	0.76	4 (0)	50	2	0.08	0.24	2 (0)
YL4.01	200889	200588	191536	10711	10691	2	0.00	0.00	2 (0)	20	3	0.27	0.47	3 (0)
7YL-s1	124032	123877	122505	36177	36016	39	0.79	0.59	39 (0)	161	6	0.79	0.91	6 (0)
7YL-s2	85188	85072	84444	668	547	10	0.75	0.70	10 (0)	121	5	0.71	0.83	5 (0)
BL2	76395	76209	73937	12	0	0	1.00	0.00	0 (NA)	12	6	0.78	0.91	8 (3)
Blw	227966	227636	218708	0	0	0	1.00	0.00	0 (NA)	0	0	1.00	0.00	0 (NA)
7BLw1	177207	177111	171404	3	0	0	1.00	0.00	0 (NA)	3	1	0.00	NA	1 (0)
7Blw2	158088	158030	151304	1	0	0	1.00	0.00	0 (NA)	1	1	0.00	NA	1 (0)
PSBL3	127518	127300	124137	8	0	0	1.00	0.00	0 (NA)	8	4	0.66	0.88	5 (1)
PSBL2	181512	180552	177288	5	3	2	0.44	0.92	2 (0)	2	1	0.00	NA	1 (0)
PS3	150482	149968	146028	146028	118980	201	0.88	0.56	313 (7)	26537	125	0.96	0.56	138 (11)
PS	304190	303656	282495	282495	274013	602	0.94	0.51	701 (22)	5883	39	0.30	0.51	50 (8)
AS5	173226	172959	165299	165299	140613	1013	0.94	0.59	1180 (33)	23983	406	0.89	0.59	544 (13)
AS5-PI	314682	314213	288299	288299	244551	1340	0.94	0.57	1441 (18)	39873	859	0.88	0.57	912 (12)
GT	303592	303083	288086	288086	274039	656	0.87	0.53	755 (25)	14047	173	0.91	0.53	193 (11)
7GT	71483	71368	64143	64143	59788	524	0.95	0.57	735 (40)	4353	147	0.92	0.57	201 (22)
7GT-pp	235497	235113	172010	172010	132967	1495	0.97	0.57	1565 (13)	39039	227	0.77	0.57	253 (10)
GC	198582	198288	189428	32807	30700	68	0.71	0.40	80 (7)	2107	2	0.00	0.01	2 (0)
GC2	94853	93926	87922	16162	15575	64	0.83	0.50	71 (5)	587	1	0.00	NA	1 (0)
7GMcon	155254	139088	139029	202	136	16	0.55	0.50	19 (5)	66	1	0.00	NA	1 (0)
7GI2	150640	150409	138429	14250	7134	33	0.64	0.43	36 (4)	6916	18	0.64	0.48	19 (1)
NEGATIVE CONTROL A	106471	105571	96568	94548	3876	19	0.69	0.49	26 (7)	91472	351	0.87	0.69	461 (56)
NEGATIVE CONTROL B	149421	148866	140175	135739	2782	16	0.83	0.72	17 (3)	132957	484	0.97	0.72	555 (20)
Eukaryotic sequences														
BAss	320243	319549	312333	307453		308	0.68	0.25	308 (2)					
Ass	148963	148396	142049	140678		122	0.65	0.31	136 (7)					
PS	83526	83207	82325	82316		50	0.06	0.04	55 (4)					
PS3	87971	87459	86745	54		6	0.36	0.46	6 (0)					
GT	15063	14998	14883	5795		4	0.25	0.34	4 (0)					
BGt	57995	57773	56359	56269		125	0.56	0.23	125 (0)					

Extended Data Fig. 6 | Sequence data and diversity measurements. *Contaminant sequences included sequences identified in negative controls and/or high similarity to human-associated bacteria; s.e., standard error. Eventual mitochondrial and chloroplast 16S rRNA gene sequences were also removed at this step.

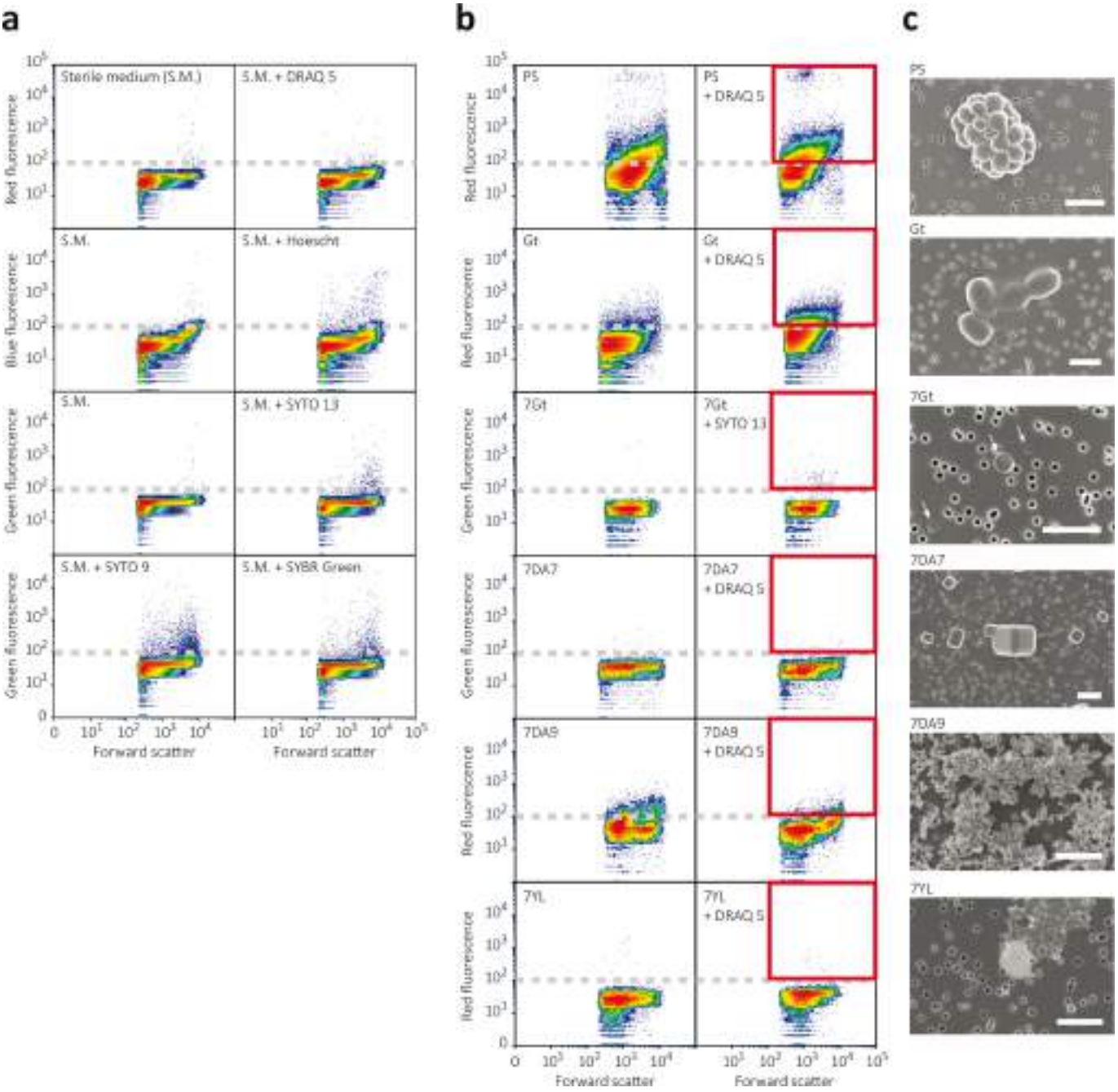


a. Peregrinibacteria
b. Saccharibacteria
c. Desulfovibrionales
d. Deferribacteres
e. Candidate division TM6
f. Candidate division NC10
g. Candidatus Latescibacteria
h. Candidatus BRC1
i. Omnitrophica
j. Zixibacteria
k. Marinimicrobia
l. Rokubacteria
m. Nitrospirae
n. Acetothermia
o. Synergistetes
p. Armatimonadetes
q. Dadabacteria
r. Deinococcus-Thermus

Extended Data Fig. 7 | Phylogenetic tree of bacterial 16S rRNA gene sequences showing the phylogenetic placement of OTUs identified in the different Dallol area samples. Sequences derived from metabarcoding studies are represented by blue lines (Illumina sequences); those derived from cloning and Sanger sequencing of environmental samples, cultures and FACS-sorted cells are labelled with a red dot. Reference sequences are in black. Concentric circles around the tree indicate the presence/absence of the corresponding OTUs in different groups of samples (groups shown in Fig. 3a). Only sequences not deemed contaminant (see Supplementary Table 5) were included in the tree. The full tree is provided as Supplementary Data 1.



Extended Data Fig. 8 | Eukaryotic presence, diversity and relative abundance in Dallol area samples. Histogram showing the phylogenetic affiliation and abundance of 18S rRNA gene amplicon reads of eukaryotes (upper panel) obtained with universal eukaryotic primers and the associated OTU diversity (lower panel). Only a few samples yielded amplicons; negative PCR controls were always negative. Sequences corresponding to macroscopic plants and fungi (probably derived from pollen or spores) were considered contaminant (light grey). The phylogenetic affiliation of dominant eukaryotic groups is colour-coded.



Extended Data Fig. 9 | Multiparametric fluorescence analyses and fluorescence-activated cell sorting (FACS) analyses of representative Dallol area samples. a, effect of DNA fluorescent dyes on background fluorescence emission; natural (sterile medium-only) and DNA dye-induced fluorescence in the sterile hypersaline SALT-YE medium used to dilute/sort Dallol samples. Fluorescence is plotted against the size of the analysed particles (forward scatter); events concentration is colour-coded, red being high concentration and blue, low concentration. DRAQ5 and SYTO13 introduced less background and were chosen for FACS of natural samples. The approximate background threshold (ca. 10^2) is indicated by a broken grey line. **b**, multiparametric fluorescence analyses of different Dallol samples before (left panels) and after (right panels) adding fluorescent DNA dyes. Events (particles) above background (red squares) were FACS-sorted and filtered on 0.1 µm pore-size filters prior to SEM observations**. c**, SEM photographs showing examples of sorted particles. Cells are observed in samples PS, Gt and 7Gt; halite crystals in 7DA7 and amorphous mineral particles in 7DA9 and 7YL. Arrows indicate ultrasmall cells. The scale bar is 1 µm.

Site	Samples	Mineral phases	
		Typical 'crystals'	Abiotic 'Biomorphs'
Cave water	Gt2016, 7Gt, 8Gt_1	Si, Ca sulfate, Fe-K sulfate, Al-Mg Fe oxides, Fe and Ca oxides	Fe-Al silicates
Lake Assale (Karum)	8Ass_2, 8Ass_3, 8Ass_4, 8Ass_6, 8Ass_7, 8Ass_8	NaCl, Na-K-Mg chloride	**Si biomorphs** (and encrustment)
Dallol dome (ponds)	Dal4.0, 7DA7_07, DAL4D, 7DA9-P1, 7CA9_P1_3, 7DA7_04, 7DA7_05, 7DA7_06, 7DA9_P1_2, 7DA9_P1_5, 7DA9_P3_10, 7DA9_P3_12	NaCl, Na-K-Mg chloride, Fe-K oxides, Ti oxides	Sulfur biomorphs, **Si biomorphs**, S-rich Na-K silicates, locally S-rich Si biomorphs, **Fe phosphates**, Fe-K phosphate, Si biomorphs – enriched in Fe, Mg, K and locally S
Yellow lake	YL1-03_4, 7YL_4, YL1-03_5, 7YL_6	Fe chloride, Mg chloride	Si, $CaCl_2$, Ca phosphate
Black lake area (ponds)	BLPS_05_5	Mg-Fe-K chloride	Mg chloride

Extended Data Fig. 10 | Mineral phases observed by SEM-EDX in precipitates of typical abiotic morphology and 'biomorphs'. Biomorphs correspond to rounded-shaped crystalline morphs resembling cell structures (cocci, rods) and compatible with cellular sizes. Observed dominant phases are highlighted in bold.

APPENDIX

B

SECOND APPENDIX

environmental microbiology reports

Environmental Microbiology Reports (2020) **12**(3), 314–323 doi:10.1111/1758-2229.12834

Brief Report

Performance of the melting seawater-ice elution method on the metabarcoding characterization of benthic protist communities

Albert Reñé,[1,2*] Adrià Auladell,[2] Guillaume Reboul,[1] David Moreira[1] and Purificación López-García[1]

[1]*Unité d'Ecologie, Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech, 91400, Orsay, France.*

[2]*Departament de Biologia Marina i Oceanografia, Institut de Ciències del Mar (CSIC), Barcelona, Catalonia, Spain.*

Summary

Massive amplicon sequencing approaches to characterize the diversity of microbial eukaryotes in sediments are scarce and controls about the effects introduced by different methods to recover DNA are lacking. In this study, we compare the performance of the melting seawater-ice elution method on the characterization of benthic protist communities by 18S rRNA gene metabarcoding with results obtained by direct cell lysis and DNA purification from sediments. Even though the most abundant operational taxonomic units were recovered by both methods, eluted samples yielded higher richness than samples undergoing direct lysis. Both treatments allowed recovering the same taxonomic groups, although we observed significant differences in terms of relative abundance for some of them. Dinoflagellata and Ciliophora strongly dominated the community in eluted samples (> 80% reads). In directly lysed samples, they only represented 37%, while groups like Fungi and Ochrophytes were highly represented (> 20% reads respectively). Our results show that the elution process yields a higher protist richness estimation, most likely as a result of the higher sample volume used to recover organisms as compared to commonly used volumes for direct benthic DNA purification. Motile groups, like dinoflagellates and ciliates, are logically more enriched during the elution process.

Received 12 April, 2019; accepted 7 March, 2020. *For correspondence. E-mail albertrene@icm.csic.es; Tel. +34932309500; Fax +34932309555.

Introduction

Studies on marine benthic protists have traditionally focused on the characterization of the diversity, distribution and function in ecosystems of morphological species based on traditional microscopy observations and cell counting (Mare, 1942; Dragesco, 1965; Fenchel, 1969). However, these studies are much scarcer than those from planktonic organisms due to difficulties in collecting, analysing and most notably, separating the cells from sediments, making it difficult to quantify them (Bak and Nieuwland, 1989). Methodologies to separate cells from soil and sediment have been developed over decades. Density gradient centrifugation has been tested to separate bacteria (Courtois *et al*., 2001) and protists (Starink *et al*., 1994) from substrate using different media. Depending on the sediment type, the cell recovery is usually high, but several limitations and uncertainties exist, like the recovery rate, biases in the recovered groups or possible adverse effects on the integrity of living cells (Robe *et al*., 2003; Parent *et al*., 2018). Some other methods to separate cells from substrate involve suspension of sediment in filtered seawater, followed by successive filtration through mesh nets of specific size pores to remove the sediments and recover and concentrate the organisms, or yet placing coverslips on top of the sediment and recovering the organisms that attach to them (Webb, 1956). However, those methods do not fully remove remaining sediment in the final sample and might result in a low and biased recovery of cells. Also frequently used, the traditional seawater ice 'Uhlig' method consists of melting seawater ice on top of a tube filled with sediment; upon melting, organisms that flow down accumulate in a Petri dish (Uhlig, 1964). Even though the recovery of cells using this method is reputed to be relatively low, it is commonly used in taxonomical studies focused on some specific groups of protists like ciliates,

dinoflagellates, diatoms and other groups of flagellates (Saburova *et al*., 1995; Azovsky *et al*., 2013; Hoppenrath *et al*., 2014).

Since the early 2000s, traditional methods used to characterize protist communities, like microscopy, have been complemented and largely displaced by molecular methods based on the use of conserved gene markers, which sidestep many difficulties associated with morphological identification (Díez *et al*., 2001; López-García *et al*., 2001; Moon-van der Staay *et al*., 2001). Currently, 18S rRNA gene metabarcoding using high-throughput sequencing (HTS) techniques provides a fast, cost effective and highly sensitive method for characterizing protist diversity in natural samples (Logares *et al*., 2012). These metabarcoding approaches are being widely applied to marine planktonic protist communities, providing insights in their diversity, composition, spatial distribution (at global or local scale) and temporal dynamics (de Vargas *et al*., 2015; Massana *et al*., 2015; Malviya *et al*., 2016; Piredda *et al*., 2016). However, studies characterizing benthic protist communities using metabarcoding are still scarce (Chariton *et al*., 2010; Quaiser *et al*., 2011; Bik *et al*., 2012; Gong *et al*., 2015; Forster *et al*., 2016; Pan *et al*., 2020; Salonen *et al*., 2019), and biases likely higher. Indeed, while genomic DNA from plankton is usually obtained from biomass retained after filtering large seawater volumes (usually litres), DNA from benthic samples is usually obtained with a direct-lysis of cells from a relatively low sediment volume or mass. Furthermore, in addition to cell lysis and DNA purification as variability sources in assessing microbial community composition, soils and marine sediments can contain detrimental amounts of potential inhibitors for downstream molecular analyses. Indeed, direct extraction methods provide higher DNA yields but lower purity, while indirect methods, which require a previous specific sample treatment to separate cells from sediment, provide lower DNA yields but of higher purity, although it is time consuming and might induce biases in microbial community characterization (Steffan *et al*., 1988; Robe *et al*., 2003). The yield of different DNA extraction methods from sediments, as well as the impact on inferences of protist diversity and community composition, has been previously assessed for cloning libraries or denaturing gel gradient electrophoresis (Lekang et al., 2015 and references therein). Some studies have focused on the effect of other factors that can greatly influence richness, sample dispersion and the structure of microbial communities. These include using different soil sample sizes (Penton *et al*., 2016), increasing DNA extraction replicates of marine sediments (Lanzén *et al*., 2017) or increasing the sequencing efforts and the number of polymerase chain reaction (PCR) replicates (Smith and Peay, 2014). Bacterial diversity studies comparing direct (direct lysis from soil) or indirect (previous cell separation using Nycodenz gradient) treatments yielded similar results for the two methods when using similar amounts of soil (Courtois *et al*., 2001; Delmont *et al*., 2011a,b). In any case, the standard methodology to obtain genomic DNA from sediment and soil consists of the direct cell lysis and DNA extraction from small amounts of sample (e.g., < 1 g) (Salonen *et al*., 2019), even though it has been proved that larger sample sizes provide a better capture of total diversity (Delmont *et al*., 2011a; Penton *et al*., 2016; Nascimento *et al*., 2018). Despite of the existing literature, the performance of the seawater-ice elution method for metabarcoding purposes has never been evaluated before. This study aims to determine how this sample treatment affects the inferred richness and composition of marine benthic protist communities. To explore this, we characterized by 18S rRNA gene metabarcoding protist communities of coastal sediment samples from the Mediterranean Sea by applying two different sample treatments: DNA purified after direct cell lysis in sediments, hereafter referred to as 'direct-lysis' samples, and DNA purified from cells separated from sediment using the 'Uhlig method', hereafter referred as 'eluted' samples.

Results and discussion

In order to test the performance of the melting seawater-ice (Ühlig) elution method to study the diversity of benthic protists, we carried out 18S rRNA gene metabarcoding analysis on a total of 72 samples issued from 18 sampling outings at different dates and three localities in the NW Mediterranean Sea (Supporting Information Table S1). Sediment from each sampling trip was subjected to the two treatments (elution and direct lysis) with two replicates per treatment, resulting in 72 samples (see *Experimental procedures* section in Supporting Information). We then generated 18S rRNA gene amplicons of approximately 550 bp encompassing the hypervariable V4 region using broad-range primers for microbial eukaryotes and sequenced them (MiSeq Illumina). Thirteen of the 72 samples issued from the direct-lysis method did not yield amplicons and were subjected to re-amplification by semi-nested PCR (Supporting Information Table S1). After quality trimming, clustering of sequences in 'swarm' operational taxonomic units (OTUs), removal of singletons and exclusion of amplicons not corresponding to protists (Supporting Information), the first inspection of the read abundances and OTU composition showed that 10 of those samples were composed of very few OTUs, mostly belonging to Fungi. We interpret this as the result of very little protist biomass per volume unit in these samples and subsequent nested-PCR-associated biases. Consequently, those samples were removed from the data set for further comparative analyses, and their counterpart replicates treated under elution method were also removed in order to have the same number of samples for each treatment. We thus retained 26 samples that corresponded to DNA purified

from small sediment volumes by direct lysis, and 26 samples that corresponded to DNA obtained from protists eluted from larger sediment volumes using the melting seawater ice method, that is, a total of 52 samples. In summary, those samples yielded 3,609,403 reads clustered into 12,518 OTUs. Samples corresponding to eluted samples yielded 1,986,751 reads and 10,447 OTUs and those corresponding to direct-lysis yielded 1,622,652 reads and 3,598 OTUs. To avoid biases in some analyses introduced by the comparison of sequence data sets of different size, we rarefied our sequence data sets when needed to the minimum number of reads observed in a sample (22,056 reads), resulting in a global data set of 1,146,912 reads and 10,142 OTUs.

Effects on the determination of the community richness

Regarding community richness, rarefaction curves showed that the diversity estimated from direct-lysis sediment samples completely saturated, while eluted samples appeared near saturation (Fig. 1A). By contrast, species accumulation curves with the addition of samples were not saturated, showing that an increase in the sampling effort would increase the observed richness (Supporting Information Fig. S1). The evaluation of OTU abundance distribution showed some OTUs comprising most reads, and many 'rare' OTUs comprising a low number of reads in eluted samples. Direct-lysis sediment samples also showed some dominant OTUs, but in comparison, the number of 'rare' OTUs was much lower (Fig. 1B). We observed higher OTU richness in eluted samples than in direct-lysis ones (analysis of variance $F_{1,50} = 88.34$; $p < 0.0001$) (Fig. 1C), and in Chao1 index ($F_{1,50} = 113.4$; $p < 0.0001$). However, they showed no significant differences when comparing Shannon ($p > 0.1$), and Simpson ($p > 0.5$) indexes (Fig. 1E and F), pointing out that even though eluted samples had higher richness, many OTUs were represented by a low number of reads. Even though DNA recovered using elution method does not correspond to all organisms present in the sediment volumes, but only to those successfully eluted, all differences observed in alpha-diversity among treatments could be attributed to the diverging initial sample volume used, with sample size of ~80 cm^3 for eluted and ~1 cm^3 for direct-lysis sediment samples, thus being close to 1:100 between both treatments.

Analysis of beta-diversity (Supporting Information Fig. S2) showed the dispersion of samples differed

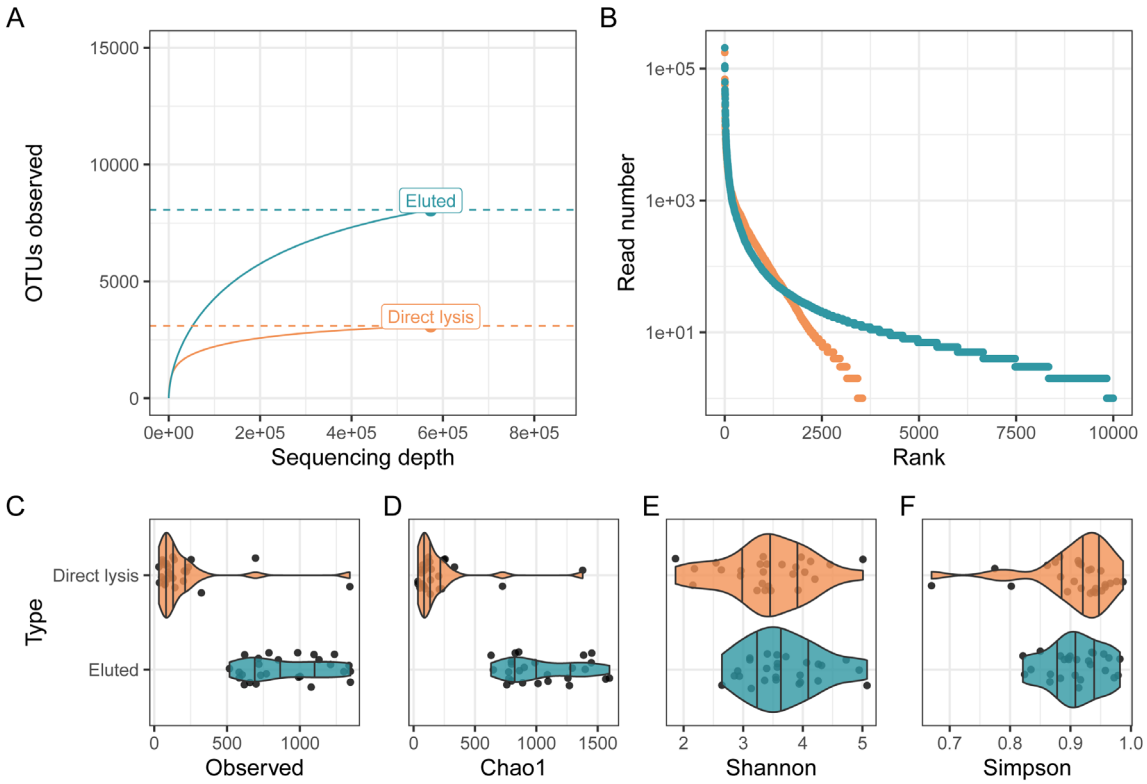


Fig. 1. Comparison of OTU richness in 'eluted' and 'direct-lysis' sediment samples. A. Rarefaction curves of both groups of samples, relating the increase in the number of reads with the number of OTUs for the complete data set. B. Distribution of the read abundances per OTU. Values of (C) richness observed, (D) Chao1 index, (E) Shannon index and (F) Simpson index for eluted and direct-lysis samples. The vertical lines in the density distribution area represent the median, 75th and 25th percentile. Direct-lysis samples are represented in orange and eluted ones in blue.

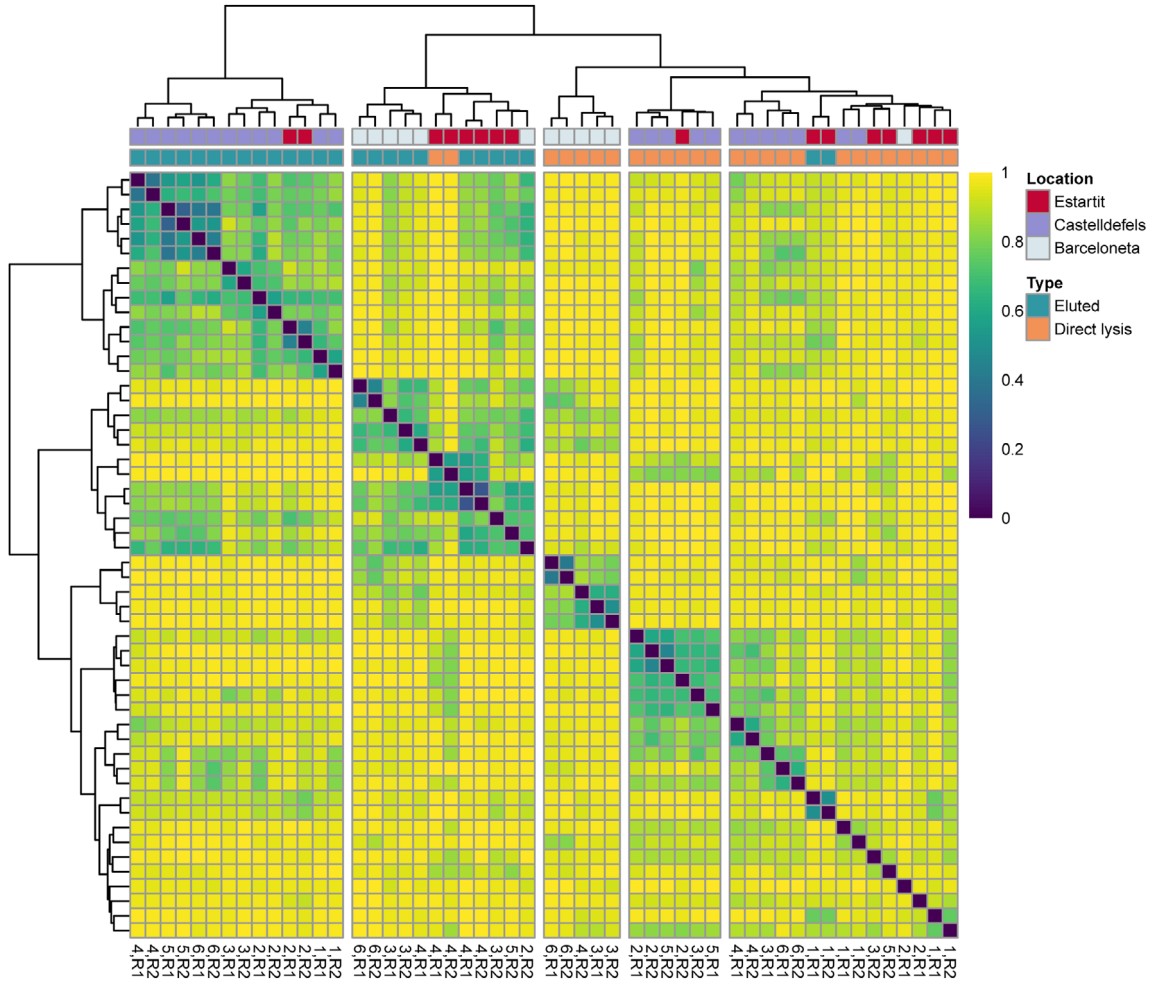


Fig. 2. Beta-diversity analyses. Heatmap showing the dissimilarity level among all samples organized by hierarchical clustering, from 0 (100% similar) to 1 (0% similar). X-axis labels indicate the month of sampling, from 1 (April) to 6 (September), and the replicates (R1 and R2) for each location and treatment type.

between treatments (ADONIS $R = 0.11$, $p < 0.001$; BETADISPER $F_{1,50} = 7.18$, $p < 0.01$), and those obtained by elution showed lower dispersion among them than the obtained by direct lysis in the multivariate dispersion analysis (MVDisp, index of multivariate dispersion = 0.562, dispersion eluted = 0.72; direct lysis = 1.28). All sample replicates clustered together. However, eluted replicates showed lower dissimilarity among them than replicates of direct-lysis samples (Fig. 2).

The used seawater-ice method does not recover all organisms present in the sample, and thus, not all DNA is eluted from the sediment volume used. However, the higher richness obtained in eluted samples confirms a higher capture of diversity when larger sample volumes are used. Given that the sequencing depth was the same for both treatments, this suggests that differences in OTU richness and saturation were due to the different volume of sample used and highlights the impact of this parameter in determining the diversity of benthic communities. For this purpose, the elution process represents around 1–2 h of time (the needed for the seawater-ice to melt and the posterior filtration of eluted sample), and does not imply remarkable extra costs than using standard direct-lysis methods.

Direct-lysis sediment samples showed a higher dispersion than eluted ones, and replicates from eluted samples, which were obtained from two different sediment cores, showed higher similarity among them than replicates obtained from subsamples from the same sediment core in samples from direct lysis. This reflects that low amounts of sediment, like those used in standard methods of direct lysis and DNA extraction (< 1 g), can lead to an incomplete characterization of the protist community, although this might ultimately depend on the density of protist cells per volume unit. Furthermore, and in contrast with previous studies (Courtois *et al*., 2001; Robe *et al*., 2003), the DNA yield obtained for eluted samples was high, probably as an effect of the higher sample volume used. The lower yield obtained for directly lysed samples might reflect a relatively

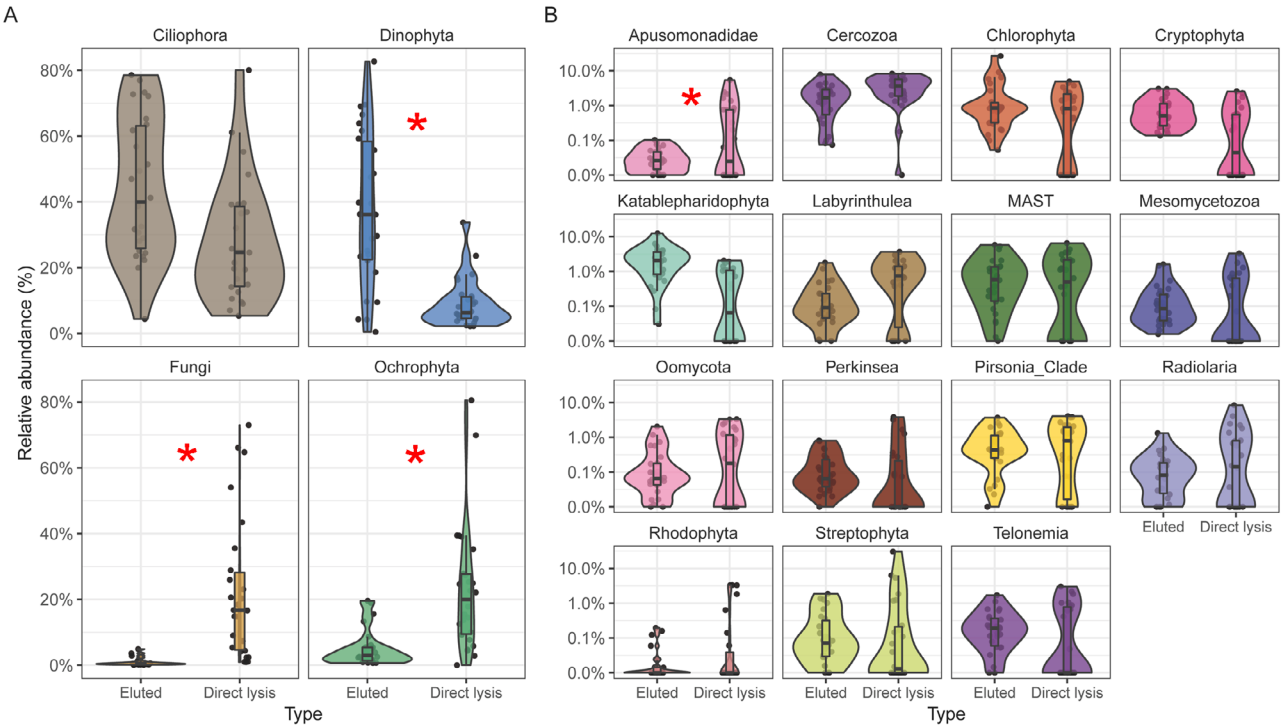


Fig. 3. Comparison of community structure for different eukaryotic phyla between eluted and direct-lysis samples. (A) Relative abundance of dominant phyla, and (B) phyla showing lower relative abundances (only those > 2% are shown). Note the logarithmic scale. Each boxplot presents the median and interquartile range of the distribution of data points shown in grey. Whiskers represent 1.5 times the interquartile range. The coloured area represents the density distribution of data points. Red asterisks indicate groups showing statistical significance in the differential abundance test between both treatments.

Table 1. Relative abundance of reads (%) of major taxonomic groups in samples pooled by sample treatment (eluted and direct-lysis).

		Total		Shared	
Superphylum	Phylum	Eluted	Direct lysis	Eluted	Direct lysis
Alveolata	Ciliophora	46.62	28.09	79.53	87.36
	Dinophyta	36.77	9.21	80.79	86.82
	Perkinsea	0.12	0.43	33.33	70.39
Archaeplastida	Chlorophyta	2.01	1.31	83.85	96.29
	Streptophyta	0.22	1.8	3.42	76.03
Hacrobia	Katablepharidophyta	2.52	0.54	31.3	96.18
	Cryptophyta	0.78	0.45	76.77	96.58
	Telonemia	0.23	0.48	84.59	58.66
Opisthokonta	Fungi	0.73	22.07	36.44	78.54
	Mesomycetozoa	0.19	0.4	92.07	99.91
Rhizaria	Cercozoa	1.66	3.78	39.1	64.3
	Radiolaria	0.23	0.92	19.09	38.23
Stramenopiles	Ochrophyta	5.03	23.03	87.95	88.1
	MAST	0.97	1.28	64.6	77.3
	Pirsonia_Clade	0.74	1.14	51.1	77.24
	Labyrinthulea	0.21	1	40.77	65.2
	Oomycota	0.21	0.76	46.22	75.48

The 'Total' columns show the percentage of reads of each taxonomic group in the whole community. The 'Shared' columns display the percentage of reads belonging to OTUs shared between both treatments. Only taxonomic groups > 0.1% in both treatments are shown.

low protist density per sediment volume unit (less eukaryotic DNA amount) and/or a lower DNA purity (e.g., metal cations or organic acids inhibiting the Taq polymerase, depending on the used DNA purification method). This likely explains the failure to amplify 18S rRNA genes by direct PCR in some samples. In agreement with our results, Penton and colleagues (2016) demonstrated that the use of larger sample sizes (e.g., 10 g in front of 0.25/1/5 g) allowed the capture of irregularly distributed abundant and rare organisms (bacterial and fungal). Nascimento and

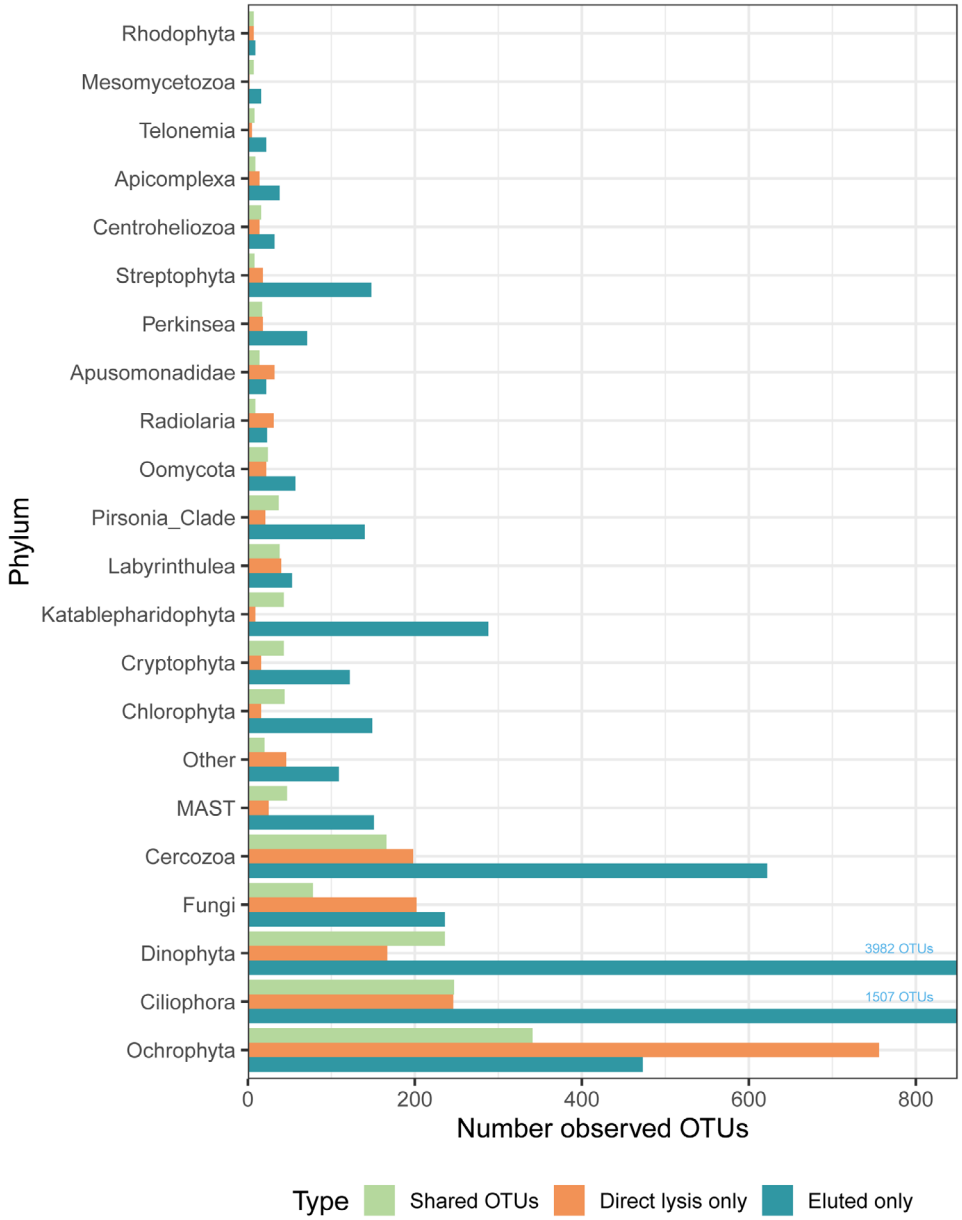


Fig. 4. Distribution of OTUs between treatments and taxonomic groups. Observed OTUs by Phylum only present in either the elution (blue), the direct-lysis treatment (orange) or in both treatments (green).

colleagues (2018) also showed that sample volume affected all protist diversity metrics investigated, being higher when increasing volumes, suggesting that sample volumes > 10 g are needed to achieve a representative assessment of alpha- and beta-diversity of microorganisms that are non-homogeneously distributed in sediments. Even if Lanzén and colleagues (2017) obtained a better representation of diversity values when increasing DNA extraction replicates than using higher amounts of sample volume, our results are congruent with those claiming that most-commonly used methods in HTS for the characterization of benthic protists lead to incomplete community determination due to the low amount of sample used and the heterogeneity of organism's distribution in the sediment. Likewise, Delmont and colleagues (2011a) concluded that the use of sample of ~100 g was sufficient to capture the majority of bacterial diversity, such that this could be used rather than increasing sampling effort, and that the major player in the estimation of community descriptors was the DNA extraction method (including direct and indirect ones). Nascimento and colleagues (2018) also demonstrated that the larger the sample volume, the more similar samples were among them. These results suggest that in protists, differences attributed to

patchy distributions might often obscure non-representative pictures of community composition due to insufficient sample volumes.

Effect on the composition of protist community

Eluted samples were clearly dominated by Alveolata: ciliates and dinoflagellates (mean of 46.6% and 36.8% reads respectively). These two groups were also abundant in direct-lysis sediment samples (28.1% and 9.2% respectively). Other groups, such as fungi (22.1%) and ochrophytes (23.0%) were also relatively abundant in direct-lysis samples, but represented lower abundances in eluted samples (Fig. 3A, Table 1). Some eukaryotic groups were present at lower percentages in both treatments without remarkable differences (Fig. 3B). The groups presenting significant differences between treatments were Dinophyta, Fungi, Ochrophyta and Apusomonadidae, (Fig. 3), confirming the selection effect (positive for Dinophyta and negative for the others) of the elution method (Supporting Information Fig. S3). The comparison of the community composition inferred from both methodologies showed that 1,459 OTUs were shared among treatments, representing 43.4% of those present in direct-lysis sediment samples and 15% of eluted ones. However, those shared OTUs represented 84% and 75% of reads, respectively, showing that most dominant OTUs were obtained by both methods, and most non-shared OTUs comprised low number of reads. This was also observed for the different taxonomic groups: shared OTUs comprised a high percentage of reads in most groups, and usually represented a fraction > 80% in direct-lysis samples (Table 1). The richness inferred for each taxonomic group (number of OTUs, regardless of their abundance) was higher (more OTUs) in eluted samples, except for Ochrophytes, Radiolaria and Apusomonadidae, which yielded a higher number of OTUs in direct-lysis samples (Fig. 4). All taxonomic groups showed similar proportions of OTUs shared between both treatments, and those unique for the direct-lysis treatment. Likewise, the exceptions were Ochrophytes and Fungi, which showed a higher proportion of OTUs unique for the direct-lysis treatment, in agreement with the higher representation in this treatment. At any rate, this difference cannot be explained by the difference in reads obtained for the two treatments. Actually, Fungi showed a similar richness of OTUs in direct-lysis samples compared to eluted samples, but the proportion of reads was much higher in direct-lysis samples (22.1%) than in eluted ones (0.73%) (Table 1). This might be explained by the lack (fungi) or limited (ochrophytes) mobility of these groups or their larger size, which might hamper their elusion from the sediment. Also, it might be that members of these groups are not (or less) active and correspond to resting stages more difficult to retrieve by the elution process.

Conversely, Dinophyta and Ciliophora showed higher richness of OTUs in eluted samples, in agreement with the relative abundances obtained (Fig. 3). Finally, some eukaryotic groups showed unexpected richness despite they were represented by a low number of reads in both treatments. This is the case of Cercozoans or Katablepharidophyta, which represent

Table 2. Relative abundance of reads (%) of different ciliate and dinoflagellate taxa in samples pooled by sample treatment (eluted and direct-lysis).

		Total		Shared	
Taxonomic group		Eluted	Direct lysis	Eluted	Direct lysis
Ciliophora	Spirotrichea	34.41	20.18	81.77	90.33
	Oligohymenophorea	5.89	1.26	63.88	76.41
	Ciliophora group 5	3.21	3.23	99.93	99.96
	Prostomatea	2.58	2.61	67.62	60.91
	Phyllopharyngea	0.22	0.41	58.97	70.98
	Colpodea	0.12	0.10	36.98	100
	Litostomatea	0.06	0.13	73.84	42.23
	Ciliophora group 7	0.02	0.08		
Dinophyceae	Peridiniales	13.35	2.34	88.44	96.06
	Uncertain Naked	12.75	0.37	78.45	99.05
	Uncertain	3.68	1.54	79.76	85.66
	Gymnodiniales	3.52	1.45	63.14	83.94
	Gonyaulacales	2.54	1.14	96.51	93.97
	Dinophysiales	0.25	0.16	46.00	97.35
	Suessiales	0.08	0.11	53.04	88.53
	Uncertain Thecate	0.07	0.08	74.55	94.69
	Prorocentrales	0.06	0.02	3.33	9.90

The 'Total' columns display the percentage of reads of each taxonomic group in the whole community. The 'Shared' columns represent the percentage of reads belonging to OTUs shared between both treatments in relation to their totality. Taxonomic groups < 0.02% in both treatments are omitted. Shaded area: no shared OTUs.

the third and fourth most diverse group in eluted samples, comprising 1.7% and 2.5% of reads respectively (Table 1).

The melting seawater-ice (Ühlig) elution method is supposed to select organisms with active motility, even though many other groups of organisms can be partially recovered just by the water flow created in the sediment column. Additionally, the mesh pore used to separate cells from the sediment only allowed to recover organisms < 60 μm. The posterior filtration on 3.0 μm filters removed those below this size. Thus, it was expected to predominantly recover motile organisms with body sizes from 3 to 60 μm, and in fact, eluted samples were dominated by taxonomic groups agreeing with those characteristics, such as ciliates and dinoflagellates (> 80% of reads). By contrast, direct DNA extraction from sediments should affect less the original composition of organisms. In direct-lysis samples, ciliates and dinoflagellates represented 37% of reads, confirming their important contribution to the community composition. However, other groups like Ochrophytes and Fungi were also highly represented in direct-lysis samples but not in eluted samples (Fig. 3), suggesting that part of this component was not recovered when using the seawater ice separation method. In any case, shared OTUs between both treatments comprised ~80% of reads in both data sets, confirming that, although differing in their relative abundance, most abundant OTUs were recovered using both methods, all lineages detected were present in both data sets and differences observed in terms of richness and taxonomic composition corresponded to low-abundant OTUs.

Dinoflagellate and ciliate community composition

Given that ciliates and dinoflagellates dominated the eluted samples and were a significant component of 'direct-lysis' sediment samples, their diversity and relative abundance was specifically compared to test possible differences among treatments (Table 2). Of all OTUs (6,385) belonging to ciliates (Ciliophora) or dinoflagellates (Dinophyceae), 483 were shared among treatments, representing 8.1% of those from eluted samples and 53.9% of those from direct-lysis in sediments (Fig. 4). All major taxonomic groups were present in both data sets, while those represented at low relative abundances (< 0.02% and all belonging to Ciliophora) were only present in eluted samples (not shown), or showed low levels of shared OTUs (e.g., Prorocentrales and Ciliophora group 7). In most cases, those shared OTUs comprised more than 85% of all reads obtained in direct-lysis samples for that taxonomic group and the percentages were generally higher than those of eluted samples.

Even though some groups like Peridiniales or 'Uncertain naked dinoflagellates' showed significant differences in their relative abundances among treatments, all taxonomic subgroups were present at abundances within the same range, confirming their dominance or rareness in the community. Thus, both treatments appear to yield a reliable characterization of dinoflagellate and ciliate communities. But, as observed for the entire community, the elution method allowed capturing higher richness of ciliates and dinoflagellates, confirming that the seawater ice 'Uhlig' treatment should be chosen when the objective of the study focuses on characterizing the community of dinoflagellates or ciliates.

Concluding remarks

We have carried out a study of the microbial eukaryotic diversity inferred by 18S rRNA gene metabarcoding in sediments after elution of protist cells by melting water in comparison with results from direct-lysis and DNA purification from sediments. We have shown that (i) alpha-diversity obtained for the elution method is much higher than the obtained for direct-lysis, likely as a result of the larger sediment volume used to obtain DNA samples. Additionally, (ii) eluted samples showed a higher similarity among them and, accordingly, reduced variability owing to stochastic subsampling effects, or patchiness of benthic communities, implying that standard methods used for metabarcoding based on small sample volumes (especially in cases of low protist density) can lead to an inadequate characterization of sample richness. We also show that although the seawater-ice elution method enriches some motile groups, it allows to recover most abundant OTUs of all taxonomic groups, although relative abundances are biased towards some of them in eluted samples. Anyway, most abundant OTUs were present in both data sets. Consequently, (iii) the seawater-ice elution seems a time and cost-efficient method that provides a more complete determination of total protist richness, especially for dinoflagellates and ciliates.

Acknowledgements

We thank R. Gallisai (ICM-CSIC) and T. Slámová (Univ. Prague) for assistance during samplings and P. Bertolino (CNRS – Univ. Paris-Sud) for technical assistance during molecular analyses. G.R., D.M. and P. L.-G. were funded by European Research Council Grant ProtistWorld (no. 322669) under the European Union's Seventh Framework Program. A.R. was granted by MINECO Grant COPAS 'Understanding top-down control in coastal bloom-forming protists' (CTM2017-86121-R), and a MECD grant 'Estancia de Movilidad en el extranjero José Castillejo' (CAS17/00237).

Conflict of interest

The authors have no conflict of interest to declare.

References

Azovsky, A., Saburova, M., Tikhonenkov, D.V., Khazanova, K., Esaulov, A., and Mazei, Y. (2013) Composition, diversity and distribution of microbenthos across the intertidal zones of Ryazhkov Island (the White Sea). *Eur J Protistol* **49**: 500–515.

Bak, R.P.M., and Nieuwland, G. (1989) Seasonal flluctuations in benthic protozoan populations at different depths in marine sediments. *Neth J Sea Res* **24**: 37–44.

Bik, H.M., Sung, W., De Ley, P., Baldwin, J.G., Sharma, J., Rocha-Olivares, A., and Thomas, W.K. (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol Ecol* **21**: 1048–1059.

Chariton, A., Court, L., Hartley, D., Collof, M., and Hardy, C. (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Front Ecol Environ* **8**: 233–238.

Courtois, S., Frostegard, A., Göransson, P., Depret, G., Jeannin, P., and Simonet, P. (2001) Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil of from cells previously released by density gradient centrifugation. *Environ Microbiol* **3**: 431–439.

Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., *et al*. (2011a) Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* **77**: 1315–1324.

Delmont, T.O., Robe, P., Clark, I.M., Simonet, P., and Vogel, T.M. (2011b) Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods* **86**: 397–400.

Díez, B., Pedrós-Alió, C., and Massana, R. (2001) Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol* **67**: 2932–2941.

Dragesco, J. (1965) Étude cytologique de quelques flagellés mésopsammiques. *Cah Biol Mar* **6**: 83–115.

Fenchel, T. (1969) The ecology of marine microbenthos. IV. Structure and function of the benthic ecosystem, its chemical and physical factors and the microfauna communities with special reference to the ciliated protozoa. *Ophelia* **6**: 1–182.

Forster, D., Dunthorn, M., Mahé, F., Dolan, J.R., Audic, S., Bass, D., *et al*. (2016) Benthic protists: the under-charted majority. *FEMS Microbiol Ecol* **92**: 1–11.

Gong, J., Shi, F., Ma, B., Dong, J., Pachiadaki, M., Zhang, X., and Edgcomb, V.P. (2015) Depth shapes α- and β-diversities of microbial eukaryotes in surficial sediments of coastal ecosystems. *Environ Microbiol* **17**: 3722–3737.

Hoppenrath, M., Murray, S., Chomérat, N., and Horiguchi, T. (2014) *Marine Benthic Dinoflagellates - Unveiling Their Worldwide Biodiversity: Kleine Senckenberg-Reihe 54*. Frankfurt am Main, Germany: Senckenberg Gesellschaft für Naturforschung.

Lanzén, A., Lekang, K., Jonassen, I., Thompson, E.M., and Troedsson, C. (2017) DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *PLoS One* **12**: e0179443.

Lekang, K., Thompson, E. M., and Troedsson, C. (2015) A comparison of DNA extraction methods for biodiversity studies of eukaryotes in marine sediments. *Aquatic Microbial Ecology* **75**: 15–25.

Logares, R., Haverkamp, T.H.A., Kumar, S., Lanzén, A., Nederbragt, A.J., Quince, C., and Kauserud, H. (2012) Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J Microbiol Methods* **91**: 106–113.

López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603–607.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., *et al*. (2016) Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci U S A* **113**: E1516–E1525.

Mare, M.F. (1942) A study of a marine benthic community with special reference to the micro-organisms. *J Mar Biol Assoc U K* **25**: 51–554.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., *et al*. (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* **17**: 4035–4049.

Moon-van der Staay, S.Y., De Watcher, R., and Vaulot, D. (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**: 607–610.

Nascimento, S.M., Lallias, D., Bik, H.M., and Creer, S. (2018) Sample size effects on the assessment of eukaryotic diversity and community structure in aquatic sediments using high-throughput sequencing. *Sci Rep* **8**: 11737.

Pan, Y., Yang, J., McManus, G.B., Lin, S., and Zhang, W. (2020) Insights into protist diversity and biogeography in intertidal sediments sampled across a range of spatial scales. *Limnol Oceanogr*. https://doi.org/10.1002/lno.11375.

Parent, B., Barras, C., and Jorissen, F. (2018) An optimised method to concentrate living (Rose Bengal-stained) benthic foraminifera from sandy sediments by high density liquids. *Mar Micropaleontol* **144**: 1–3.

Penton, C.R., Gupta, V.V.S.R., Yu, J., and Tiedje, J.M. (2016) Size matters: assessing optimum soil sample size for fungal and bacterial community structure analyses using high throughput sequencing of rRNA gene amplicons. *Front Microbiol* **7**: 824.

Piredda, R., Tomasino, M.P., D'Erchia, A.M., Manzari, C., Pesole, G., Montresor, M., *et al*. (2016) Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiol Ecol* **93**: fiw200.

Quaiser, A., Zivanovic, Y., Moreira, D., and López-García, P. (2011) Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J* **5**: 285–304.

Robe, P., Nalin, R., Capellano, C., Vogel, T.M., and Simonet, P. (2003) Extraction of DNA from soil. *Eur J Soil Biol* **39**: 183–190.

Saburova, M., Polikarpov, I.G., and Burkovsky, I.V. (1995) Spatial structure of an intertidel sandflat microphytobenthic community as related to different spatial scales. *Mar Ecol Prog Ser* **129**: 229–239.

Salonen, I.S., Chronopoulou, P.M., Leskinen, E., and Koho, K.A. (2019) Metabarcoding successfully tracks temporal changes in eukaryotic communities in coastal sediments. *FEMS Microbiol Ecol* **95**: fiy226.

Smith, D.P., and Peay, K.G. (2014) Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS One* **9**: e90234.

Starink, M., Bär-Gilissen, M.-J., Bak, R.P.M., and Cappenberg, T. (1994) Quantitative centrifugation to extract benthic protozoa from freshwater sediments. *Appl Environ Microbiol* **60**: 167–173.

Steffan, R.J., Goksoyr, J., Bej, A.K., and Atlas, R.M. (1988) Recovery of DNA from soils and sediments. *Appl Environ Microbiol* **54**: 2908–2915.

Uhlig, G. (1964) Eine einfache methode zur extraktion der vagilen, mesopsammalen microfauna. *Helgol Wiss Meeresunters* **11**: 178–185.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., *et al*. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1–11.

Webb, M.G. (1956) An ecological study of brackish water ciliates. *J Anim Ecol* **25**: 148–175.

Supporting Information

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

Appendix S1. Supporting Information.

APPENDIX

C

THIRD APPENDIX

Environmental Microbiology

Core microbial communities of lacustrine microbialites sampled along an alkalinity gradient

Miguel Iniesto[1], David Moreira[1], Guillaume Reboul[1], Philippe Deschamps[1], Karim Benzerara[2], Paola Bertolino[1], Aurélien Saghaï[1,3], Rosaluz Tavera[4] and Purificación López-García[1]

[1] Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France

[2] Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, CNRS, Muséum National d'Histoire Naturelle, Sorbonne Université, Paris, France

[3] Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Sweden

[4] Departamento de Ecología y Recursos Naturales, Universidad Nacional Autónoma de México, DF Mexico, Mexico

For correspondence: puri.lopez@u-psud.fr

Running title: Microbialite core communities in crater lakes

25

Originality and Significance Statement

Microbialites are rocks formed by microbial communities under particular physicochemical conditions. Although they are important as the oldest reliable life traces and for their capacity to sequester CO_2 as biomass and carbonates, the specific drivers influencing carbonatogenesis are not well understood. We compare the prokaryotic and eukaryotic communities associated to microbialites sampled in lakes of increasing alkalinity in the Trans-Mexican volcanic belt. We identify a conserved core microbial community populating microbialites that is more abundant in the most conspicuous microbialites, which occur in lakes with the highest alkalinity. This helps constraining microbialite formation conditions and opens interesting perspectives for the use of subsampled core communities for carbon sequestration experiments.

Summary

Microbialites are usually carbonate-rich sedimentary rocks formed by the interplay of phylogenetically and metabolically complex microbial communities with their physicochemical environment. Yet, the biotic and abiotic determinants of microbialite formation remain poorly constrained. Here, we analyzed the structure of prokaryotic and eukaryotic communities associated with microbialites occurring in several crater lakes of the Trans-Mexican volcanic belt along an alkalinity gradient. Microbialite size and community structure correlated with lake physicochemical parameters, notably alkalinity. Although microbial community composition varied across lake microbialites, major taxa-associated functions appeared quite stable with both, oxygenic and anoxygenic photosynthesis and, to less extent, sulfate reduction, as major putative carbonatogenic processes. Despite inter-lake microbialite community differences, we identified a microbial core of 247 operational taxonomic units conserved across lake microbialites, suggesting a prominent ecological role in microbialite formation. This core mostly encompassed Cyanobacteria and their typical associated taxa (Bacteroidetes, Planctomycetes) and diverse anoxygenic photosynthetic bacteria, notably Chloroflexi, Alphaproteobacteria (Rhodobacteriales, Rhodospirilalles), Gammaproteobacteria (Chromatiaceae), and minor proportions of Chlorobi. The conserved core represented up to 40% (relative abundance) of the total community in lakes Alchichica and Atexcac, displaying the highest alkalinities and the most conspicuous microbialites. Core microbialite communities associated with carbonatogenesis might be relevant for inorganic carbon sequestration purposes.

Keywords: 16S/18S rRNA metabarcoding; stromatolite; carbonate precipitation; biomineralization; cyanobacteria; anoxygenic photosynthesis

Introduction

Microbialites are organosedimentary structures formed under the influence of phylogenetically and functionally diverse microbial communities in particular physicochemical environments (Riding, 2000; Dupraz and Visscher, 2005). These geobiological structures have a double interest in ecology and evolution. First, these lithifying microbial mats are easily preserved in the fossil record and, when laminated at the macroscale (stromatolites), provide a simple morphological diagnosis for biogenicity. Applying this criterion, fossil stromatolites from the early Archaean (~3.5 Ga) are included among the oldest (almost) unambiguous life traces on Earth (Awramik, 1990; Altermann, 2004; Tice and Lowe, 2004; Allwood et al., 2006; Allwood et al., 2009). Second, formed by conspicuous photosynthetic microbial communities and being generally carbonate-rich, they constitute carbon reservoirs in the form of both, biomass and carbonates. Yet, although microbialites are thought to result from the interplay of biotic and abiotic factors (Dupraz et al., 2009), the specific identity and functions of associated microorganisms and the local environmental conditions resulting in their formation are still poorly understood.

In modern systems, both the trapping and binding of detritic particles and the *in situ* precipitation of minerals, mostly carbonates, contribute to microbialite growth. Carbonate precipitation in microbialites requires nucleation centers as well as solutions supersaturated with carbonate mineral phases, i.e. relatively rich in carbonate anions and e.g. Ca^{2+} and/or Mg^{2+} cations (Dupraz and Visscher, 2005). Exopolymeric substances (EPS), abundantly produced by many cyanobacteria, may be a source of both, cations (liberated during their degradation) and nucleation centers (Benzerara et al., 2006; Dupraz et al., 2009; Obst et al., 2009). Some microbial activities, such as oxygenic and anoxygenic photosynthesis (Dupraz and Visscher, 2005; Bundeleva et al., 2012), sulfate reduction (Visscher et al., 2000; Gallagher

et al., 2012), nitrate-driven sulfide oxidation (Himmler et al., 2018) or anaerobic methane oxidation coupled to sulfate reduction (Michaelis et al., 2002), can increase the pH and/or alkalinity ([HCO_3^-]) and, hence, the local supersaturation of the solution with carbonate phases and the precipitation kinetics. The occurrence of these activities in microbialites can be recorded in the form of isotopic signatures. Values of $\delta^{13}C$ in modern microbialites from lakes Clifton (Southwestern Australia) (Warden et al., 2016) and Alchichica (Mexico) (Chagas et al., 2016), and of $\delta^{13}C$ and $\delta^{18}O$ from Highborne Cay microbialites (Bahamas) (Louyakis et al., 2017) support the implication of these microbial activities (e.g. oxygenic and anoxygenic photosynthesis) in the formation of these lithified structures. On the contrary, other metabolisms, such as aerobic respiration, complete sulfide oxidation to sulfates and fermentation (Dupraz and Visscher, 2005) tend to promote dissolution by acidification. Carbonate precipitation would result from the balance of the different metabolisms in complex microbial communities. However, although very different taxa can display metabolisms potentially sustaining such an 'alkalinity engine', microbialite-associated microbial communities are extremely diverse (e.g. (Mobberley et al., 2012; Russell et al., 2014; Saghaï et al., 2015; Suosaari et al., 2016)) and it is difficult to determine which members have an effective role in microbialite formation. For instance, both oxygenic (cyanobacteria, eukaryotic microalgae) and anoxygenic (Chloroflexi, Chlorobi, some Alphaproteobacteria and Gammaproteobacteria) photosynthesizers should favor carbonate precipitation (Saghaï et al., 2015). However, some cyanobacterial species do favor carbonate dissolution (Guida and Garcia-Pichel, 2016; Cam et al., 2018) and others, such as cyanobacteria from the order Pleurocapsales, seem significantly more carbonatogenic than others in some systems (Couradeau et al., 2013; Gerard et al., 2013), suggesting taxon-specific effects.

Currently growing microbialites are found in a few marine sites (Logan, 1961; Dravis, 1983; Awramik and Riding, 1988; Reid and Browne, 1991; Casaburi et al., 2016; Suosaari et al., 2016) and in a variety of inland water bodies. These include saline lagoons (Saint Martin and Saint Martin, 2015), thalassohaline crater lakes (Gerard et al., 2018) and hypersaline ponds (Farias et al., 2013; Farias et al., 2014) but also freshwater systems. Freshwater microbialites raise particular interest because they appear to be more abundant in the fossil record than initially thought (e.g., (Fedorchuk et al., 2016)) and they form essentially by *in situ* mineral precipitation, like many Archean microbialites (Grotzinger, 1990). By contrast, modern marine microbialite formation involves considerable particle trapping and binding (Awramik and Riding, 1988; Reid et al., 2000). The number of discovered living microbialites in freshwater lakes is continuously increasing, with reports of microbialites displaying different morphologies and microfabrics in more than 50 lakes worldwide. Examples exist in karst areas, such as the Pavilion Lake (Laval et al., 2000), Cuatro Ciénegas (Breitbart et al., 2009) or Ruidera Pools (Santos et al., 2010), but also in volcanic terrains, such as Lake Van in Turkey (Kempe et al., 1991; López-García et al., 2005) or crater lakes (Couradeau et al., 2011; Kazmierczak et al., 2011; Zeyen et al., 2015; Johnson et al., 2018) and lagoons (Johnson et al., 2018) in Mexico. Freshwater microbialites form in lakes with very diverse hydrochemistries and usually contain one or several carbonate phases (monohydrocalcite, hydromagnesite, aragonite, calcite, dolomite) (Arp et al., 1999; Kazmierczak et al., 2011; Last et al., 2012) and often, authigenic Mg-silicates (e.g. (Arp et al., 2003; López-García et al., 2005; Souza-Egipsy et al., 2005; Reimer et al., 2009; Zeyen et al., 2015; Gerard et al., 2018; Zeyen et al., 2019)). Some studies have tried to relate microbialite mineralogy and water chemistry in individual lakes (e.g. (Lim et al., 2009; Power et al., 2011)) but comparative analyses including microbial diversity analyses are rare and limited to few systems (Centeno et al., 2012; Valdespino-Castillo et al.,

2018), such that inferring possible universal mechanisms derived from the interplay between biotic and abiotic factors is still lacking.

In a recent survey, Zeyen and co-workers (Zeyen et al., 2017) identified the occurrence of microbialites in several crater lakes (*maars*) from the Trans-Mexican volcanic belt exhibiting contrasted chemical conditions (e.g., pH, alkalinity, Mg/Ca ratios, [SO_4^{2-}]). The intensity of microbialite formation and their mineralogical composition (Mg-calcite vs aragonite vs monohydrocalcite vs hydromagnesite) strongly correlated with lake hydrochemistry (Zeyen et al., 2017). Among these lakes, the most conspicuous microbialites formed in Lake Alchichica, an alkaline (pH~9 and [HCO_3^-]~40 mM) and relatively Mg-rich ([Mg^{2+}] ~17 mM) crater lake located at high altitude (2,300 m above sea level). Lake Alchichica microbialites are dominated by hydromagnesite ($Mg_5(CO_3)_4(OH)_2.4(H_2O)$) and aragonite ($CaCO_3$) (Kazmierczak et al., 2011; Couradeau et al., 2013), and several studies have focused on the associated microbial communities (Couradeau et al., 2011; Valdespino-Castillo et al., 2018) and their functional potential derived from metagenomic analyses (Saghaï et al., 2016). Here, we characterize the prokaryotic and eukaryotic community composition of microbialites detected in several Trans-Mexican volcanic belt crater lakes following an alkalinity gradient (Zeyen et al., 2017) by massive 16S/18S rRNA gene amplicon sequencing. Comparative analyses reveal the existence of a common core of microbial taxa associated with these microbialites, which might play a determinant role in their formation.

Experimental procedures

Sampling

Microbialite samples were identified and collected during two field trips (January 2012 and May 2014) from 9 out of 11 visited lakes located in the Trans-Mexican volcanic belt (Fig. 1 and Supporting Information Fig.S1). The physicochemical parameters of lake waters (Supporting Information Table S1) were measured in situ using a multiparameter probe (Multi 350i, WTW). Alkalinity and cation/anion concentrations were analyzed from water samples collected during the 2014 expedition and reported by Zeyen et al. (Zeyen et al., 2017). Parameters for Rincon del Parangueo were obtained from Armienta et al. (Armienta et al., 2008). To limit potential biases linked to microbialite heterogeneity, microbialite fragments were collected in replicates and, for some lakes, at different locations along the shore and/or at different depths or season, with the help of a hammer and sterile chisels/forceps. In total, we collected and analyzed 30 microbialite and mineral-associated biofilm samples (Table 1) as well as two non-calcifying microbial mat samples from Rincon del Parangueo. Sample fragments were fixed in situ with EtOH (>80% v/v) and subsequently stored at –20°C.

DNA purification and amplicon sequencing

Microbialite fragments were ground using a sterile agate mortar. DNA purification was carried out as previously described (Saghaï et al., 2015), using the Power Biofilm™ DNA Isolation Kit (MoBio, Carlsbad, CA, USA) with extended incubation in the kit resuspension buffer (>2h at 4°C for rehydration) and bead-beating steps. Archaeal and bacterial 16S rRNA gene fragments (~290 bp long) covering the V4-hypervariable region were amplified using the prokaryote-specific primer set U515F (5'-GTGCCAGCMGCCGCGGTAA) and U806R (5'-GGACTACVSGGGTATCTAAT). Eukaryotic 18S rRNA gene fragments (~600 bp long) also encompassing the V4-hypervariable region were PCR amplified using the primers EK-448F (5'- CTGAYWCAGGGAGGTAGTRA) and 18s-EUK-1134-R_UNonMet (5'-

TTTAAGTTTCAGCCTTGCG) biased against Metazoa (Bower et al., 2004). Forward and reverse primers were tagged with different 10-bp molecular identifiers (MIDs) to allow pooling and later identification of amplicons from different samples. The 25-µl PCR-amplification reaction contained 0.5-3 µl of eluted DNA, 1.5 mM $MgCl_2$, 0.2 mM of deoxynucleotide (dNTP) mix, 0.3 µM of each primer and 0.5 U of the hot-start Platinum Taq DNA Polymerase (Invitrogen, Carlsbad, CA). PCR reactions were carried out for 35 cycles (94°C for 30 s, 55-58°C for 30-45 s, 72°C for 90 s) preceded by 2 min denaturation at 94°C, and followed by 5 additional minutes of polymerization at 72°C. To minimize PCR bias, 5 different PCR reactions were pooled for each sample. Amplicons were then purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany). Amplicons were massively sequenced using Illumina MiSeq (2x300 bp, paired-end) by Eurofins Genomics (Ebersberg, Germany). Sequences have been deposited in GenBank under the BioProject number PRJNA625182. Individual biosample accessions are listed in Supporting Information Table S2.

Sequence analysis

We obtained 2 270 503 and 4 886 605 sequence-reads of 16S and 18S rDNA amplicons, respectively. Raw sequences were processed using an in-house bioinformatic pipeline. High-quality raw 16S rDNA paired-end reads were merged together according to strict criterions using FLASH (Magoc and Salzberg, 2011). Cleaned merged reads with correct MIDs at each extremity were attributed to their original samples and pruned of primer+MID sequences using 'cutadapt' (Martin, 2011). In the case of 18S rDNA sequences, we used high-quality forward reads since, due to the amplicon size, too few read pairs could be assembled reliably. High-quality (merged) reads were dereplicated to retain unique sequences for further analyses while keeping trace of their corresponding amounts using VSEARCH (Rognes et al., 2016). Chimeric

high-quality reads were detected de novo with VSEARCH and excluded from further analyses. Non-chimeric (merged) high-quality reads were then pooled together in order to define inter-sample Operational Taxonomic Units (OTUs) using SWARM (Mahe et al., 2015) and CD-HIT (Fu et al., 2012) at 97 and 98% sequence identity (Table 1; Supporting Information Table S2). The number of prokaryotic OTUs obtained was of the same order of magnitude for the two approaches. However, CD-HIT resulted in an inflation of eukaryotic OTUs as compared with SWARM and previous results based on whole Alchichica microbialite metagenomes (Saghaï et al., 2016). Therefore, we chose SWARM-derived OTUs for subsequent analyses. Singletons (OTUs composed of one sequence) were removed from subsequent analyses. OTUs were phylogenetically classified based on sequence similarity with sequences from cultured/described organisms and environmental surveys retrieved from SILVAv128 for prokaryotic and eukaryotic rDNA sequences (Quast et al., 2013) and additionally from PR2v4.5 for eukaryotic rDNA, (Guillou et al., 2013) and stored in a local database. OTUs corresponding to chloroplasts, mitochondria and Metazoa were removed from subsequent analyses. Sequences with low identity values were manually blasted and assigned to their best hit's taxon when they combined coverage and identity values >80% and >85%, respectively. Prokaryotic OTUs (103) whose identity with their best hit ranged between 75 and 85% were placed in a reference phylogenetic tree and, upon manual inspection to verify their placement within a robust monophyletic group, reassigned accordingly (trees in Newick format are provided as supplementary files). To this end, 16S/18S rDNA reference sequences covering the tree-of-life diversity (Hug et al., 2016) and near-complete OTU best-hit sequences were aligned using MAFFT (Katoh and Standley, 2013); ambiguously aligned sites were removed from the alignment using trimAl (Capella-Gutierrez et al., 2009). The reference phylogenetic tree was then built with IQtree (Nguyen et al., 2015) using the GTR+G+I model of sequence evolution.

To align our OTU reads to the reference alignment, we used the --addfragments function of MAFFT (with the highly accurate option L-INS-I). Finally, reads were placed into the reference phylogenetic tree using the alignment files and the reference tree with the EPA-ng tool (Barbera et al., 2019). Genesis library (Czech et al., 2020) was used to create a NEWICK format tree out of the resulting EPA-ng JPLACE-format tree. When the phylogenetic affiliation in the reference tree was not conclusive, the OTUs remained 'uncertain'.

Predictive functional profiling of microbial communities

Several microbial taxa (down to the family or genus) are systematically associated to particular broad metabolisms and their relative abundance can be therefore used for tentative metabolic prediction (Langille et al., 2013) (Martiny et al., 2015). Based on this approach, we established 10 broad metabolic categories readily attributable to specific taxa: oxygenic photosynthesis, anoxygenic photosynthesis (subdivided according to whether it was carried out by green non-sulfur bacteria (GNSB, Chloroflexi), purple sulfur bacteria (PSB, photosynthetic Gammaproteobacteria) or purple non-sulfur bacteria (PNSB, photosynthetic Alphaproteobacteria), sulfate reduction, nitrification, denitrification, hydrogen oxidation, heterotrophy and fermentation. The different OTUs, including relative abundance data, were subsequently distributed in these categories based on the known metabolism of the family or genus it was confidently affiliated to (Supporting Tables S4-S5). Whenever this was not confidently possible they were included in one additional category comprising OTUs of uncertain metabolism.

Statistical analyses

Statistical analyses were carried out in R (R Development Core Team, 2017). Diversity indexes and non-metric multidimensional scaling (NMDS) ordination analyses were conducted using the 'Vegan' R package (Oksanen et al., 2011). Community structures across microbialite samples were compared using Bray–Curtis (BC) dissimilarities (Bray and Curtis, 1957) based on Wisconsin-standardized OTU relative frequencies to balance the weight of abundant versus rare OTUs. To test whether microbial diversity was significantly correlated to environmental variables, we carried out a Mantel test (Legendre and Legendre, 1998) between the BC distance matrix and a matrix of Euclidean distances of physicochemical parameters (mineral composition and depth) using the 'Vegan' package. Canonical Correspondence Analyses (CCA) to explore the cross-variance of our datasets were calculated with the 'Ade4' package (Dray and Dufoour, 2007). Permutational multivariate analysis of variance (PERMANOVA) (Legendre and Legendre, 1998) tests were also carried out with 'Vegan' to quantify the influence of individual variables on community structure.

Results and discussion

Microbialites in lakes of the Trans-Mexican volcanic belt

Microbialites in the alkaline (pH ~9) crater Lake Alchichica are meter-sized and their chemical and mineralogical composition, microbial diversity and metagenome-derived functional potential have been studied for several years (Couradeau et al., 2011; Kazmierczak et al., 2011; Centeno et al., 2012; Couradeau et al., 2013; Gerard et al., 2013; Saghaï et al., 2015; Saghaï et al., 2016; Valdespino-Castillo et al., 2018; Zeyen et al., 2019). However, calcifying microbial communities in other alkaline lakes with comparable hydrochemistry from the same volcanic area (Armienta et al., 2008; Mancilla Villa et al., 2014; Zeyen et al., 2017) remain

largely understudied. We carried out two field campaigns to explore and eventually collect microbialites from other lakes in the Trans-Mexican volcanic belt. In total, we visited eleven lakes in the Puebla and Michoacan regions, nine of which harbored calcifying microbial structures (Fig.1; Supporting Information Fig.S1 and Table S1). Based on their hydrochemistry, these lakes locate along an alkalinity gradient (Zeyen et al., 2017) (Fig.1), with more developed microbialites in lakes showing a higher alkalinity (e.g. Alchichica, Atexcac). Lower alkalinity systems, such as La Alberca de Michoacan, harbored calcifying biofilms growing on basalt rocks. Neither Lake Zirahuen, with the lowest alkalinity value, nor Rincon del Parangueo, an almost completely evaporated lake with residual hypersaline ponds (conductivity 165 mS/cm; Table S1), harbored actively growing calcifying communities (Rincon del Parangueo exhibited subfossil, dried microbialites) (Supporting Information Fig.S1 and Table S1). We analyzed samples of floating, non-calcifying halophilic microbial mats from Rincon del Parangueo, as well as 30 microbialite samples from microbialite-containing lakes. These samples included replicates and, in some cases, were collected at different depths and location along the shore (Table 1). This allowed comparing microbial community composition across lakes with different hydrochemistries and studying the abiotic factors determining it.

Overall microbialite community structures

After DNA purification from microbialite samples, we amplified and high-throughput-sequenced 16S and 18S rRNA gene amplicons. High-quality sequences were used to define operational taxonomic units (OTUs), with a total of 17 559 prokaryotic OTUs (766 archaeal, 16 793 bacterial) and 3 769 eukaryotic OTUs, excluding singletons (Table 1; Supporting Information Tables S2,S4-S5). The diversity of microbialite communities was high and even, as reflected by indices of richness (chao1 and ACE), diversity (Shannon and Simpson) and evenness

(Pielou) (Supporting Information Table S3). For both, prokaryotes and eukaryotes, the relative proportions of OTUs belonging to high-rank taxa were more similar than the relative abundance of reads (Fig.2). This likely reflects the high heterogeneity of these structures with local abundance (but not OTU diversity) changing at local spatial scale. Nonetheless, in general, replicate samples exhibited consistent profiles reflecting similar trends in terms of community structure.

We identified OTUs belonging up to 112 different prokaryotic phyla or equivalent high-rank taxa, most of them bacterial. Four major groups dominated, albeit in different proportions, three of which include photosynthetic members: Cyanobacteria, Alphaproteobacteria, Chloroflexi and Planctomycetes. Altogether, they averaged 66 ± 16% of total reads, with a maximum of 88% at Alberca de los Espinos. However, in some microbialites other groups were also relatively abundant (up to ca. 15-25%), such as Gammaproteobacteria in Tecuitlapa, Deltaproteobacteria in La Preciosa and Actinobacteria in La Alberca de Michoacan (Fig.2A). Cyanobacteria were, on average, the most represented group, especially in lakes Alchichica and Atexcac, often comprising more than 50% of the reads. We identified 712 cyanobacterial OTUs mostly belonging to the Oscillatoriales and diverse lineages in the polyphyletic order Synechococcales (notably *Leptolyngbya*) (Supporting Information Fig.S2A and Table S3). Pleurocapsales were present, but were not the most abundant cyanobacterial group in the collected surface microbialites. This agreed with previous observations in Alchichica showing that members of this group increased in abundance at higher lake depth (Couradeau et al., 2011; Saghaï et al., 2015). Alphaproteobacteria were highly diverse and included an important proportion (often >50%) of likely photosynthetic Rhodobacterales and Rhodospirillales (Supporting Information Fig.S2B). In addition, many other bacterial lineages appeared in smaller amounts, including anoxygenic photosynthetic Chlorobi and various typically

heterotrophic taxa (Supporting Information Fig.S3A). Archaea were detected only in very minor proportions (generally <1 to 5%), in agreement with previous observations (Saghaï et al., 2015; Saghaï et al., 2016). However, in a few replicate samples (Alberca de los Espinos, Patzcuaro) they represented up to ~10%. Diverse Euryarchaeota (including several methanogenic lineages), Thaumarchaeota and Woesearchaeota were the most abundant archaea (Supporting Information Fig.S3B).

Microbial eukaryotes (metazoan sequences were excluded from the analysis) were also very diverse, although they represent a minor fraction (ca. 5-10%) of the bacteria-dominated microbialite communities, as shown by metagenomic studies in Alchichica (Saghaï et al., 2015; Saghaï et al., 2016), (Fig.2C-D; Supporting Information Fig.S4). Photosynthetic lineages dominated (>50%) both in terms of OTU diversity and, especially, relative sequence read abundance in most microbialites. Chlorophyta (Archaeplastida) and Ochrophyta (Stramenopila, mostly diatoms) were highly represented. Dinoflagellates, haptophytes and euglenozoans were also present. Only in the case of Alberca de Michoacan, the relative amount of reads in the two replicates suggested a higher dominance of heterotrophic eukaryotes, consistent with a high grazing activity and the presence of relatively thin calcifying biofilms (Supporting Information Fig.S1H). Ciliates were the most abundant grazers (although their diversity and abundance were likely inflated by the presence of intraspecific variation and multiple gene copies (Wang et al., 2017), followed by cercozoans and heterotrophic stramenopiles, depending on samples. Together with ciliates, fungi were the most abundant eukaryotic heterotrophs (Fig.2). The observed eukaryotic diversity needs to be interpreted with caution due to potential intra-species or intracellular 18S rRNA gene variation (Weisse, 2002; Decelle et al., 2014).

The overall observed community composition across microbialite samples is consistent with that observed by previous studies of Lake Alchichica microbialites (Saghaï et al., 2015; Saghaï et al., 2016). At the level of high-rank taxa, the high relative abundance of Cyanobacteria and Alphaproteobacteria within bacteria, green algae and diatoms within eukaryotes and the minor presence of archaea are general trends observed in marine and other lacustrine microbialites (López-García et al., 2005; Papineau et al., 2005; Havemann and Foster, 2008; Foster and Green, 2011; Centeno et al., 2012) but also in many non-lithifying microbial mats (Harris et al., 2013; Wong et al., 2016; Gutierrez-Preciado et al., 2018). In the non-calcifying mats sampled in the terminal desiccating system of Rincon del Parangueo, although Cyanobacteria were the most abundant bacterial group, Firmicutes and Deinococcus-Thermus were also very abundant, together with Bacteroidetes and Gammaproteobacteria (Supporting Information Fig.S5). Since the diversity of these non-calcifying mats was significantly different from that of microbialites in other Trans-Mexican crater lakes, these samples were excluded from subsequent comparisons.

Comparison of microbialite community structures across lakes and influence of abiotic parameters

To evaluate the degree of similarity of microbial communities associated with the different Mexican microbialites, we built a correlation matrix using Bray-Curtis (BC) distances taking into account OTU presence/absence and frequency (Supporting Information Fig.S6). We then applied ordination methods based on these BC distances, such as NMDS and hierarchical cluster analysis (HCA). NMDS showed most microbialite samples scattered between the two main axes, although there is a clear trend distributing lake samples according to their relative alkalinity along axis 1 (Fig.3; Supporting Information Fig.S7). Notably, all Alchichica and

Atexcac samples were situated on the left of axis 1, with two Atexcac samples tightly clustered with Alchichica microbialites (Fig.3A). This trend was equally observed in the cluster analysis. Replicate samples always clustered together (Fig.3B). PERMANOVA tests showed that differences between microbialites from different lakes were significant (p-value < 0.001, R^2=0.8499). Differences between microbialites of the various lakes were associated with differences in their prokaryotic communities. Indeed, HCA and NMDS excluding eukaryotic taxa from the test resulted in almost the same ordination and clustering pattern. By contrast, ordination analysis of eukaryotic OTUs produced mixed patterns instead (Supporting Information Fig.S8). This likely reflects the more random capture of grazing protists in the different samples, which superposes to that of the integral members of the microbialite biofilms (e.g. green algae, diatoms).

A Mantel test showed a significant correlation between the physicochemical parameters and the prokaryotic community structure matrices (p-value = 0.006). Canonical Correspondence Analyses (CCA) further revealed the influence of different physicochemical parameters on the microbialite community structure across the different lakes. The correlations observed were mostly driven by the response of prokaryotic communities, as shown by CCA including or excluding the eukaryotic component and taking into account all the measured abiotic parameters (Supporting Information Fig.S9). Among the measured physicochemical parameters of the lakes, pH, conductivity, alkalinity (i.e. [HCO_3^-]), [Ca^{2+}] and the [Mg^{2+}]/[Ca^{2+}] ratio appeared the most relevant, explaining up to 22.7% of the variance (Fig.4). The microbial community composition in Alchichica and Atexcac microbialites was most influenced by high conductivities and alkalinities. The difference in microbial community structure of Alberca de los Espinos and Patzcuaro microbialites compared with other microbialites correlated with

400 [Ca^{2+}], while the structures of the microbialite communities in Alberca de Michoacan correlated with pH.

Taxon-based metabolic profiling of microbialite communities

405 Some microbial metabolisms, notably photosynthesis and sulfate reduction, can promote carbonate precipitation, based on the general consideration that they usually consume protons (Dupraz et al., 2009) as well as observations in the field (Visscher et al., 2000; Couradeau et al., 2013; Gerard et al., 2013; Pace et al., 2016). These metabolisms, unlike others, can be phylogenetically associated with specific microbial taxa (Martiny et al., 2015). Recent studies 410 showed a strong correlation between the phylogenetic composition of microbial communities and their predicted metabolic activities (Morrissey et al., 2019). These predictions of broad metabolic classes (photosynthesis, sulfate reduction, heterotrophy) are consistent with predictions made from protein-coding genes in previous metagenomic analyses of Alchichica microbialites (Saghaï et al., 2015; Saghaï et al., 2016). Therefore, taxon-based metabolic 415 profiling provides a reasonable working hypothesis about dominant metabolisms, which should be further validated by metagenomic and/or metatranscriptomic analyses. As shown in Fig.5A, potential carbonatogenic metabolisms (essentially photosynthesis and sulfate reduction in our microbialites) were clearly dominant (>50% reads and up to ~70%) in several lakes, including Atexcac and Alchichica, harboring the most apparent microbialites, but also Quechulac and 420 Alberca de los Espinos. Microbialites from Alberca de Michoacan, Aljojuca and La Preciosa harbored between 40-50% of prokaryotes carrying out typical carbonatogenic metabolisms, whereas Patzcuaro showed the lowest values (25%). These are minimal values, since part of the organisms within the “uncertain” category might also promote carbonate precipitation. Also,

although eukaryotes represent relatively minor proportions (5-10%) of the total community, at least in Alchichica microbialites (Saghaï et al., 2015), photosynthetic eukaryotes may also contribute to it. At the same time, these values only correspond to metabolic potential and need to be taken as cautionary proxies for carbonatogenesis for two reasons. First, not all the organisms carrying out one of those metabolic activities do actually promote carbonate precipitation in situ (for instance, some cyanobacterial borers dissolve rather than trigger carbonate precipitation). Second, these values correspond to the relative abundance of OTU sequence reads (as a proxy for organisms) and not to direct activity. Although in principle dominant community members are likely active in the community, the intensity of these activities may vary and, therefore, transcriptomic or direct metabolic measurements will be needed to validate or refine their actual contribution to these different metabolisms.

It is interesting to note that anoxygenic photosynthesis was well represented in all the observed microbialites, with photosynthetic Chloroflexi and Alphaproteobacteria members appearing as dominant players, except in Tecuitlapa, a more eutrophic, less oxygenated lake, where photosynthetic gammaproteobacteria (Chromatiaceae) slightly dominated over photosynthetic alphaproteobacteria. Actually, the relative contribution of anoxygenic over oxygenic photosynthesis seemed more important in some systems (Quechulac, La Preciosa, Tecuitlapa). Overall, our observations in Transmexican belt volcanic lake microbialites confirm and extend previous studies suggesting an important potential contribution of anoxygenic photosynthesis to microbialite formation (Ionescu et al., 2014; Saghaï et al., 2015; Gerard et al., 2018).

Based on BC distances calculated on metabolic profiles, microbialite samples appeared interspersed in NMDS analysis (Fig.5B). In agreement, differences in the metabolic potential profiles between lakes were not significant according to pairwise PERMANOVA tests

(Supporting Information Table S6). The same trend was observed when microbialites were grouped in categories according to their massiveness (well developed –Alchichica and Atexcac–, medium-to-modest structures –Alberca de los Espinos, La Preciosa, Aljojuca, Quechulac, Patzcuaro and Tecuitlapa–, and thin calcifying biofilms –Alberca de Michoacan–)(Supporting Information Table S7). These observations suggest a stability of broad metabolic functions expressed at the microbialite ecosystem level, despite variations of microbialite communities between the different crater lakes (Fig.3). Similar trends have been observed in other types of settings (Louca et al., 2016). Our metabolic profile results complement others obtained in marine systems and collectively highlight the importance of community metabolisms in interplay with local conditions for microbialite formation (Casaburi et al., 2016; Ruvindy et al., 2016). In addition, the influence of photosynthesis (both oxygenic and anoxygenic) or sulfate reduction (especially at Tecuitlapa, La Preciosa and Aljojuca) is consistent with isotopic signatures detected in modern microbialites (Chagas et al., 2016; Louyakis et al., 2017; Foster et al., 2020) from different locations.

Shared microbial core across lake microbialites

Although the microbialite-associated prokaryotic and eukaryotic communities were different among lakes (Figs.2-4), we asked whether a conserved microbial core existed across these calcifying communities as this core might play a relevant role in microbialite formation. To limit biases due to local heterogeneity, we compared the collection of microbialite-associated OTUs collectively identified in each lake (only 10 OTUs were actually shared by the 30 samples considered independently). We detected a ‘restricted core’ of 106 microbialite-associated OTUs shared by the nine lakes (24 prokaryotic, 82 eukaryotic; Fig.6). We then slightly relaxed our criteria and search for OTUs shared by microbialites from eight out of the nine sampled

lakes. This defined an 'extended core' comprising 247 OTUs (91 prokaryotes, 156 eukaryotes; Fig.6). The prokaryotic extended core included 17 cyanobacterial OTUs (7 *Leptolyngbya*-related, 2 *Synechococcus*-like, 1 member of Pleurocapsales) and 13 alphaproteobacterial OTUs (with at least 6 OTUs from families of anoxygenic photosynthesizers), among others, including one methanogenic archaeon (Supporting Information Table S8). In total, 23 OTUs corresponded to prokaryotes carrying out potentially carbonatogenic metabolisms (essentially oxygenic and anoxygenic photosynthesis). Interestingly, OTUs belonging to the prokaryotic core represented up to ~40% in relative abundance of the total microbialite community in lakes Alchichica and Atexcac, where the most massive structures are found (Fig.6A). These values fell to 20-25% for Tecuitlapa, La Preciosa, Alberca de los Espinos and Alberca de Michoacan and 15% or less in Aljojuca, Quechulac and Patzcuaro. This suggests that those OTUs represent community members associated with actively growing microbialites. Some of them might actually trigger carbonatogenesis via their metabolic activities, notably the photosynthetic members, but other core OTUs, such as those of Planctomycetes or Bacteroidetes, might simply be specifically associated with the core photosynthetic OTUs as degraders of exopolymeric substances.

The extended eukaryotic core included 82 OTUs of photosynthetic members, mostly diatoms and green algae, but also a few representatives of other groups (stramenopiles, dinoflagellates, haptophytes and cryptophytes; Supporting Information Table S9). The rest of eukaryotic OTUs corresponded to some fungi and to typical grazers that are not strictly associated with the microbialites but might be common predators on biofilm surfaces in the different crater lakes. The shared eukaryotic OTUs represented a high proportion of the total eukaryotic community (>60 and up to ~90% reads; Fig.6B). However, eukaryotes are likely minor components (<5-10%) in the total community as suggested by metagenomic studies in

Alchichica microbialites (Saghaï et al., 2015; Saghaï et al., 2016). In addition, the relatively high diversity of core eukaryotic OTUs associated with microbialites might be inflated due to 18S rDNA intraspecific (Weisse, 2002) and/or intracellular variation (Decelle et al., 2014) and the higher number of eukaryotic sequences analyzed.

The occurrence of a distinct microbial core in microbialites as compared to plankton has been previously noted in some freshwater systems (White et al., 2016). However, to our knowledge, this is the first time that a core of prokaryotic and eukaryotic OTUs is detected in microbialites from lakes of varying physicochemistries using the same criteria across samples treated in the same way, thus minimizing confounding factors. Therefore, the identified microbial core across freshwater microbialites is ecologically relevant and corresponds to microorganisms that are intimately associated with calcifying mats, some of which likely trigger carbonatogenesis (e.g. photosynthesizers), and others specifically depending on them (EPS-degraders, calcifying biofilm grazers). A similar approach has been applied to the study of coral microbiomes to identify important microbial components in coral holobionts, also including potentially carbonatogenic members (karHernandez-Agreda et al., 2017).

Concluding remarks

Microbialite formation results from the fine-tuned interplay of biotic and abiotic factors. To better understand and constrain those factors, we have analyzed the composition of both, prokaryotic and eukaryotic communities associated with microbialites sampled in a series of crater lakes from the Trans-Mexican volcanic belt that follow an alkalinity gradient. We identify a clear correlation between the composition of calcifying communities and lake alkalinity, accompanying the observation that more massive structures actively form in high-alkalinity lakes Alchichica and Atexcac (Figs.1 and 3). Although the microbial communities differ across

lake microbialites, there are conserved trends. These include the high relative abundance of Cyanobacteria and their typical EPS-degrading associated taxa (Bacteroidetes, Planctomycetes) and that of anoxygenic photosynthetic bacteria, notably Chloroflexi and some Alphaproteobacteria (Rhodobacterales, Rhodospirillales), but also some Gammaproteobacteria (Chromatiaceae) and minor proportions of Chlorobi. Green algae and diatoms, together with ciliate and cercozoan grazers are the most relatively abundant eukaryotes. Based on the metabolic potential of the dominant microbial taxa, it clearly appears that both, oxygenic and anoxygenic photosynthesis are important players in carbonatogenesis, with minor contributions from sulfate reduction (Fig.5). However, although the photosynthesis-related carbonatogenic metabolic potential appears higher in the most conspicuous microbialites (Alchichica, Atexcac), it is also the case in other, less massive calcifying structures. This suggests that local physicochemical conditions play a crucial role and that the specific components of the microbial community contribute differently to carbonatogenesis, either due to different phylogenetic components and/or to different expression levels. Transcriptomic and/or functional analyses in situ should help to better constrain these contributions (Mobberley et al., 2015). Despite these differences, we identified a shared conserved core of prokaryotic and eukaryotic OTUs across lake microbialites. Interestingly, this microbial core represents a higher relative abundance (up to 40% of the total community) in lakes with more conspicuous microbialites (Fig.6). This advocates for a relevant, if not causal, role of these microorganisms in microbialite formation.

The identification of microbialite communities that actively favor carbonate precipitation under certain abiotic conditions has potential applied implications in the context of global climate change. Capture and storage of carbon is a serious option to mitigate the effects of atmospheric greenhouse gas emission and climate change. While some vegetated

ecosystems are active carbon sinks (Blue Carbon ecosystems), the contribution of microbial communities is not yet well constrained (Macreadie et al., 2019). The ability of microbialite communities to fix CO_2 as biomass and, especially, carbonates makes them interesting as potential sequestration systems. The biomineralization of calcium carbonates by bacteria has long been used for the remediation of concrete and damaged heritage buildings (Dhami et al., 2013; Seifan and Berenjian, 2019) and some tests using cyanobacterial mats favoring hydromagnesite precipitation have been carried out in laboratory (McCutcheon et al., 2014). Our study suggests that microbial consortia similar to the microbial core community identified in Mexican microbialites may be used for carbon sequestration following a more biomimetic approach than the use of axenic strains. For that purpose, future studies should identify which of the two strategies, axenic culture versus consortium-based, are the most efficient in carbon sequestration.

Acknowledgments We thank Anabel Lopez-Archilla, Nina Zeyen and Eleonor Cortes for help and discussions during the 2014 field trip. This research was funded by the European Research Council Grants ProtistWorld (322669, PL-G) and PlastEvol (787904, DM) and the French ANR project Microbialites (ANR-18-CE02-0013-01; PL-G/KB).

Conflict of interest The authors declare that they have no conflicts of interest.

References

Allwood, A.C., Walter, M.R., Kamber, B.S., Marshall, C.P., and Burch, I.W. (2006) Stromatolite reef from the Early Archaean era of Australia. *Nature* **441**: 714-718.

Allwood, A.C., Grotzinger, J.P., Knoll, A.H., Burch, I.W., Anderson, M.S., Coleman, M.L., and Kanik, I. (2009) Controls on development and diversity of Early Archean stromatolites. *Proc Natl Acad Sci U S A* **106**: 9548-9555.

Altermann, W. (2004) Precambrian stromatolites: problems in definition, classification, morphology and stratigraphy. In *The Precambrian Earth: Tempos and Events*. Eriksson, P.G., Altermann, W., Nelson, D.R., Mueller, W.U., and Catuneanu, O. (eds). Amsterdam: Elsevier, pp. 564-574.

Armienta, M.A., Vilaclara, G., De la Cruz-Reyna, S., Ramos, S., Ceniceros, N., Cruz, O. et al. (2008) Water chemistry of lakes related to active and inactive Mexican volcanoes. *Journal of Volcanology and Geothermal Research* **178**: 249-258.

Arp, G., Reimer, A., and Reitner, J. (1999) Calcification in cyanobacterial biofilms of alkaline salt lakes. *Eur J Phycol* **34**: 393 - 403.

Arp, G., Reimer, A., and Reitner, J. (2003) Microbialite formation in seawater of increased alkalinity, Satonda crater lake, Indonesia. *Journal of Sedimentary Research* **73**: 105-127.

Awramik, S.M. (1990) Stromatolites. In *Paleobiology: A synthesis*. Briggs, D.E.G., and Crowther, P.R. (eds). Oxford: Blackwell Scientific Publications, pp. 336-341.

Awramik, S.M., and Riding, R. (1988) Role of algal eukaryotes in subtidal columnar stromatolite formation. *Proc Natl Acad Sci U S A* **85**: 1327-1329.

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2019) EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Syst Biol* **68**: 365-369.

Benzerara, K., Menguy, N., Lopez-Garcia, P., Yoon, T.H., Kazmierczak, J., Tyliszczak, T. et al. (2006) Nanoscale detection of organic signatures in carbonate microbialites. *Proc Natl Acad Sci U S A* **103**: 9440-9445.

Bower, S.M., Carnegie, R.B., Goh, B., Jones, S.R., Lowe, G.J., and Mak, M.W. (2004) Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. *J Eukaryot Microbiol* **51**: 325-332.

Bray, J.R., and Curtis, J.T. (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs* **27**: 325-349.

Breitbart, M., Hoare, A., Nitti, A., Siefert, J., Haynes, M., Dinsdale, E. et al. (2009) Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environ Microbiol* **11**: 16-34.

Bundeleva, I.A., Shirokova, L.S., Benezeth, P., Pokrovsky, O.S., Kompantseva, E.I., and Balor, S. (2012) Calcium carbonate precipitation by anoxygenic phototrophic bacteria. *Chemical Geology* **291**: 116-131.

Cam, N., Benzerara, K., Georgelin, T., Jaber, M., Lambert, J.F., Poinsot, M. et al. (2018) Cyanobacterial formation of intracellular Ca-carbonates in undersaturated solutions. *Geobiology* **16**: 49-61.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Casaburi, G., Duscher, A.A., Reid, R.P., and Foster, J.S. (2016) Characterization of the stromatolite microbiome from Little Darby Island, The Bahamas using predictive and whole shotgun metagenomic analysis. *Environ Microbiol* **18**: 1452-1469.

Centeno, C.M., Legendre, P., Beltran, Y., Alcantara-Hernandez, R.J., Lidstrom, U.E., Ashby, M.N., and Falcon, L.I. (2012) Microbialite genetic diversity and composition relate to environmental variables. *FEMS Microbiol Ecol* **82**: 724-735.

Chagas, A.A.P., Webb, G.E., Burne, R.V., and Southam, G. (2016) Modern lacustrine microbialites: Towards a synthesis of aqueous and carbonate geochemistry and mineralogy. *Earth-Science Reviews* **162**: 338-363.

Couradeau, E., Benzerara, K., Moreira, D., Gerard, E., Kazmierczak, J., Tavera, R., and Lopez-Garcia, P. (2011) Prokaryotic and eukaryotic community structure in field and cultured microbialites from the alkaline Lake Alchichica (Mexico). *PLoS ONE* **6**: e28767.

Couradeau, E., Benzerara, K., Gérard, E., Estève, I., Moreira, D., Tavera, R., and López-García, P. (2013) In situ microscale cyanobacterial calcification in modern microbialites. *Biogeosciences* **10**: 5255-5266.

Czech, L., Barbera, P., and Stamatakis, A. (2020) Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics*.

Decelle, J., Romac, S., Sasaki, E., Not, F., and Mahe, F. (2014) Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. *PLoS One* **9**: e104297.

Dhami, N.K., Reddy, M.S., and Mukherjee, A. (2013) Biomineralization of calcium carbonates and their engineered applications: a review. *Front Microbiol* **4**: 314.

Dravis, J.J. (1983) Hardened subtidal stromatolites, Bahamas. *Science* **219**: 385-386.

Dray, S., and Dufoour, A.B. (2007) The ade4 Package : Implementing the duality diagram for ecologists. *Journal of Statistical Software* **22-4**.

Dupraz, C., and Visscher, P.T. (2005) Microbial lithification in marine stromatolites and hypersaline mats. *Trends Microbiol* **13**: 429-438.

Dupraz, C., Reid, R.P., Braissant, O., Decho, A.W., Norman, R.S., and Visscher, P.T. (2009) Processes of carbonate precipitation in modern microbial mats. *Earth-Science Reviews* **96**: 141-162.

Farias, M.E., Contreras, M., Rasuk, M.C., Kurth, D., Flores, M.R., Poire, D.G. et al. (2014) Characterization of bacterial diversity associated with microbial mats, gypsum evaporites and carbonate microbialites in thalassic wetlands: Tebenquiche and La Brava, Salar de Atacama, Chile. *Extremophiles* **18**: 311-329.

Farias, M.E., Rascovan, N., Toneatti, D.M., Albarracin, V.H., Flores, M.R., Poire, D.G. et al. (2013) The discovery of stromatolites developing at 3570 m above sea level in a high-altitude volcanic lake Socompa, Argentinean Andes. *PLoS One* **8**: e53497.

Fedorchuk, N.D., Dornbos, S.Q., Corsetti, F.A., Isbell, J.L., Petryshyn, V.A., Bowles, J.A., and Wilmeth, D.T. (2016) Early non-marine life: Evaluating the biogenicity of Mesoproterozoic fluvial-lacustrine stromatolites. *Precambrian Research* **275**: 105-118.

Foster, J.S., and Green, S.J. (2011) Microbial diversity in modern stromatolites. In *Stromatolites: Interaction of microbes with sediments*. Tewari, V.C., and Seckbach, J. (eds): Springer Science & Business Media, pp. 383–405.

Foster, J.S., Babilonia, J., Parke-Suossari, E., and Reid, R.P. (2020) Chapter 4: Stromatolites, biosignatures and astrobiological implications. In *Astrobiology and Cuatro Ciénagas basin as analog of early Earth*. Souza, V., Segura, A., and Foster, J.S. (eds): Springer International Publishing, pp. 89-105.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150-3152.

Gallagher, K.L., Kading, T.J., Braissant, O., Dupraz, C., and Visscher, P.T. (2012) Inside the alkalinity engine: the role of electron donors in the organomineralization potential of sulfate-reducing bacteria. *Geobiology* **10**: 518-530.

Gerard, E., Menez, B., Couradeau, E., Moreira, D., Benzerara, K., Tavera, R., and Lopez-Garcia, P. (2013) Specific carbonate-microbe interactions in the modern microbialites of Lake Alchichica (Mexico). *ISME J* **7**: 1997-2009.

Gerard, E., De Goeyse, S., Hugoni, M., Agogue, H., Richard, L., Milesi, V. et al. (2018) Key role of Alphaproteobacteria and Cyanobacteria in the formation of stromatolites of Lake Dziani Dzaha (Mayotte, Western Indian Ocean). *Front Microbiol* **9**: 796.

Grotzinger, J.P. (1990) Geochemical model for Proterozoic stromatolite decline. *Amer J Sci* **290-A**: 80-103.

Guida, B.S., and Garcia-Pichel, F. (2016) Extreme cellular adaptations and cell differentiation required by a cyanobacterium for carbonate excavation. *Proceedings of the National Academy of Sciences of the United States of America* **113**: 5712-5717.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L. et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: D597-D604.

Gutierrez-Preciado, A., Saghai, A., Moreira, D., Zivanovic, Y., Deschamps, P., and Lopez-Garcia, P. (2018) Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nat Ecol Evol* **2**: 1700-1708.

Harris, J.K., Caporaso, J.G., Walker, J.J., Spear, J.R., Gold, N.J., Robertson, C.E. et al. (2013) Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. *ISME J* **7**: 50-60.

Havemann, S.A., and Foster, J.S. (2008) Comparative characterization of the microbial diversities of an artificial microbialite model and a natural stromatolite. *Appl Environ Microbiol* **74**: 7410-7421.

Himmler, T., Smrzka, D., Zwicker, J., Kasten, S., Shapiro, R.S., Bohrmann, G., and Peckmann, J. (2018) Stromatolites below the photic zone in the northern Arabian Sea formed by calcifying chemotrophic microbial mats. *Geology* **46**: 339-342.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J. et al. (2016) A new view of the tree of life. *Nat Microbiol* **1**: 16048.

Ionescu, D., Spitzer, S., Reimer, A., Schneider, D., Daniel, R., Reitner, J. et al. (2014) Calcium dynamics in microbialite-forming exopolymer-rich mats on the atoll of Kiritimati, Republic of Kiribati, Central Pacific. *Geobiology* **doi: 10.1111/gbi.12120. [Epub ahead of print]**.

Johnson, D.B., Beddows, P.A., Flynn, T.M., and Osburn, M.R. (2018) Microbial diversity and biomarker analysis of modern freshwater microbialites from Laguna Bacalar, Mexico. *Geobiology* **16**: 319-337.

karHernandez-Agreda, A., Gates, R.D., and Ainsworth, T.D. (2017) Defining the core microbiome in corals' microbial soup. *Trends Microbiol* **25**: 125-140.

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

Kazmierczak, J., Kempe, S., Kremer, B., López-García, P., Moreira, D., and Tavera, R. (2011) Hydrochemistry and microbialites of the alkaline crater lake Alchichica, Mexico. *Facies* **57**: 543-570.

Kempe, S., Kazmierczak, J., Landmann, G., Konuk, T., Reimer, A., and Lipp, A. (1991) Largest known microbialites discovered in Lake Van, Turkey. *Nature* **349**: 605-608.

Langille, M.G., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A. et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814-821.

Last, F.M., Last, W.M., and Halden, N.M. (2012) Modern and late Holocene dolomite formation: Manito Lake, Saskatchewan, Canada. *Sedimentary Geology* **281**: 222-237.

Laval, B., Cady, S.L., Pollack, J.C., McKay, C.P., Bird, J.S., Grotzinger, J.P. et al. (2000) Modern freshwater microbialite analogues for ancient dendritic reef structures. *Nature* **407**: 626-629.

Legendre, P., and Legendre, L. (1998) *Numerical ecology*. Amsterdam: Elsevier Science BV.

Lim, D.S.S., Laval, B.E., Slater, G., Antoniades, D., Forrest, A.L., Pike, W. et al. (2009) Limnology of Pavilion Lake, B. C., Canada - Characterization of a microbialite forming environment. *Fundamental and Applied Limnology* **173**: 329-351.

Logan, B.W. (1961) *Cryptozoon* and associated stromatolites from the Recent, Shark Bay, Western Australia. *J Geol* **69**: 517-533.

López-García, P., Kazmierczak, J., Benzerara, K., Kempe, S., Guyot, F., and Moreira, D. (2005) Bacterial diversity and carbonate precipitation in the giant microbialites from the highly alkaline Lake Van, Turkey. *Extremophiles* **9**: 263-274.

Louca, S., Jacques, S.M.S., Pires, A.P.F., Leal, J.S., Srivastava, D.S., Parfrey, L.W. et al. (2016) High taxonomic variability despite stable functional structure across microbial communities. *Nat Ecol Evol* **1**: 15.

Louyakis, A.S., Mobberley, J.M., Vitek, B.E., Visscher, P.T., Hagan, P.D., Reid, R.P. et al. (2017) A study of the microbial spatial heterogeneity of Bahamian thrombolites using molecular, biochemical, and stable isotope analyses. *Astrobiology* **17**: 413-430.

Macreadie, P.I., Anton, A., Raven, J.A., Beaumont, N., Connolly, R.M., Friess, D.A. et al. (2019) The future of Blue Carbon science. *Nat Commun* **10**: 3998.

Magoc, T., and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957-2963.

Mahe, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**: e1420.

Mancilla Villa, O.R., Bautista Olivas, A.L., Ortega Escobar, H.M., Sánchez Bernal, E.I., Can Chulim, Á., Guevara Gutiérrez, R.D., and Ortega Mikolaev, Y.M. (2014) Hidrogeoquímica de salinas Zapotitlán y los lagos-cráter Alchichica y Atexcac, Puebla and Atexcac, Puebla. *Idesia (Arica)* **32**: 55-69.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJournal* **17**: 10-12.

Martiny, J.B., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**: aac9323.

McCutcheon, J., Power, I.M., Harrison, A.L., Dipple, G.M., and Southam, G. (2014) A greenhouse-scale photosynthetic microbial bioreactor for carbon sequestration in magnesium carbonate minerals. *Environ Sci Technol* **48**: 9142-9151.

Michaelis, W., Seifert, R., Nauhaus, K., Treude, T., Thiel, V., Blumenberg, M. et al. (2002) Microbial reefs in the Black Sea fueled by anaerobic oxidation of methane. *Science* **297**: 1013-1015.

Mobberley, J.M., Ortega, M.C., and Foster, J.S. (2012) Comparative microbial diversity analyses of modern marine thrombolitic mats by barcoded pyrosequencing. *Environ Microbiol* **14**: 82-100.

Mobberley, J.M., Khodadad, C.L., Visscher, P.T., Reid, R.P., Hagan, P., and Foster, J.S. (2015) Inner workings of thrombolites: spatial gradients of metabolic activity as revealed by metatranscriptome profiling. *Sci Rep* **5**: 12601.

Morrissey, E.M., Mau, R.L., Hayer, M., Liu, X.A., Schwartz, E., Dijkstra, P. et al. (2019) Evolutionary history constrains microbial traits across environmental variation. *Nat Ecol Evol* **3**: 1064-1069.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.

Obst, M., Dynes, J.J., Lawrence, J.R., Swerhone, G.D.W., Benzerara, K., Karunakaran, C. et al. (2009) Precipitation of amorphous CaCO3 (aragonite-like) by cyanobacteria: A STXM study of the influence of EPS on the nucleation process. *Geochimica Et Cosmochimica Acta* **73**: 4180-4198.

Oksanen, J., Blanchet, G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L. et al. (2011) Vegan: Community Ecology Package. R package version 1.17-9. In. http://CRAN.R-project.org/package=vegan (ed): http://CRAN.R-project.org/package=vegan.

Pace, A., Bourillot, R., Bouton, A., Vennin, E., Galaup, S., Bundeleva, I. et al. (2016) Microbial and diagenetic steps leading to the mineralisation of Great Salt Lake microbialites. *Scientific Reports* **6**.

Papineau, D., Walker, J.J., Mojzsis, S.J., and Pace, N.R. (2005) Composition and structure of microbial communities from stromatolites of Hamelin Pool in Shark Bay, Western Australia. *Appl Environ Microbiol* **71**: 4822-4832.

Power, I.M., Wilson, S.A., Small, D.P., Dipple, G.M., Wan, W.K., and Southam, G. (2011) Microbially mediated mineral carbonation: roles of phototrophy and heterotrophy. *Environmental Science & Technology* **45**: 9061-9068.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-596.

R Development Core Team (2017) R: A language and environment for statistical computing. In. http://www.r-project.org (ed). Vienna, Austria: R Foundation for Statistical Computing.

Reid, R.P., and Browne, K.M. (1991) Intertidal stromatolites in a fringing Holocene reef complex in the Bahamas. *Geology* **19**: 15-18.

Reid, R.P., Visscher, P.T., Decho, A.W., Stolz, J.F., Bebout, B.M., Dupraz, C. et al. (2000) The role of microbes in accretion, lamination and early lithification of modern marine stromatolites. *Nature* **406**: 989-992.

Reimer, A., Landmann, G., and Kempe, S. (2009) Lake Van, Eastern Anatolia, Hydrochemistry and History. *Aquatic Geochemistry* **15**: 195-222.

Riding, R. (2000) Microbial carbonates: the geological record of calcified bacterial-algal mats and biofilms. *Sedimentology* **47**: 179-214.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahe, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Russell, J.A., Brady, A.L., Cardman, Z., Slater, G.F., Lim, D.S., and Biddle, J.F. (2014) Prokaryote populations of extant microbialites along a depth gradient in Pavilion Lake, British Columbia, Canada. *Geobiology* **12**: 250-264.

Ruvindy, R., White, R.A., 3rd, Neilan, B.A., and Burns, B.P. (2016) Unravelling core microbial metabolisms in the hypersaline microbial mats of Shark Bay using high-throughput metagenomics. *ISME J* **10**: 183-196.

Saghaï, A., Zivanovic, Y., Moreira, D., Benzerara, K., Bertolino, P., Ragon, M. et al. (2016) Comparative metagenomics unveils functions and genome features of microbialite-associated communities along a depth gradient. *Environ Microbiol* **18**: gut4990-5004.

Saghaï, A., Zivanovic, Y., Zeyen, N., Moreira, D., Benzerara, K., Deschamps, P. et al. (2015) Metagenome-based diversity analyses suggest a significant contribution of non-cyanobacterial lineages to carbonate precipitation in modern microbialites. *Front Microbiol* **6**: 797.

Saint Martin, J.P., and Saint Martin, S. (2015) Discovery of calcareous microbialites in coastal ponds of Western Sardinia, Italy. *Geo-Eco-Marina* **21**: 35-53.

Santos, F., Pena, A., Nogales, B., Soria-Soria, E., Del Cura, M.A., Gonzalez-Martin, J.A., and Anton, J. (2010) Bacterial diversity in dry modern freshwater stromatolites from Ruidera Pools Natural Park, Spain. *Syst Appl Microbiol* **33**: 209-221.

Seifan, M., and Berenjian, A. (2019) Microbially induced calcium carbonate precipitation: a widespread phenomenon in the biological world. *Appl Microbiol Biotechnol* **103**: 4693-4708.

Souza-Egipsy, V., Wierzchos, J., Ascaso, C., and Nealson, K.H. (2005) Mg-silica precipitation in fossilization mechanisms of sand tufa endolithic microbial community, Mono Lake (California). *Chemical Geology* **217**: 77-87.

Suosaari, E.P., Reid, R.P., Playford, P.E., Foster, J.S., Stolz, J.F., Casaburi, G. et al. (2016) New multi-scale perspectives on the stromatolites of Shark Bay, Western Australia. *Sci Rep* **6**: 20557.

Tice, M.M., and Lowe, D.R. (2004) Photosynthetic microbial mats in the 3,416-Myr-old ocean. *Nature* **431**: 549-552.

Valdespino-Castillo, P.M., Hu, P., Merino-Ibarra, M., Lopez-Gomez, L.M., Cerqueda-Garcia, D., Gonzalez-De Zayas, R. et al. (2018) Exploring biogeochemistry and microbial diversity of extant microbialites in Mexico and Cuba. *Front Microbiol* **9**: 510.

Visscher, P.T., Reid, P.R., and Bebout, B.M. (2000) Microscale observations of sulfate reduction: Correlation of microbial activity with lithified micritic laminae in modern marine stromatolites. *Geology* **28**: 919–922.

Wang, C., Zhang, T., Wang, Y., Katz, L.A., Gao, F., and Song, W. (2017) Disentangling sources of variation in SSU rDNA sequences from single cell analyses of ciliates: impact of copy number variation and experimental error. *Proc Roy Soc B* **284**: 20170425.

Warden, J.G., Casaburi, G., Omelon, C.R., Bennett, P.C., Breecker, D.O., and Foster, J.S. (2016) Characterization of microbial mat microbiomes in the modern thrombolite ecosystem of Lake Clifton, Western Australia using shotgun metagenomics. *Front Microbiol* **7**: 1064.

Weisse, T. (2002) The significance of inter- and intraspecific variation in bacterivorous and herbivorous protists. *Antonie Van Leeuwenhoek* **81**: 327-341.

White, R.A., 3rd, Chan, A.M., Gavelis, G.S., Leander, B.S., Brady, A.L., Slater, G.F. et al. (2016) Metagenomic analysis suggests modern freshwater microbialites harbor a distinct core microbial community. *Front Microbiol* **6**: 1531.

Wong, H.L., Ahmed-Cox, A., and Burns, B.P. (2016) Molecular ecology of hypersaline microbial mats: current insights and new directions. *Microorganisms* **4**.

Zeyen, N., Daval, D., Lopez-Garcia, P., Moreira, D., Gaillardet, J., and Benzerara, K. (2017) Geochemical conditions allowing the formation of modern lacustrine microbialites. *Procedia Earth and Planetary Science* **17**: 380-383.

Zeyen, N., Benzerara, K., Li, J., Groleau, A., Balan, E., Robert, J.-L. et al. (2015) Formation of low-T hydrated silicates in modern microbialites from Mexico and implications for microbial fossilization. *Frontiers in Earth Science* **3**: 64.

Zeyen, N., Benzerara, K., Menguy, N., Brest, J., Templeton, A.S., Webb, S.M. et al. (2019) Fe-bearing phases in modern lacustrine microbialites from Mexico. *Geochim Cosmochim Acta* **253**: 201-230.

Figure legends

Fig.1. Mexican lakes sampled for this study. **A**, location of the different lakes on the Trans-Mexican volcanic belt (pink area). **B**, Mexican lakes displaying microbialites (green-shaded area) as a function of alkalinity and conductivity. All lakes except Zirahuen and Patzcuaro are crater (*maar*) lakes.

Fig.2. Histograms showing the phylogenetic diversity and relative proportion of 16S and 18S rRNA genes amplified from microbialite samples collected from Mexican lakes along an alkalinity gradient. **A**, relative abundance of prokaryotic sequences. **B**, relative abundance of prokaryotic operational taxonomic units (OTUs). **C**, relative abundance of eukaryotic sequences. **D**, relative abundance of eukaryotic OTUs. Detailed histograms of the categories 'Other Bacteria', Archaea and 'Other eukaryotes' are provided in, respectively, Supporting Information Figs. S3A, 3B and S4. Sample descriptions are provided in Table 1.

Fig.3. Comparison of microbialite samples according to their associated prokaryotic and eukaryotic communities based on Bray Curtis distances. **A**, Non-metric multidimensional scaling (NMDS) biplot. A variant of this figure including a projection of the most influential parameters is shown in Fig. S7. **B**, Hierarchical clustering based on 16S and 18S rRNA gene-based community composition. The green-shaded area indicates closely grouping samples from Alchichica and Atexcac SE samples.

Fig.4. Canonical-correlation analysis biplot showing the studied microbialite samples as a function of pH, conductivity (Cond), alkalinity [HCO_3^-], [Ca^{2+}] and the ratio [Mg^{2+}]/[Ca^{2+}]. CCAs

showing additional abiotic parameters are shown in Supporting Information Fig.S9. Microbialites from the different lakes are color-coded as indicated.

875

Fig.5. Taxon-based metabolic profiling of microbialite-associated prokaryotic communities across different Mexican crater lakes. **A**, phylogeny-based relative abundance of different metabolic pathways potentially influencing microbialite formation inferred from the number of 16S rRNA genes reads for specific taxa known to carry out a particular metabolism. Values
880 correspond to average proportions from replicate samples for each lake. Metabolic categories to the left of 'Uncertain' are potentially carbonatogenic, those on the right, favor carbonate dissolution. **B**, NMDS biplot showing the distribution of the different samples according to their inferred metabolic pattern. Anox., anoxygenic; GNSB, green non-sulfur bacteria; PNSB, purple non-sulfur bacteria; PSB, purple sulfur bacteria.

885

Fig.6. Prokaryotic and eukaryotic core communities shared by Trans-Mexican volcanic belt lake microbialites. **A**, UpSet plot showing prokaryotic OTUs shared by the different lake microbialites. The number, phylogenetic affiliation and relative abundance of OTUs within the core shared by all the lakes or all the lakes but one (light grey dot) are provided in the upper
890 histogram. The histogram on the right shows the relative proportion (sequence reads) of the prokaryotic core community in the total prokaryotic community of each lake microbialite. **B**, UpSet plot as in (A) showing eukaryotic OTUs shared by the different lake microbialites.

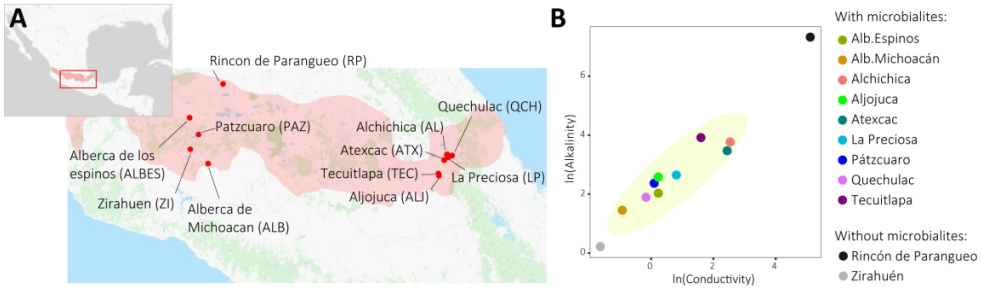


Figure 1. Iniesto et al.

185x96mm (300 x 300 DPI)

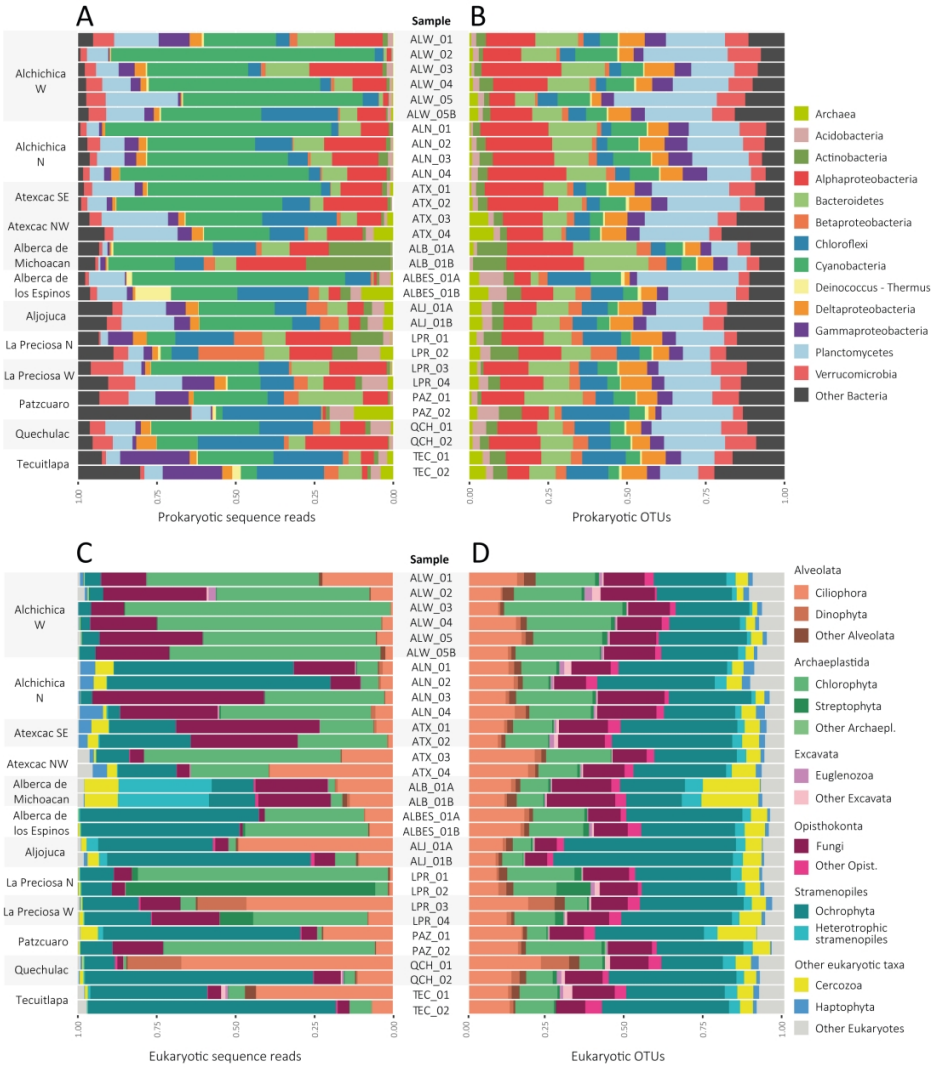


Figure 2. Iniesto et al.

206x260mm (300 x 300 DPI)

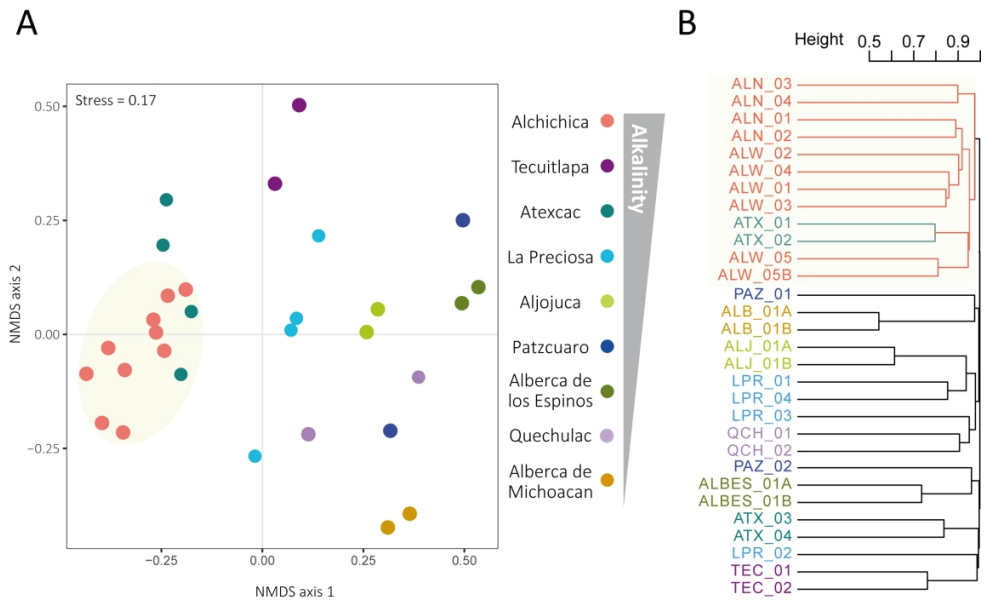


Figure 3. Iniesto et al.

194x162mm (300 x 300 DPI)

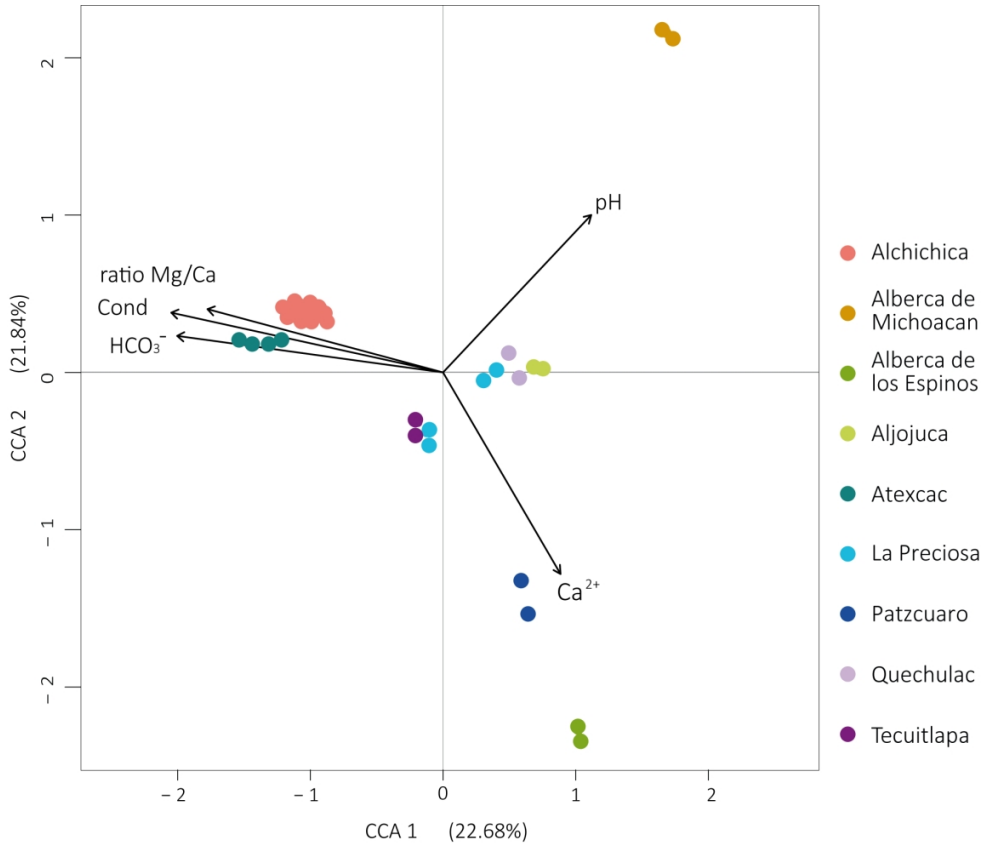


Figure 4. Iniesto et al.

169x166mm (300 x 300 DPI)

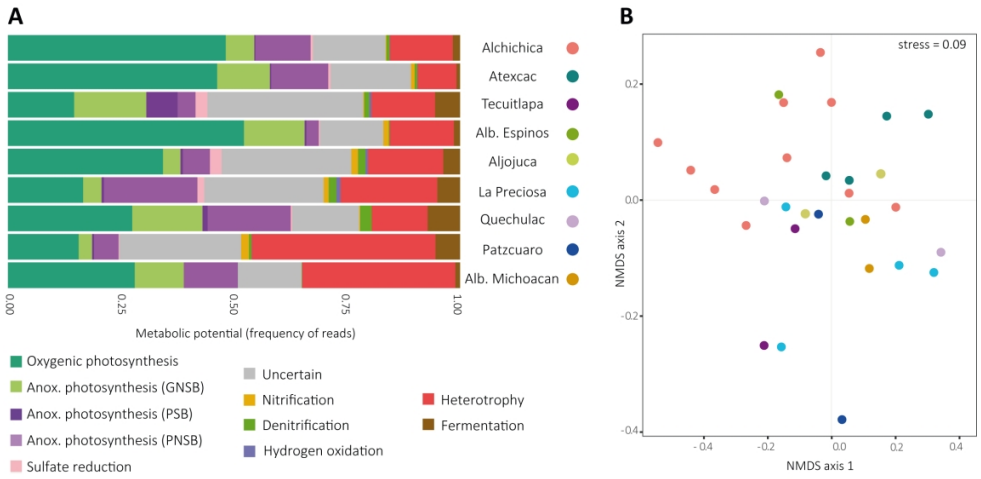


Figure 5. Iniesto et al.

202x149mm (300 x 300 DPI)

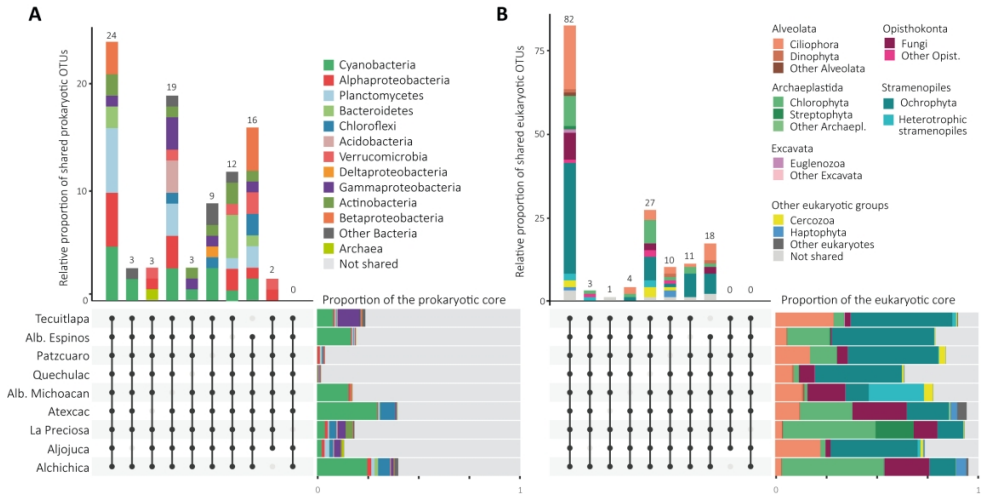


Figure 6. Iniesto et al.

206x139mm (300 x 300 DPI)

Table 1. Sample information and selected sequence statistics for microbialite samples collected at 9 Mexican crater lakes. OTUs were defined usign SWARM. For additional sequence statistics and diversity indices, see Supporting Information Table S2. H, hydromagnesite; A, aragonite; MgSi, talc; MgC, magnesian calcite.

Sample name	Lake	Coordinates / Location	Collection date	Description/Dominant mineral facies*	Total retained high-quality reads 16S	Total retained high-quality reads 18S	Bacterial OTUs	Archaeal OTUs	Eukaryotic OTUs
ALW_01	Alchichica (West)	19°24.299' N 97°24.389'W	5/6/2014	Microbialite sampled at 0.4m depth; H-A	14899	42299	1521	7	341
ALW_02	Alchichica (W)		5/6/2014	Microbialite sampled at 1.5m depth; H-A	26599	43087	737	8	432
ALW_03	Alchichica (W)		5/6/2014	Microbialite sampled at 3m depth; A-H-MgSi	45340	33661	1574	9	392
ALW_04	Alchichica (W)		5/6/2014	Microbialite in column; H-A	20681	49676	1193	13	478
ALW_05	Alchichica (W)		5/6/2014	Microbialite in column; H-A	28623	31176	1567	43	432
ALW_05B	Alchichica (W)		5/6/2014	Microbialite in column; H-A	81811	54069	2537	74	503
ALN_01	Alchichica (Nord)	19°25.147' N 97°4.162' W	5/6/2014	Surface microbialite; H-A	17602	34203	632	2	376
ALN_02	Alchichica (N)		5/6/2014	Surface microbialite; H-A	9158	51100	750	5	476
ALN_03	Alchichica (N)		5/6/2014	Microbialite partially exposed to light; H-A	2795	19654	498	4	421
ALN_04	Alchichica (N)		5/6/2014	Microbialite totally exposede to light; H-A	4415	33340	450	2	342
ALN_F_01A	Alchichica (N)		5/6/2014	Colonization experiment - nascent microbialite on Nylon mesh	155782	47832	1971	14	450
ALN_F_01B	Alchichica (N)		5/6/2014	Colonization experiment - nascent microbialite on Nylon mesh	13722	32160	864	1	460
ATX_01	Atexcac (NW)	19°20'6.92"N 97°27'12.31"W	5/8/2014	Surface microbialite; A-MgSi	111202	60642	1752	10	613
ATX_02	Atexcac (NW)		5/8/2014	0.5m depth microbialite; A-MgSi	70265	49061	1137	11	547
ATX_03	Atexcac (SE)	19°19'53.57"N 97°27'3.19"W	5/8/2014	Surface microbialite; A-MgSi	39593	17408	1957	124	334
ATX_04	Atexcac (SE)		5/8/2014	Surface microbialite; A-MgSi	60827	41952	2735	225	496
ALB_01A	Alberca de Michoacán	19°12'41.15"N 101°27'25.18"W	5/15/2014	Thin calcifying biofilm on basalt rocks	55141	62453	1807	25	694
ALB_01B	Alberca de Michoacán		5/15/2014	Thin calcifying biofilm on basalt rocks	60550	28245	1805	1	549
ALBES_01A	Alberca de los Espinos	19°54'27.39"N 101°46'1.14"W	5/15/2014	Surface microbialite; MgC	17624	17905	1079	34	338
ALBES_01B	Alberca de los Espinos		5/15/2014	Surface microbialite; MgC	60695	28710	2183	138	334
ALJ_01A	Aljojuca	19° 5'39.32"N 97°31'56.15"W	5/9/2014	Surface microbialite; MgC	33099	39631	2461	98	1007
ALJ_01B	Aljojuca		5/9/2014	Surface microbialite; MgC	79206	57226	3298	137	1143
LPR_01	La Preciosa (N)	19°22'31.90"N 97°23'19.11"W	1/12/2012	Surface microbialite; A-MgSi	142641	77804	2781	64	574
LPR_02	La Preciosa (N)		1/12/2012	Carbonate-like crust; A-MgSi	23357	36469	1211	43	321
LPR_03	La Preciosa (W)	19°22'20.70"N 97°22'57.55"W	5/7/2014	Microbialite sampled at 1.50m depth; A-MgSi	22499	2978	1303	32	159
LPR_04	La Preciosa (W)		5/7/2014	Surface microbialite; A-MgSi	96952	96341	3704	109	899
PAZ_01	Patzcuaro	19°37'06.2"N 101°38'29.6"W	5/12/2014	Surface microbialite; MgC-A-MgSi	35756	33213	1519	10	493
PAZ_02	Patzcuaro		5/12/2014	Carbonate-like crust; MgC-A-MgSi	48152	7989	1305	33	201
QCH_01	Quechulac	19°22'30.76"N 97°21'17.68"W	1/13/2012	Surface microbialite; A	19660	2250	1631	14	127
QCH_02	Quechulac		1/13/2012	Surface microbialite with Nostoc-like colonies; A	144271	60922	2202	20	631
TEC_01	Tecuitlapa	19°07'30.9"N 97°32'39.0"W	5/9/2014	Surface microbialite; Ca	25314	12979	1044	45	281
TEC_02	Tecuitlapa		5/9/2014	Fragment of surface microbialite from the likely noxic zone; Ca	20034	8408	871	48	177

* Dominant mineral facies from Zeyen et al (2017)

APPENDIX

D

FOURTH APPENDIX

Environmental drivers of plankton protist communities along latitudinal and vertical gradients in the oldest and deepest freshwater lake

Gwendoline M. David[1], David Moreira[1], Guillaume Reboul[1], Nataliia V. Annenkova[2], Luis J. Galindo[1], Paola Bertolino[1], Ana I. López-Archilla[3], Ludwig Jardillier[1] and Purificación López-García[1]

[1] Ecologie Systématique Evolution, Centre National de la Recherche Scientifique - CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France

[2] Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

[3] Departamento de Ecología, Universidad Autónoma de Madrid, Madrid, Spain

For correspondence: puri.lopez@u-psud.fr

Running title: Drivers of protist communities in Lake Baikal

Originality and Significance Statement

Lake Baikal is the oldest, deepest and most voluminous freshwater lake on Earth, offering a unique opportunity to test the effects of horizontal versus vertical gradients on microbial community structure. Using a metabarcoding approach, we studied planktonic microbial eukaryotes from Baikal water columns (5 up to 1,400 m depth) across a North-South latitudinal gradient (~600 km), including coastal and pelagic areas. Our results show that depth has a strong effect on protist community assemblage, but not latitude (minor effect) or coastal vs. open water sites (no effect). Co-occurrence analyses also point to specific biotic interactions as drivers of community structure. This comprehensive survey constitutes a useful reference for monitoring active climate change effects in this ancient lake.

Summary

Identifying which abiotic and biotic factors determine microbial community assembly is crucial to understand ecological processes and predict how communities will respond to environmental change. While global surveys aim at addressing this question in the world's oceans, equivalent studies in large freshwater systems are virtually lacking. Being the oldest, deepest and most voluminous freshwater lake on Earth, Lake Baikal offers a unique opportunity to test the effect of horizontal versus vertical gradients in community structure. Here, we characterized the structure of planktonic microbial eukaryotic communities (0.2-30 µm cell size) along a North-South latitudinal gradient (~600 km) from samples collected in coastal and pelagic waters and from surface to the deepest zones (5-1400 m) using an 18S rRNA gene metabarcoding approach. Our results show complex and diverse protist communities dominated by alveolates (ciliates and dinoflagellates), ochrophytes and holomycotan lineages, with cryptophytes, haptophytes, katablepharids and telonemids in moderate abundance and many low-frequency lineages, including several typical marine members, such as diplonemids, syndinians and radiolarians. Depth had a strong significant effect on protist community stratification. By contrast, the effect of the latitudinal gradient was marginal and no significant difference was observed between coastal and surface open water communities. Co-occurrence network analyses showed that epipelagic communities are much more interconnected than meso- and bathypelagic communities and suggest specific biotic interactions between autotrophic, heterotrophic and parasitic lineages that influence protist community structure. Since climate change is rapidly affecting Siberia and Lake Baikal, our comprehensive protist survey constitutes a useful reference to monitor ongoing community shifts.

Keywords: Lake Baikal; protist; 18S rRNA gene metabarcoding; marine-freshwater transition; light stratification; network analysis

Introduction

Of all ecosystems, freshwater reservoirs are the most dynamic and concentrate a high biodiversity (Rolls et al., 2018). Freshwater ecosystems are particularly vulnerable to climate change owing to a higher exposure and sensitivity to increasing temperature and other altered conditions, limited dispersal across these fragmented habitats and little-known, but likely modest, resilience potential (Woodward et al., 2010; Markovic et al., 2017). Since microorganisms are crucial in biogeochemical cycles, the impact of climate change will strongly depend on how they will respond to environmental challenge (Cavicchioli et al., 2019). Permafrost-covered areas in the Arctic region (Schuur et al., 2015) and forest-steppe ecotones in Siberia are among the most heavily impacted regions by global warming (Mackay et al., 2017). This includes Lake Baikal, in southern Siberia, which is the oldest (ca. 30 Myr), deepest, and most capacious freshwater lake on Earth (Müller et al., 2001). Lake Baikal is rapidly changing, as can be told from trends in hydrological and hydrochemical processes (Moore et al., 2009; Shimaraev and Domysheva, 2012). The lake sediments represent a continuous record of past climate for over 12 million years (Kashiwaya et al., 2001; Prokopenko et al., 2002) such that Lake Baikal is a unique model to understand and predict microbial community change and how this is linked to carbon cycling and hydrological processes.

A mandatory prerequisite for such a task is to have comprehensive information about the existing microbial community structure. However, if the broad biodiversity of Lake Baikal metazoans, including many endemisms (1455 out of 2595 species described), has been amply documented in the past two centuries, that of microbial life is highly fragmentary. One of the reasons relates to the large dimensions of the lake, which is around 640 km long, attains a depth of ca. 1650 m and contains around 20% of the Earth's unfrozen freshwater (Sherstyankin et al., 2006; UNDP-GEF, 2015). This, together with its geographical location and its association to a rifting zone make Lake Baikal unique and listed as UNESCO World Heritage Site (UNDP-GEF, 2015). The lake is divided in three major basins (Northern, Central, Southern) by, respectively, the Academician Ridge and the Selenga river delta (Mats and Perepelova, 2011). Its surface freezes in winter for several months, favoring coastal downwelling and deep-water oxygenation (Schmid et al., 2008; Moore et al., 2009). As a result, Lake Baikal ultra-oligotrophic waters are globally cold (~4°C) and oxygen-rich down to the bottom (Schmid et al., 2008; Moore et al., 2009; Shimaraev and Domysheva, 2012; Troitskaya et al., 2015). Baikal also uniquely hosts methane hydrates, which are stabilized by the low temperatures and high pressures (De Batist et al., 2002; Granin et al., 2019). All these features make Lake Baikal akin a freshwater sea.

Microbial diversity in Lake Baikal plankton was first studied by classical observation and cultural approaches (Maksimova and Maksimov, 1972; Maksimov et al., 2002; Bel'kova et al., 2003) before molecular tools started to be applied at the beginning of the century (Glöckner et al., 2000) and

expanded more recently with the generalization of high-throughput sequencing. Several 16S rRNA gene-based metabarcoding studies have targeted pelagic bacteria diversity (Kurilkina et al., 2016; Belikov et al., 2019; Zakharenko et al., 2019; Wilburn et al., 2020) and, more recently, metagenomic analyses have been used to characterize planktonic prokaryotic communities from sub-ice (Cabello-Yeves et al., 2018) and deep waters (Cabello-Yeves et al., 2020), virus-bacteria assemblages in coastal waters (Butina et al., 2019) or viruses from the pelagic zone (Potapov et al., 2019). Microbial eukaryotes have only been partially studied by 18S rRNA gene metabarcoding. Several of these studies focused on phytoplankton, either on specific groups, such as diatoms (Zakharova et al., 2013) or dinoflagellates (Annenkova et al., 2011), or on whole communities, from winter sub-ice waters (Bashenkhaeva et al., 2015) to spring blooms (Mikhailov et al., 2015; Mikhailov et al., 2019b; Mikhailov et al., 2019a). Remarkably few studies have aimed at charactering the diversity of all microbial eukaryotes, especially in a comparative manner. Yi et al. analyzed protist diversity by 454 sequencing of 18S rRNA gene V9-region amplicons along the Southern basin water column (52-1450 m) (Yi et al., 2017). More recently, Annenkova et al. determined the community structure of small protists (0.45-8 µm cell-size fraction) from surface waters (1-15-50 m) across the lake via 18S rRNA gene V4-region metabarcoding and suggested that some clades within known protist groups might be endemic (Annenkova et al., 2020). Nonetheless, we still lack a comprehensive view about how microbial eukaryotes distribute in the lake plankton, across basins and throughout the complete water column and, crucially, which are the most influential parameters determining community structure.

In this work, we carry out a wide-ranging comparative study of Lake Baikal planktonic protist communities in the 0.2-30 µm cell-size range using a 18S rRNA gene metabarcoding approach to study distribution patterns and to test whether depth, latitude or the coastal versus pelagic location determine community structure. With this aim, we analyze 65 samples from 17 sites across a ~600 km latitudinal North-South transect along the three lake basins and from littoral shallow areas to deep water columns covering the epi-, meso- and bathypelagic region. Our results show complex and diverse protist communities that are mostly structured by depth and that include several typical marine lineages in low abundance. Network analyses show that epipelagic communities are much more interconnected than meso- and bathypelagic communities, suggesting potential specific biotic interactions between autotrophs, heterotrophs and parasites.

Experimental procedures

Sample collection

Lake Baikal water samples were collected at different depths from seventeen sites distributed along a North-South transect during a French-Russian research cruise in the summer of 2017. Sites were chosen

to cover littoral (8) and open water (9) samples, including the deepest zones in the three major basins of the lake. In total, 65 water samples were collected from depths ranging from 5 to 1400 m; deep samples were collected far from the bottom to avoid sediment disturbance (Supplementary Table 1). Samples were collected with Niskin bottles (5 l for epipelagic waters, 10 l for meso- and bathypelagic waters). The physicochemical parameters of lake waters were measured with a multiparameter probe Multi 350i (WTW, Weilheim, Germany). The water was sequentially filtered onboard immediately after collection through 30-µm and 0.22-µm pore-size Nucleopore filters (Whatman, Maidstone, UK) and 0.2 µm pore-size Cell-Trap units (MEM-TEQ Ventures Ltd, Wigan, UK). Volumes of water samples filtered through Cell-Traps were smaller (samples indicated with an asterisk in Fig.1). The recovered biomass and biomass-containing filters were fixed in absolute ethanol and stored at -20°C until processed.

DNA purification, 18S rRNA gene-fragment amplification and sequencing

DNA was purified using the Power Soil™ DNA purification kit (Qiagen, Hilden, Germany). 18S rRNA gene fragments (~530 bp) encompassing the V4 region were PCR-amplified using EK-565F-NGS (5'-GCAGTTAAAAAGCTCGTAGT-3') and UNonMet (5'-TTTAAGTTTCAGCCTTGCG-3'), the latter biased against metazoans (Bower et al., 2004). Primers were tagged with specific 10-bp molecular identifiers (MIDs) for multiplexed sequencing. To minimize PCR-associated biases, five PCR reaction products per sample were pooled. PCR reactions were conducted in 25-µl reaction mixtures containing 0.5-3 µl of eluted DNA, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, 0.3 µM primers and 0.5 U Platinum Taq DNA Polymerase (Invitrogen, Carlsbad, CA) for 35 cycles (94°C for 30 s, 55-58°C for 30-45 s, 72°C for 90 s) preceded by 2 min denaturation at 94°C and followed by 5 min extension at 72°C. Pooled amplicons were purified using QIAquick PCR purification kit (Qiagen, Hilden,Germany). Amplicons were sequenced using paired-end (2x300 bp) Illumina MiSeq (Eurofins Genomics, Ebersberg, Germany). Sequences have been deposited in GenBank under the BioProject number PRJNA657482 (BioSamples SAMN15830589 to SAMN15830657).

Sequence and phylogenetic analyses

We used an in-house bioinformatic pipeline to process raw sequences. Paired-end reads were merged with FLASH (Magoc and Salzberg, 2011) under strict criteria and assigned to specific samples based on their MIDs. MID and primer sequences were trimmed using CUTADAPT (Martin, 2011). Cleaned merged reads were next dereplicated to unique sequences using VSEARCH (Rognes et al., 2016), which was also used to detect and eliminate potential chimeras. Non-chimeric sequences from all samples were pooled together to define operational taxonomic units (OTUs) at a conservative threshold of 95% identity for 18S rRNA genes using CD-HIT-EST (Fu et al., 2012) and SWARM (Mahe et al., 2015). Singletons were

excluded from subsequent analyses. OTUs were assigned to taxa based on their similarity with a local 18S rRNA database build from SILVA v128 (Quast et al., 2013) and PR2 v4.5 (Guillou et al., 2013). OTUs less than 80% identical to their best environmental hit were blasted against the GenBank *nr* database (https://ncbi.nlm.nih.gov/) and assigned manually by phylogenetic placement analyses. Briefly, the closest hits to our OTUs in SILVA and PR2 were aligned with full 18S rDNA reference sequences covering the eukaryotic diversity using MAFFT (Katoh and Standley, 2013). After removal of uninformative sites with trimAl (Capella-Gutierrez et al., 2009), we built a tree with full reference sequences with IQ-tree (Nguyen et al., 2015) under a GTR-+G+I sequence evolution model. OTU sequences were aligned to the reference alignment and then placed in the reference phylogenetic tree using EPA-ng (Barbera et al., 2019). OTUs with no reliable affiliation were maintained as 'Unclassified'. Maximum likelihood phylogenetic trees of diplonemid and radiolarian OTUs were reconstructed from specific MAFFT alignments including their closest blast hits and reference sequences with PhyML (Guindon and Gascuel, 2003) applying a GTR+G+I (4 categories) model of sequence evolution. Bootstrap values were obtained from 100 replicates.

Statistical analyses

We generated a table of eukaryotic OTU read abundance in the different samples of Lake Baikal for diversity and statistical analyses (Supplementary Tables 1-2). To avoid biases due to differences in absolute numbers of reads per sample, we rarefied our sequences to the second smaller number of reads (9771 in BK16.500m). BK28.100m was excluded from this process due to its lower number of reads. Statistical analyses were conducted on these data with R (R Development Core Team, 2017). Richness and diversity indices were calculated using the vegan package (Oksanen et al., 2011). Evenness was calculated according to Pielou (Pielou, 1966). To see if these indices were significantly different between sampling depths and basins, we performed Wilcoxon tests between the groups distributions using R. Likewise, to test the effect of sampling point, basin and depth class on protist community composition across samples, we conducted permutational multivariate analyses of variance (PERMANOVA) based on Wisconsin-standardized Bray-Curtis dissimilarities, using the adonis function of the vegan package. Across-sample community composition differences were visualized using non-metric multidimensional scaling (NMDS) analysis, also on Wisconsin-standardized Bray-Curtis dissimilarities. To connect communities according to specific origin we drew ellipses with the ade4 package (Dray and Dufoour, 2007). To test the significance of groups revealed by NMDS, we applied analysis of similarity (ANOSIM) tests with 999 permutations. Principal component analysis (PCA) of abiotic parameters based on centered and scaled data was performed with FactoMineR (Lê et al., 2008).

Network analysis

We built co-occurrence networks for each depth category (epipelagic, mesopelagic, bathypelagic using a multivariation Poisson lognormal model with the PLNmodels R-package (Chiquet et al., 2018) in order to account for depth-class differences between samples and potential additional covariables (specifically the sampling basin). We retained for the analysis OTUs present in more than 20% of samples and abundances higher than 0.01%. For model selection, we used Bayesian information criteria with a 50-size grid of penalties. Networks were visualized with the ggnet R-package (Chiquet et al., 2018). To further analyze network structure, we carried out a block model analysis using a stochastic block model approach on the binary co-occurrence network using the blockmodel R-package (Leger, 2016), which synthetizes the overall network structure by gathering nodes in groups with similar modes of interactions. Network properties were calculated using the igraph R-package (Csardi and Nepusz, 2006). Properties included the number of positive and negative edges, the total number of nodes and number of connected nodes. Network mean degrees correspond to the average number of established edges. The average path length indicates the mean number of edges necessary to link a given node randomly to another. Network complexity was estimated using two indicators: connectance and clustering coefficient. The connectance was calculated as $c = \frac{2E}{N \times (N-1)}$, where E is the number of edges and N the number of nodes (Barrat et al., 2008). Connectance is 1 when all possible links are established. The clustering coefficient is the probability that two nodes having a similar neighbor are connected to each other (Delmas et al., 2019). It varies between 0 and 1; low values indicated poor connectivity.

Results and discussion

Abiotic variables across sampling sites

We collected Lake Baikal water samples along the Northern, Central and Southern basins from the same established depths in the water column (except for the deepest sample, which was collected close to the bottom but at sufficient distance –minimum 45 m– to avoid sediment influence) (Fig.1A; Supplementary Table 1). Samples from coastal areas were always collected at 5 m depth in the water column. The measured physicochemical parameters were remarkably stable across sites and depths. Temperature ranged from 3.6 to 15.3°C, but was globally low (average 5.7°C; only five surface samples exceeded 9.5°C), and significantly higher in epipelagic samples (Supplementary Fig.1). pH ranged from 7.45 to 8.47. Salinity was extremely low (always 0.0 PSU) as, accordingly, conductivity and total dissolved solids (TDS). Dissolved oxygen was high (mean 79.5%). Like temperature, pH, conductivity and dissolved oxygen in mesopelagic waters were significantly lower than in epipelagic samples.

Bathypelagic parameters were similar to those of the mesopelagic zone but more variable. In terms of basins, temperature, pH, conductivity and dissolved oxygen were higher in the Southern basin, which is also more impacted by human activities and pollution, notably aromatic hydrocarbons and mercury brought by the Selenga river (Adams et al., 2018; Roberts et al., 2020), although only oxygen and, marginally, conductivity were significantly different (Supplementary Fig.1). The two main axes of a PCA considering these abiotic parameters explained 58% of the variance (Fig.1B). Surface samples correlated with higher temperature, conductivity and, to a lower extent, pH and dissolved oxygen. These observations suggest that depth, as a proxy for light accessibility but also temperature and other abiotic parameters, might be a strong environmental driver for community structure.

Composition of planktonic protist communities

To study the diversity and relative abundance of microbial eukaryotes in Lake Baikal plankton, we concentrated cells in the 0.2-30 µm diameter fraction by successive filtration steps. This fraction thus integrated pico- (0.2-2 µm), nano- (2-20 µm) and small microplankton (20-30 µm), covering a wider protistan spectrum than some previous comparative studies (Annenkova et al., 2020). We purified DNA and massively sequenced (MiSeq Illumina, 2x300 bp) multiplexed 18S rRNA gene V4-region amplicons. After discarding low-quality reads, we generated 6 405 343 high-quality merged paired-end sequences that we clustered in operational taxonomic units (OTUs) at different thresholds. We determined 27 504 OTUs and 9 700 OTUs at, respectively, 98% and 95% sequence identity (CD-HIT). SWARM yielded 11 590 OTUs (Supplementary Table 1), only slightly higher than the number of OTUs defined at the latter cut-off. For our subsequent comparative analyses, we deliberately retained OTUs defined at 95% sequence identity threshold. Many diversity studies focus on exact sequence variants after sequence error correction (Callahan et al., 2016) that can inform about individual strain variation. However, for the purpose of this comparative study, we chose to use conservatively defined OTUs that, on average (this varies across phylogenetic groups), correspond to the genus or species-genus level (Caron et al., 2009). This taxonomy cut-off level is relevant for broad comparative ecological studies (members of the same genus are likely to have similar general functions, despite inter- strain or species-specific niche differences), while operationally diminishing the number of handled OTUs. Based on sequence MIDs, the abundance of the different OTUs was determined for each sample (Supplementary Table 1). To avoid potential biases in diversity and relative abundance estimates linked to differences in the total number of reads, we rarefied sequences to the same number across samples, which resulted in a global number of 4 570 genus-level OTUs. Nonetheless, accumulation curves showed that the diversity of planktonic protists was far from reaching saturation, even at the conservative genus level (Supplementary Fig.2). Richness significantly decreased in deep as compared to surface waters; so did

evenness (Supplementary Fig.3). A lower evenness may be partly explained by the lower cell abundance in deeper waters, as the counts for each OTU become more aleatory. We did not observe richness differences across lake basins, but evenness appeared significantly higher in the Northern basin.

From a phylogenetic perspective, our defined OTUs affiliated to at least 27 eukaryotic phyla belonging to several major eukaryotic supergroups (Fig.1C; Supplementary Table 2): the SAR clade (Stramenopiles, Alveolata, Rhizaria), Amebozoa, Archaeplastida, Excavata, Opisthokonta and Hacrobia. Although we considered Hacrobia as originally described (Okamoto et al., 2009), they should be possibly split in two or more groups as the eukaryotic phylogeny progressively resolves (Burki et al., 2020). Ciliates and dinoflagellates (Alveolata), Ochrophyta (Stramenopiles) and Holomycota (Fungi and related lineages within the Opisthokonta) dominated plankton samples representing, respectively 48.4%, 21.5%, 12.6% and 8% relative sequence abundance. Cryptophyta, Haptophyta, Kathablepharida and Tenomemida displayed moderate abundances (0.5 to 5% reads) and were followed by a long tail of lower-frequency taxa in rank:abundance curves (Supplementary Fig.4). The major dominant groups were similar in all depths, with small variations in the deepest waters. Ciliates were by far the most abundant in terms of sequence reads. However, this observation is to be pondered by the fact that, in ciliate somatic macronuclei, rRNA genes are amplified several thousand times (e.g. ~9000 copies in *Tetrahymena thermophila* (Ward et al., 1997)), such that their relative abundance in term of cells is certainly much lower. Although diatoms (Bacillariophyta, Ochrophyta), several of them considered endemic, are well known in Lake Baikal plankton (Moore et al., 2009; Zakharova et al., 2013; Bashenkhaeva et al., 2015; Roberts et al., 2018; Mikhailov et al., 2019b), they represented only 6.1% ochrophyte reads distributed in 64 OTUs. Optical microscopy on board showed that diatoms were numerous, but their long frustules prevented most of them from being retained in the analyzed plankton fraction. Members of the Holomycota were very diverse. Classical fungi represented ca. 60% holomycotan sequences, most of them corresponding to chytrids, although the Dicarya (Ascomycota, Basidiomycota) were relatively abundant too (Supplementary Fig.5). Most Dicarya belonged to typical terrestrial fungi entering the lake waters with river in-flow or from the surrounding land. However, chytrids (flagellated fungi) are more likely to be truly planktonic organisms. Interestingly, members of Rozellida (Cryptomycota) and Aphelida, were also relatively abundant, making up to almost 40% of the holomycotan sequences. Rozellids and aphelids, together with their microsporidian relatives are parasites (Karpov et al., 2014; Bass et al., 2018). Although rozellids (cryptomycotes) are often included within fungi, they are phagotrophic organisms, unlike fungi (which are osmotrophs), and they branch more deeply than aphelids in the Holomycota tree (Torruella et al., 2018). Our data suggest that the majority of actual fungal-like planktoners in Lake Baikal are parasites.

Overall, despite methodological differences, our identified plankton protist communities were consistent with previous studies in surface waters or in a water column previously sampled in the Southern basin, with ciliates, dinoflagellates and ochrophytes being highly represented (Annenkova et al., 2020) (Yi et al., 2017).

Marine signature taxa

Although marine-freshwater transitions are thought to be rare (Mukherjee et al., 2019) and salinity, a major driver of microbial community composition (Lozupone and Knight, 2007), high-throughput environmental studies are revealing an increasing number of typically marine eukaryotic lineages in freshwater systems. Among those are members of the parasitic perkinsids (Brate et al., 2010), haptophytes (Simon et al., 2013), Bolidophyceae (Richards and Bass, 2005; Annenkova et al., 2020) and several Marine Stramenopiles (MAST) clades (Massana et al., 2004; Massana et al., 2006), such as MAST-2, MAST-12, MAST-3 and possibly MAST-6 (Simon et al., 2015a). Recently, diplonemids, a cosmopolitan group of oceanic excavates particularly abundant and diverse in the deep ocean (Lara et al., 2009; de Vargas et al., 2015) were identified in deep freshwater lakes (Yi et al., 2017; Mukherjee et al., 2019). Likewise, Syndiniales, a clade of parasitic alveolates (often parasitizing their dinoflagellate relatives) widely distributed in oceans (López-García et al., 2001; Guillou et al., 2008), were recently identified in Baikal surface plankton (Annenkova et al., 2020). We identified members of all these lineages in our large Lake Baikal plankton dataset, albeit mostly in low proportions (Fig.2A; Supplementary Table 3). Bolidophytes and, collectively, MAST clades were nonetheless relatively abundant in the lake. However, MAST clades are not monophyletic and they exhibited different abundance patterns. Clades previously detected in freshwater systems, MAST-2, MAST-6, MAST-12 and to a lesser extent MAST-3, were relatively abundant. But MAST clades not previously observed in other freshwater systems, including MAST-1, MAST-4, MAST-8 and MAST-20 occurred in very low proportions in a few samples. In addition to the rare diplonemids, which were widely but sporadically present across Lake Baikal samples (Fig. 2A-B), we identified OTUs belonging to the emblematic Radiolaria, to our knowledge never before identified in freshwater plankton. These OTUs were members of the Polycystinea (Fig.2C) and exhibited extremely low frequencies.

The low abundance of some of these typically marine lineages partly explains the fact that they failed to be detected in previous studies of freshwater systems, suggesting that these ecological transitions have been so far underestimated (Paver et al., 2018). However, an additional explanation might be found in the particular features of the Lake Baikal, including its considerable depth, marked oligotrophy and even the presence of deep-venting (Müller et al., 2001; Sherstyankin et al., 2006; UNDP-GEF, 2015), which make it qualify in all points but salinity as a freshwater sea.

Environmental drivers of protist community structure

To test whether planktonic protist communities were influenced by abiotic factors (clearly correlated to sample spatial origin; Fig.1), we carried out permutational multivariate analysis of variance (PERMANOVA) of Wisconsin-standardized Bray-Curtis distances between communities as a function of sample spatial origin. PERMANOVA tests revealed significant differences in microbial eukaryotic communities as a function of basin (latitudinal region), sampling site (coordinates) and depth within sampling sites (Table 1). However, the most influential effects were those of the water column location, 23.7%, which combine latitudinal and vertical determinants, and depth within each single water column, i.e. vertical variation alone (16.3%). The effect of the sampling basin was significant but small (5.3%). To better visualize differences between communities, we carried out an NMDS analysis on the global Bray-Curtis distance matrix. Points from most water columns did not show a marked differentiation, as most water columns overlapped to some extent (Supplementary Fig.6). Likewise, samples from different basins did not show a clear differentiation, although samples from the Southern basin tended to segregate from the two other basins (Fig.3A). Samples from coastal versus open waters did not segregate at all (Fig.3B). However, planktonic communities clearly segregated as a function of the water column zonation, with epipelagic, mesopelagic and bathypelagic communities well separated in the NMDS plot (Fig.3C). NMDS analyses based on SWARM-defined OTUs yielded very similar results (Supplementary Fig.7). These observations were statistically supported by ANOSIM tests, which showed significant and marked differences among communities according to depth, significant but weak differences according to basin origin, and no correlation at all between coastal and pelagic samples (Supplementary Table 4).

These results suggest that depth is the major environmental factor structuring Lake Baikal protist communities. Depth is in turn a proxy for a variety of abiotic parameters, notably light, but also, despite their limited variation, temperature, dissolved oxygen, conductivity and pH (Fig.1). These environmental variables and others, such as the nature of dissolved organic matter (TDS amount does not vary significantly; Fig.1B), are likely to influence prokaryotic communities as well (Kurilkina et al., 2016). Consequently, the nature of prey available for bacterivorous protists is possibly different. This may, in turn, select for protists with particular preying affinities, such that biotic interactions with other planktonic members may be also important determinant factors of community structure and function.

Functional groups and biotic interactions

To look for potential ecological interactions between members of protist communities, we first explored the distribution of major functional classes with depth. We attributed protists to three major

categories based on knowledge about the lifestyle and ecological function of the corresponding phylogenetic lineages: autotrophs, free-living heterotrophs and parasites (Supplementary Table 5). We acknowledge that these are very broad categories and that many photosynthetic organisms can be mixotrophs (Massana, 2011; Mitra et al., 2016). However, information about mixotrophy is still scarce and it is difficult to predict this ability from sequence data only. Therefore, our category 'autotrophs' included also photosynthetic organisms that can additionally use heterotrophic feeding modes. Free-living heterotrophs include predatory protists but also osmotrophic organisms feeding on organic matter, such as fungi or some Stramenopiles. The relative abundance of the three functional categories in Lake Baikal significantly followed the same trend in the three water column zones, with autotrophs being less abundant than heterotrophs and parasites being in much lower proportion (Fig.4). Low proportions of parasitic protists are consistent with affordable parasite loads for an ecosystem, as was previously observed (Simon et al., 2015b). Nonetheless, the relative amount of parasitic lineages diminished with depth, potentially suggesting that a relatively important proportion of protists identified in deep waters might be inactive. This is indeed likely the case for most photosynthetic organisms that were identified below the epipelagic region. Although the proportion of autotrophs diminished with depth, they still made up to 30% of the total in bathypelagic waters. As mentioned, some of these protists may be mixotrophic and prey on bacteria or other protists in the dark water column. However, the majority of photosynthetic lineages may simply be inactive, dormant or on their way to decay, serving as food for the heterotrophic component of microbial communities. The presence of relatively abundant photosynthetic protists in the Baikal dark water column and sediments is well documented (Zakharova et al., 2013; Yi et al., 2017), low temperatures possibly helping their preservation during sedimentation. Finally, free-living heterotrophs were the most abundant functional category throughout the water column. This might seem at odds with a pyramidal food-web structure whereby primary producers should be more abundant than consumers. However, several factors might explain this. First, the presence of ciliates likely introduces a positive bias in this functional category. Second, many autotrophs might be, on average, larger than heterotrophic protists and their biomass exceed that of consumers. Finally, many heterotrophic protists might depend on bacteria or on larger organisms (e.g. fungi degrading decaying plant material).

To further explore biotic interactions, and given that protist community differences were essentially seen throughout the water column, we reconstructed co-occurrence networks of OTUs found in epipelagic, mesopelagic and bathypelagic zones. To build the networks, we retained OTUs present in more than 20% of samples at relative abundances higher than 0.01% (Supplementary Table 6). The structure of the three networks was markedly different (Fig.5). The epipelagic network was denser, having more interconnected OTUs, more positive interactions and several hub-type OTUs that

interact with many OTUs. Mean node degrees were also higher in the epipelagic network (Supplementary Table 7). Meso- and bathypelagic networks had less connected nodes and most correlations were negative. Only one positive interaction was observed in mesopelagic waters (ciliate-fungus) and in bathypelagic waters (rozellid-ochrophyte). The latter might suggest a specific parasitic interaction. Although bathypelagic waters exhibited the least connected nodes, both the connectance and the clustering coefficient of the network were the highest. A block-model representation of the three networks indicated the occurrence of pairs of OTU sets sharing similar properties that were highly interconnected with each other and only loosely to other sets (Supplementary Fig.8).

Collectively, our network data suggest that epipelagic communities are more active and have more positive and negative interactions, whereas in very deep waters, communities are more stable with a restricted but strongly connected core.

Concluding remarks

Lake Baikal in Southern Siberia is a unique freshwater system by its volume, maximum depth (1 642 m) and topographical features that include rifting associated with hydrothermalism. With its highly oligotrophic waters, it amounts to an inner freshwater sea in all points but an extremely low salinity. Freshwater ecosystems are particularly threatened by climate change and, being located in Southern Siberia, one of the most rapidly changing zones, Lake Baikal is being severely impacted (Mackay et al., 2017). Yet, despite the importance of the lake and its uniqueness, its microbial planktonic communities have been only partially studied and we lack reference comprehensive comparative community data to assess ongoing and future change and infer how it may affect microbial functions and the ecology of the lake. In this study, we have analyzed the composition of microbial eukaryotic communities in plankton collected from different water columns along a transect of ~600 km across the three lake basins, from surface (5 m) to high depth (1 400 m) and from littoral to open waters. Our study shows widely diverse communities covering all eukaryotic supergroups, with ciliates, dinoflagellates, chrysophytes and flagellated fungi plus related lineages (rozellids, aphelids) being the most relatively abundant, together with cryptophytes, haptophytes, katablepharids, telonemids and several MAST lineages. Interestingly, confirming previous observations in Lake Baikal, we observed members of typically marine lineages, including bolidophytes, syndineans, diplonemids and, for the first time, radiolarians. These observations suggest that the salinity barrier is relatively easy to cross and that the 'marine' determinants might be more related with the oligotrophic nature of the system and the occurrence of a deep water column than with salinity itself. Despite the relatively homogeneous values of several physicochemical parameters, planktonic protist communities were highly and significantly stratified in Lake Baikal, suggesting that depth, as a proxy for light but also temperature, pH, oxygen

and conductivity, is a major determinant of community structure. By contrast, the effect of latitude (basins) was minor, if not negligible. Consistent with vertical stratification, the relative proportion of autotrophs, free-living heterotrophs and parasites is altered with depth, where photosynthetic lineages are still present but, like parasites, in lower proportions. Biotic factors are also important in structuring Lake Baikal communities. Co-occurrence network analyses showed highly interconnected communities in surface waters, with positive and negative interactions. By contrast, deep, bathypelagic communities exhibit much less connected OTUs, although they are strongly, and mostly negatively, correlated. This might be suggestive of much more diluted and potentially inactive populations, but with a conserved core of highly interconnected OTUs. Our results pave the way for future comparative analyses of protist communities through time, notably in the context of rapid climate change that is affecting Siberia and Lake Baikal.

Acknowledgments

We thank the crew of the R/V G. Titov for their professionalism and efficiency onboard, the director of the Limnological Institute at Irkusk for logistical assistance and Philippe Deschamps for technical bioinformatic support. This research was funded by the European Research Council Grants ProtistWorld (322669, PL-G) and PlastEvol (787904, DM) as well as the Russian State grant 0345-2016-0009 (NVA).

Author contributions

PLG, DM and NVA designed the work and organized the limnological cruise. PLG, PB, AILA, LG, GR and NVA collected and processed water samples during the cruise. PB purified DNA and carried out PCR reactions for metabarcoding analysis. GR carried out the initial bioinformatic analysis of amplicon sequences. GD carried out metabarcoding, statistical and network analyses, with help from LJ. PLG wrote the manuscript with input from co-authors. All authors read, critically commented and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

Adams, J.K., Martins, C.C., Rose, N.L., Shchetnikov, A.A., and Mackay, A.W. (2018) Lake sediment records of persistent organic pollutants and polycyclic aromatic hydrocarbons in southern Siberia mirror the changing fortunes of the Russian economy over the past 70 years. *Environ Pollut* **242**: 528-538.

Annenkova, N.V., Lavrov, D.V., and Belikov, S.I. (2011) Dinoflagellates associated with freshwater sponges from the ancient lake baikal. *Protist* **162**: 222-236.

Annenkova, N.V., Giner, C.R., and Logares, R. (2020) Tracing the origin of planktonic protists in an ancient lake. *Microorganisms* **8**.

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2019) EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Syst Biol* **68**: 365-369.

Barrat, A., Barthélemy, M., and Vespignani, A. (2008) *Dynamical Processes on Complex Networks*. Cambridge: Cambridge University Press.

Bashenkhaeva, M.V., Zakharova, Y.R., Petrova, D.P., Khanaev, I.V., Galachyants, Y.P., and Likhoshway, Y.V. (2015) Sub-ice microalgal and bacterial communities in freshwater Lake Baikal, Russia. *Microb Ecol* **70**: 751-765.

Bass, D., Czech, L., Williams, B.A.P., Berney, C., Dunthorn, M., Mahe, F. et al. (2018) Clarifying the relationships between Microsporidia and Cryptomycota. *J Eukaryot Microbiol* **65:** 773-782.

Bel'kova, N.L., Parfenova, V.V., Kostopnova, T., Denisova, L., and Zaichikov, E.F. (2003) [Microbial biodiversity in the Lake Baikal water]. *Mikrobiologiia* **72**: 239-249.

Belikov, S., Belkova, N., Butina, T., Chernogor, L., Martynova-Van Kley, A., Nalian, A. et al. (2019) Diversity and shifts of the bacterial community associated with Baikal sponge mass mortalities. *PLoS One* **14**: e0213926.

Bower, S.M., Carnegie, R.B., Goh, B., Jones, S.R.M., Lowe, G.J., and Mak, M.W.S. (2004) Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues. *J Euk Microbiol* **51**: 325-332.

Brate, J., Logares, R., Berney, C., Ree, D.K., Klaveness, D., Jakobsen, K.S., and Shalchian-Tabrizi, K. (2010) Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *ISME J* **4**: 1144-1153.

Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020) The New Tree of Eukaryotes. *Trends in Ecology & Evolution* **35**: 43-55.

Butina, T.V., Bukin, Y.S., Krasnopeev, A.S., Belykh, O.I., Tupikin, A.E., Kabilov, M.R. et al. (2019) Estimate of the diversity of viral and bacterial assemblage in the coastal water of Lake Baikal. *FEMS Microbiol Lett* **366**.

Cabello-Yeves, P.J., Zemskaya, T.I., Rosselli, R., Coutinho, F.H., Zakharenko, A.S., Blinov, V.V., and Rodriguez-Valera, F. (2018) Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. *Appl Environ Microbiol* **84**.

Cabello-Yeves, P.J., Zemskaya, T.I., Zakharenko, A.S., Sakirko, M.V., Ivanov, V.G., Ghai, R., and Rodriguez-Valera, F. (2020) Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. *Limnol Oceanogr* **n/a**.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**: 581-583.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D. et al. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75**: 5797-5808.

Cavicchioli, R., Ripple, W.J., Timmis, K.N., Azam, F., Bakken, L.R., Baylis, M. et al. (2019) Scientists' warning to humanity: microorganisms and climate change. *Nat Rev Microbiol* **17**: 569-586.

Chiquet, J., Mariadassou, M., and Robin, S. (2018) Variational inference for probabilistic Poisson PCA. *Ann Appl Stat* **12**: 2674-2698.

Csardi, G., and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**: 1-9.

De Batist, M., Klerkx, J., Van Rensbergen, P., Vanneste, M., Poort, J., Golmshtok, A.Y. et al. (2002) Active hydrate destabilization in Lake Baikal, Siberia? *Terra Nova* **14**: 436-442.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R. et al. (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Delmas, E., Besson, M., Brice, M.-H., Burkle, L.A., Dalla Riva, G.V., Fortin, M.-J. et al. (2019) Analysing ecological networks of species interactions. *Biological Reviews* **94**: 16-36.

Dray, S., and Dufoour, A.B. (2007) The ade4 Package : Implementing the duality diagram for ecologists. *Journal of Statistical Software* **22-4**.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150-3152.

Glöckner, F.O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. (2000) Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl Environ Microbiol* **66**: 5053-5065.

Granin, N.G., Aslamov, I.A., Kozlov, V.V., Makarov, M.M., Kirillin, G., McGinnis, D.F. et al. (2019) Methane hydrate emergence from Lake Baikal: direct observations, modelling, and hydrate footprints in seasonal ice cover. *Sci Rep* **9**: 19361.

Guillou, L., Viprey, M., Chambouvet, A., Welsh, R.M., Kirkham, A.R., Massana, R. et al. (2008) Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* **10**: 3349-3365.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L. et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: D597-D604.

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

Karpov, S.A., Mamkaeva, M.A., Aleoshin, V.V., Nassonova, E., Lilje, O., and Gleason, F.H. (2014) Morphology, phylogeny, and ecology of the aphelids (Aphelidea, Opisthokonta) and proposal for the new superphylum Opisthosporidia. *Front Microbiol* **5**: 112.

Kashiwaya, K., Ochiai, S., Sakai, H., and Kawai, T. (2001) Orbit-related long-term climate cycles revealed in a 12-Myr continental record from Lake Baikal. *Nature* **410**: 71-74.

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

Kurilkina, M.I., Zakharova, Y.R., Galachyants, Y.P., Petrova, D.P., Bukin, Y.S., Domysheva, V.M. et al. (2016) Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing. *FEMS Microbiol Ecol* **92**.

Lara, E., Moreira, D., Vereshchaka, A., and Lopez-Garcia, P. (2009) Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. *Environ Microbiol* **11**: 47-55.

Lê, S., Josse, J., and Husson, F. (2008) FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software* **25**: 1-18.

Leger, J.B. (2016) Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates. *http://arxivorg/abs/160207587*.

López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603-607.

Lozupone, C.A., and Knight, R. (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* **104**: 11436-11440.

Mackay, A.W., Seddon, A.W., Leng, M.J., Heumann, G., Morley, D.W., Piotrowska, N. et al. (2017) Holocene carbon dynamics at the forest-steppe ecotone of southern Siberia. *Glob Chang Biol* **23**: 1942-1960.

Magoc, T., and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957-2963.

Mahe, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**: e1420.

Maksimov, V.V., Shchetinina, E.V., Kraikivskaia, O.V., Maksimov, V.N., and Maksimova, E.A. (2002) [The classification and the monitoring of the state of mouth riverine and lacustrine ecosystems in lake Baikal based on the composition of local microbiocenoses and their activity]. *Mikrobiologiia* **71**: 690-696.

Maksimova, E.A., and Maksimov, V.N. (1972) [Vertical distribution of microbial plankton in the Southern part of Lake Baikal in 1969]. *Mikrobiologiia* **41**: 896-902.

Markovic, D., Carrizo, S.F., Kärcher, O., Walz, A., and David, J.N.W. (2017) Vulnerability of European freshwater catchments to climate change. *Glob Chang Biol* **23**: 3567-3580.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJournal* **17**: 10-12.

Massana, R. (2011) Eukaryotic picoplankton in surface oceans. *Annu Rev Microbiol* **65**: 91-110.

Massana, R., Terrado, R., Forn, I., Lovejoy, C., and Pedros-Alio, C. (2006) Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environ Microbiol* **8**: 1515-1522.

Massana, R., Castresana, J., Balague, V., Guillou, L., Romari, K., Groisillier, A. et al. (2004) Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl Environ Microbiol* **70**: 3528-3534.

Mats, V.D., and Perepelova, T.I. (2011) A new perspective on evolution of the Baikal Rift. *Geoscience Frontiers* **2**: 349-365.

Mikhailov, I.S., Zakharova, Y.R., Bukin, Y.S., Galachyants, Y.P., Petrova, D.P., Sakirko, M.V., and Likhoshway, Y.V. (2019a) Co-occurrence networks among bacteria and microbial eukaryotes of Lake Baikal during a spring phytoplankton bloom. *Microb Ecol* **77**: 96-109.

Mikhailov, I.S., Zakharova, Y.R., Galachyants, Y.P., Usoltseva, M.V., Petrova, D.P., Sakirko, M.V. et al. (2015) Similarity of structure of taxonomic bacterial communities in the photic layer of Lake Baikal's three basins differing in spring phytoplankton composition and abundance. *Dokl Biochem Biophys* **465**: 413-419.

Mikhailov, I.S., Bukin, Y.S., Zakharova, Y.R., Usoltseva, M.V., Galachyants, Y.P., Sakirko, M.V. et al. (2019b) Co-occurrence patterns between phytoplankton and bacterioplankton across the pelagic zone of Lake Baikal during spring. *J Microbiol* **57**: 252-262.

Mitra, A., Flynn, K.J., Tillmann, U., Raven, J.A., Caron, D., Stoecker, D.K. et al. (2016) Defining planktonic protist functional groups on mechanisms for energy and nutrient acquisition: incorporation of diverse mixotrophic strategies. *Protist* **167**: 106-120.

Moore, M.V., Hampton, S.E., Izmest'eva, L.R., Silow, E.A., Peshkova, E.V., and Pavlov, B.K. (2009) Climate change and the world's "Sacred Sea"—Lake Baikal, Siberia. *BioScience* **59**: 405-417.

Mukherjee, I., Hodoki, Y., Okazaki, Y., Fujinaga, S., Ohbayashi, K., and Nakano, S.I. (2019) Widespread dominance of kinetoplastids and unexpected presence of diplonemids in deep freshwater lakes. *Front Microbiol* **10**: 2375.

Müller, J., Oberhänsli, H., Melles, M., Schwab, M., Rachold, V., and Hubberten, H.W. (2001) Late Pliocene sedimentation in Lake Baikal: implications for climatic and tectonic change in SE Siberia. *Palaeogeogr Palaeoclimatol Palaeoecol* **174**: 305-326.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268-274.

Okamoto, N., Chantangsi, C., Horak, A., Leander, B.S., and Keeling, P.J. (2009) Molecular phylogeny and description of the novel katablepharid Roombia truncata gen. et sp. nov., and establishment of the Hacrobia taxon nov. *PLoS ONE* **4**: e7080.

Oksanen, J., Blanchet, G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L. et al. (2011) Vegan: Community Ecology Package. R package version 1.17-9. In. http://CRAN.R-project.org/package=vegan (ed): http://CRAN.R-project.org/package=vegan.

Paver, S.F., Muratore, D., Newton, R.J., and Coleman, M.L. (2018) Reevaluating the salty divide: phylogenetic specificity of transitions between marine and freshwater systems. *mSystems* **3**.

Pielou, E.C. (1966) Species-diversity and pattern-diversity in the study of ecological succession. *J Theor Biol* **10**: 370-383.

Potapov, S.A., Tikhonova, I.V., Krasnopeev, A.Y., Kabilov, M.R., Tupikin, A.E., Chebunina, N.S. et al. (2019) Metagenomic analysis of virioplankton from the pelagic zone of Lake Baikal. *Viruses* **11**.

Prokopenko, A.A., Karabanov, E.B., and Williams, D.F. (2002) Age of long sediment cores from Lake Baikal. *Nature* **415**: 976.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-596.

R Development Core Team (2017) R: A language and environment for statistical computing. In. http://www.r-project.org (ed). Vienna, Austria: R Foundation for Statistical Computing.

Richards, T.A., and Bass, D. (2005) Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis. *Curr Opin Microbiol* **8**: 240-252.

Roberts, S., Adams, J.K., Mackay, A.W., Swann, G.E.A., McGowan, S., Rose, N.L. et al. (2020) Mercury loading within the Selenga River basin and Lake Baikal, Siberia. *Environ Pollut* **259**: 113814.

Roberts, S.L., Swann, G.E.A., McGowan, S., Panizzo, V.N., Vologina, E.G., Sturm, M., and Mackay, A.W. (2018) Diatom evidence of 20th century ecosystem change in Lake Baikal, Siberia. *PLoS One* **13**: e0208765.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahe, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Rolls, R.J., Heino, J., Ryder, D.S., Chessman, B.C., Growns, I.O., Thompson, R.M., and Gido, K.B. (2018) Scaling biodiversity responses to hydrological regimes. *Biol Rev Camb Philos Soc* **93**: 971-995.

Schmid, M., Budnev, N.M., Granin, N.G., Sturm, M., Schurter, M., and Wüest, A. (2008) Lake Baikal deepwater renewal mystery solved. *Geophys Res Lett* **35**: L09605-L09605.

Schuur, E.A., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J. et al. (2015) Climate change and the permafrost carbon feedback. *Nature* **520**: 171-179.

Sherstyankin, P.P., Alekseev, S.P., Abramov, A.M., Stavrov, K.G., De Batist, M., Hus, R. et al. (2006) Computer-based bathymetric map of Lake Baikal. *Dokl Earth Sci* **408**: 564-569.

Shimaraev, M.N., and Domysheva, V.M. (2012) Trends in hydrological and hydrochemical processes in Lake Baikal under conditions of modern climate change. In. Chichester, UK: John Wiley & Sons, Ltd, pp. 43-66.

Simon, M., Lopez-Garcia, P., Moreira, D., and Jardillier, L. (2013) New haptophyte lineages and multiple independent colonizations of freshwater ecosystems. *Environ Microbiol Rep* **5**: 322-332.

Simon, M., Jardillier, L., Deschamps, P., Moreira, D., Restoux, G., Bertolino, P., and Lopez-Garcia, P. (2015a) Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. *Environ Microbiol* **17**: 3610-3627.

Simon, M., Lopez-Garcia, P., Deschamps, P., Moreira, D., Restoux, G., Bertolino, P., and Jardillier, L. (2015b) Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. *ISME J* **9**: 1941-1953.

Torruella, G., Grau-Bove, X., Moreira, D., Karpov, S.A., Burns, J.A., Sebe-Pedros, A. et al. (2018) Global transcriptome analysis of the aphelid *Paraphelidium tribonemae* supports the phagotrophic origin of fungi. *Commun Biol* **1**: 231.

Troitskaya, E., Blinov, V., Ivanov, V., Zhdanov, A., Gnatovsky, R., Sutyrina, E., and Shimaraev, M. (2015) Cyclonic circulation and upwelling in Lake Baikal. *Aquatic Sciences* **77**: 171-182.

UNDP-GEF (2015) The ecological atlas of the Baikal basin. In. http://baikal.iwlearn.org/en: United Nations Office for Project Sercives (UNOPS), p. 145.

Ward, J.G., Blomberg, P., Hoffman, N., and Yao, M.C. (1997) The intranuclear organization of normal, hemizygous and excision-deficient rRNA genes during developmental amplification in Tetrahymena thermophila. *Chromosoma* **106**: 233-242.

Wilburn, P., Shchapov, K., Theriot, E.C., and Litchman, E. (2020) Environmental drivers define contrasting microbial habitats, diversity, and community structure in Lake Baikal, Siberia. *bioRxiv*: 605899.

Woodward, G., Perkins, D.M., and Brown, L.E. (2010) Climate change and freshwater ecosystems: impacts across multiple levels of organization. *Philos Trans R Soc Lond B Biol Sci* **365**: 2093-2106.

Yi, Z., Berney, C., Hartikainen, H., Mahamdallie, S., Gardner, M., Boenigk, J. et al. (2017) High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations. *FEMS Microbiol Ecol* **93**: 10.

Zakharenko, A.S., Galachyants, Y.P., Morozov, I.V., Shubenkova, O.V., Morozov, A.A., Ivanov, V.G. et al. (2019) Bacterial communities in areas of oil and methane seeps in pelagic of Lake Baikal. *Microb Ecol* **78**: 269-285.

Zakharova, Y.R., Galachyants, Y.P., Kurilkina, M.I., Likhoshvay, A.V., Petrova, D.P., Shishlyannikov, S.M. et al. (2013) The structure of microbial community and degradation of diatoms in the deep near-bottom layer of Lake Baikal. *PLoS One* **8**: e59977.

Figure Legends

Fig. 1. Sampling sites and overall planktonic protist community composition in Lake Baikal. **A**, map of Lake Baikal showing the different sampling sites across the three major lake basins (indicated by colors). **B**, Principal component analysis (PCA) of samples according to their associated physicochemical parameters. The number near the points correspond to the sampling site, and the color of the points indicates their sampling depth. TDS, total dissolved solids; DO, dissolved oxygen; ORP, oxidation-reduction potential. Blue tones indicate the sampling depth in the water column, as indicated. **C**, Relative abundance of different high-rank eukaryotic taxa in Baikal plankton based on read counts for the defined OTUs. The asterisk indicates samples retrieved from Cell-Traps (Methods). Color codes for sample basin and depth origin as well as for the different taxa are indicated.

Fig. 2. Marine signature taxa detected in Lake Baikal plankton. **A**, heat map showing the relative abundance of different typically marine taxa across Baikal plankton samples. The frequency of the different phylogenetic groups is indicated by different shades of blue. **B**, Maximum Likelihood (ML) phylogenetic tree of OTUs belonging to diplonemids and a related group of euglenozoan excavates (594 unambiguously aligned positions). **C**, ML tree of radiolarian OTUs (534 unambiguously aligned positions).

Fig. 3. Non-metric multidimensional scaling (NMDS) analysis of Lake Baikal plankton samples as a function of protist community similarities. The NMDS plot was constructed with Wisconsin-standardized Bray-Curtis dissimilarities between all samples. A, plankton samples highlighted by basin origin. B, plankton samples from coastal, shallow sites versus open water sites. C, samples grouped according to their depth origin in the water column; epipelagic (<200 m), mesopelagic (200-500 m), bathypelagic (>500 m).

Fig. 4. Box plots showing the distribution of relative abundances of major functional categories of protists in Lake Baikal plankton. The three plots show the relative abundance of sequences affiliated to autotrophic, heterotrophic and parasitic protists for each sampling depth class. The thickest line inside each box represents the median on the distribution, bottom and top borders of boxes correspond to the first and the third quartiles and whiskers extend to minimal and maximal distances. Significant differences between distributions are indicated with stars (p-values <0.05, <0.005 and <0.0005 are respectively indicated by one, two and three stars). For the assignation of taxa to functional categories, see Supplementary Table 5.

Fig. 5. Co-occurrence networks of planktonic protists in the Lake Baikal water column. A, network obtained from epipelagic (<200 m) samples across the lake. B, network obtained from mesopelagic (200-500 m) samples. C, network obtained from bathypelagic (>500 m samples). Networks were built on OTUs present in more than 20% samples and having a relative abundance higher than 0.01%. OTUs are represented by nodes and direct covariations between them, by edges. Nodes and taxa labeled with an asterisk correspond to putative parasites.

Table 1. Permutational multivariate analyse of variance (PERMANOVA) of Lake Baikal planktonic protist communities across basins, sampling site and depth. PERMANOVA was calculated using Wisconsin standardization on rarefied OTUs belonging to the 65 studied plankton samples. Df, degrees of freedom.

Effect	Df	F.Model	R^2	P value
Region	2	1.8369	5.30%	***
Sampling site	15	1.1857	23.70%	***
Depth | Sampling site	9	1.2861	16.30%	***

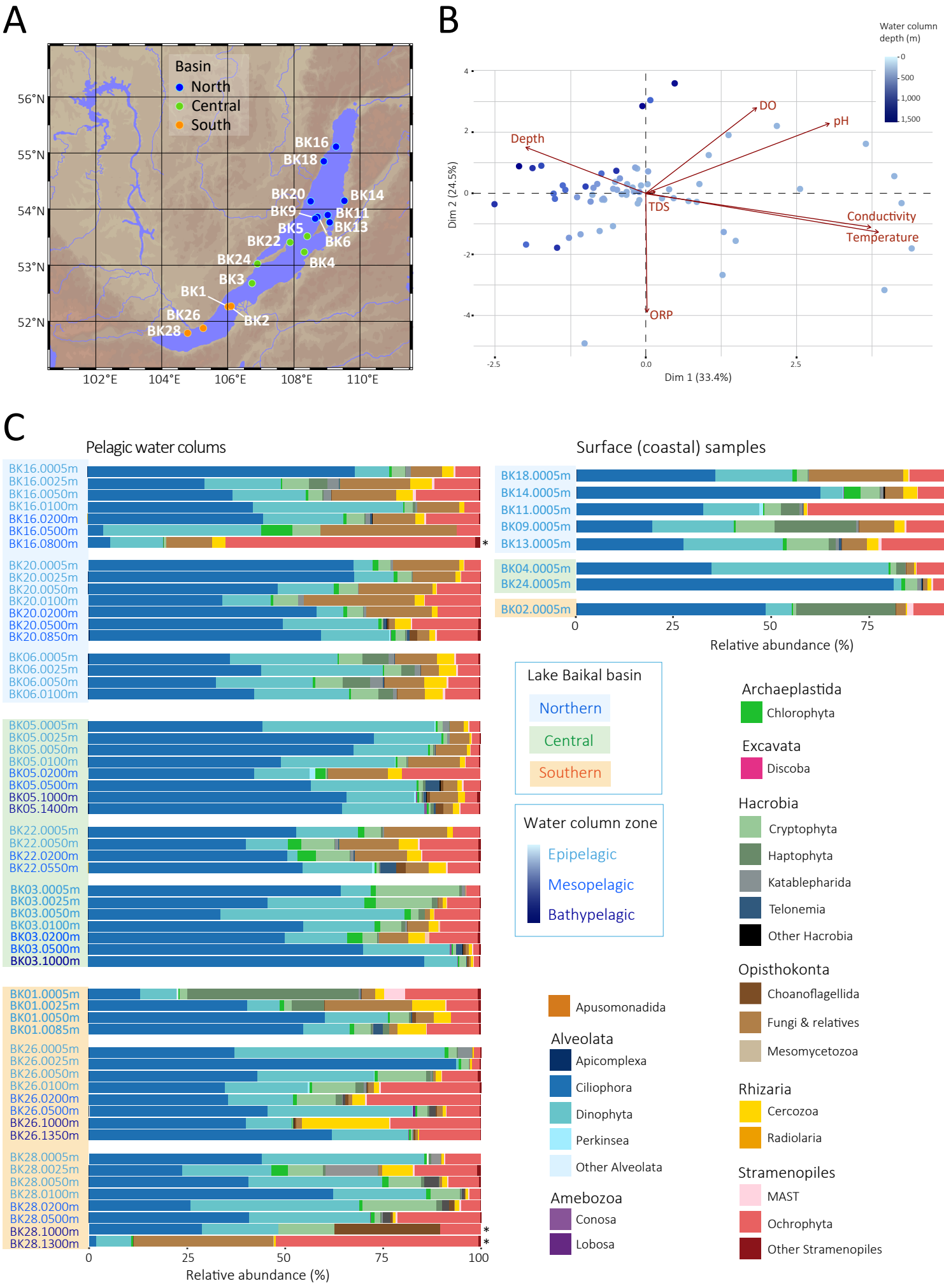


Figure 1. G. David et al.

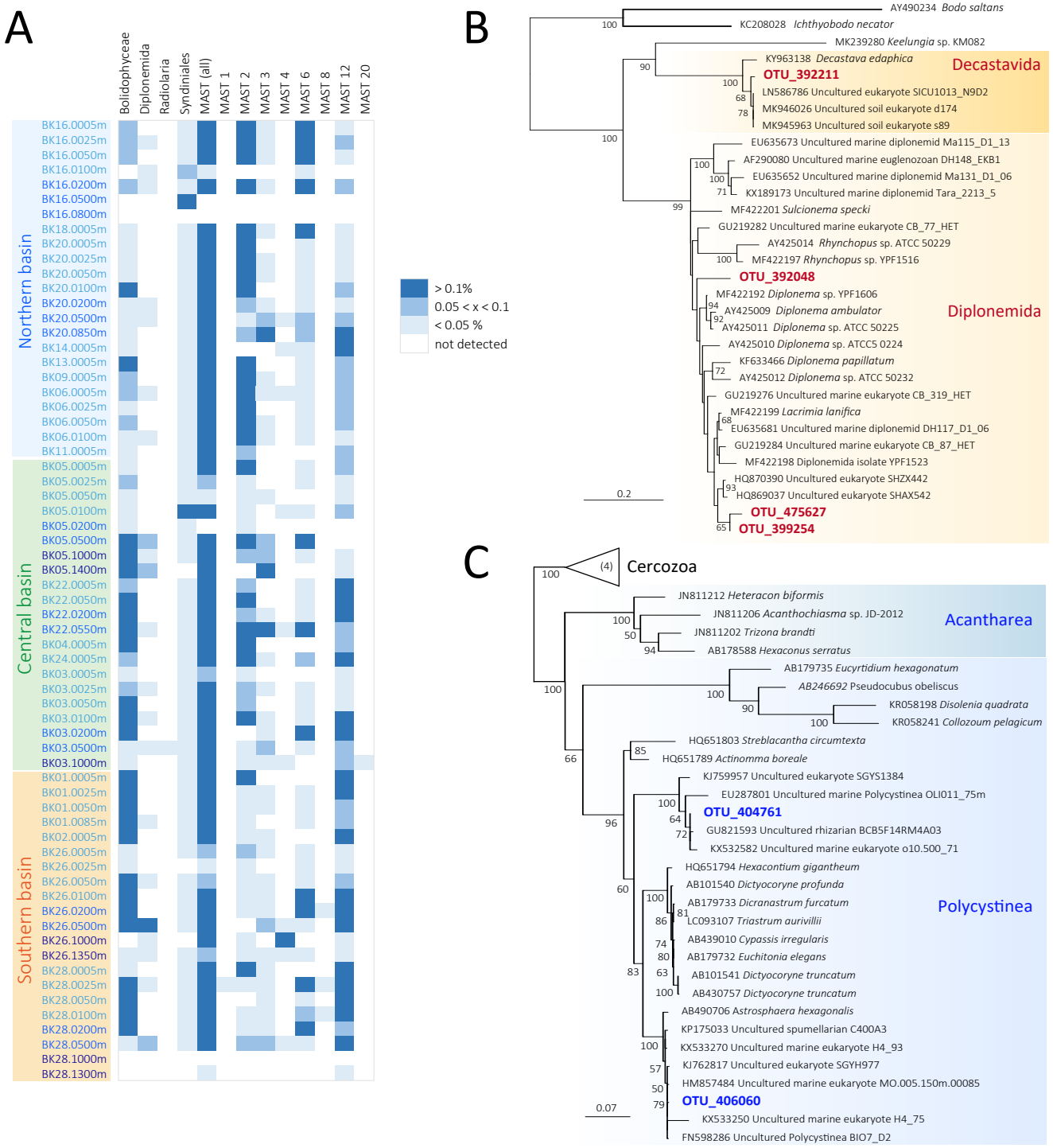


Figure 2. G. David et al.

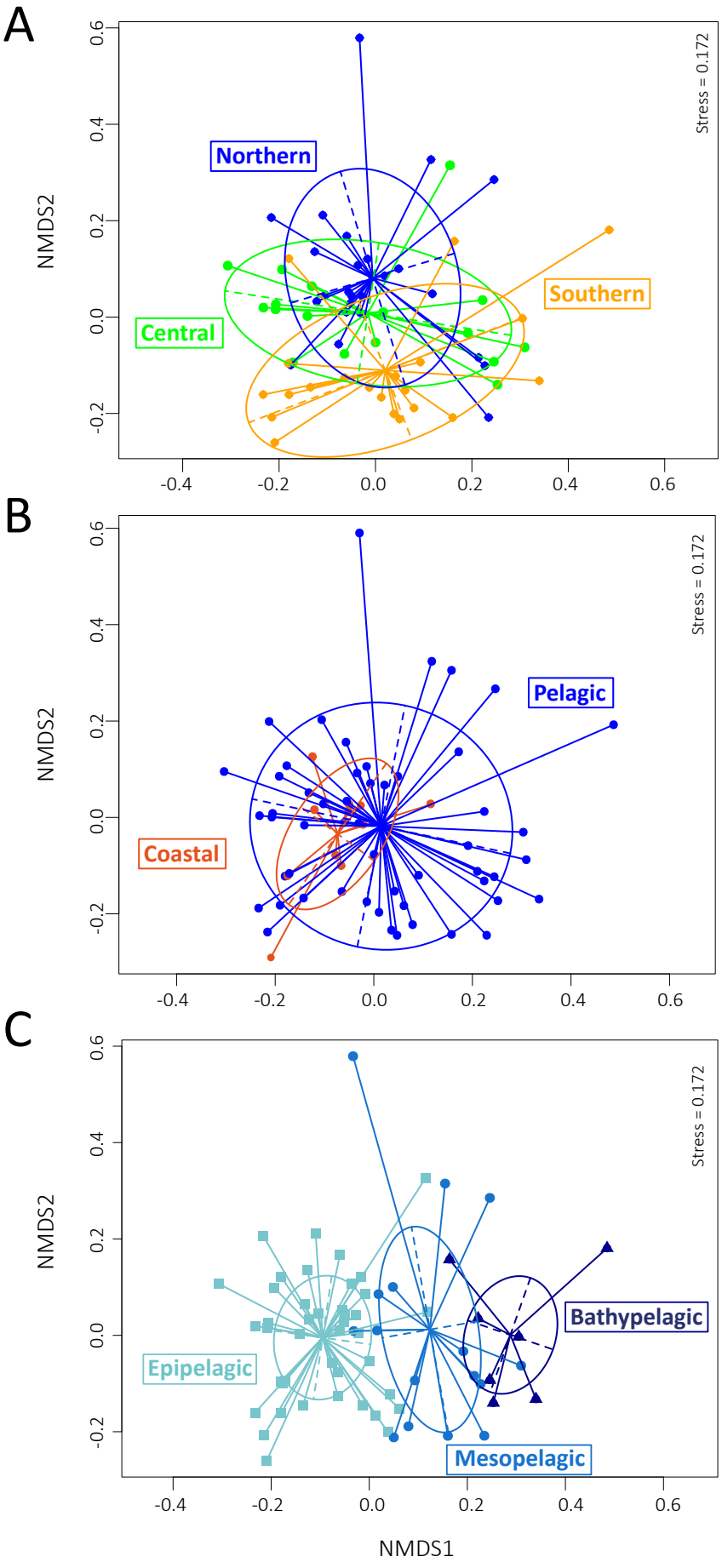


Figure 3. G. David et al.

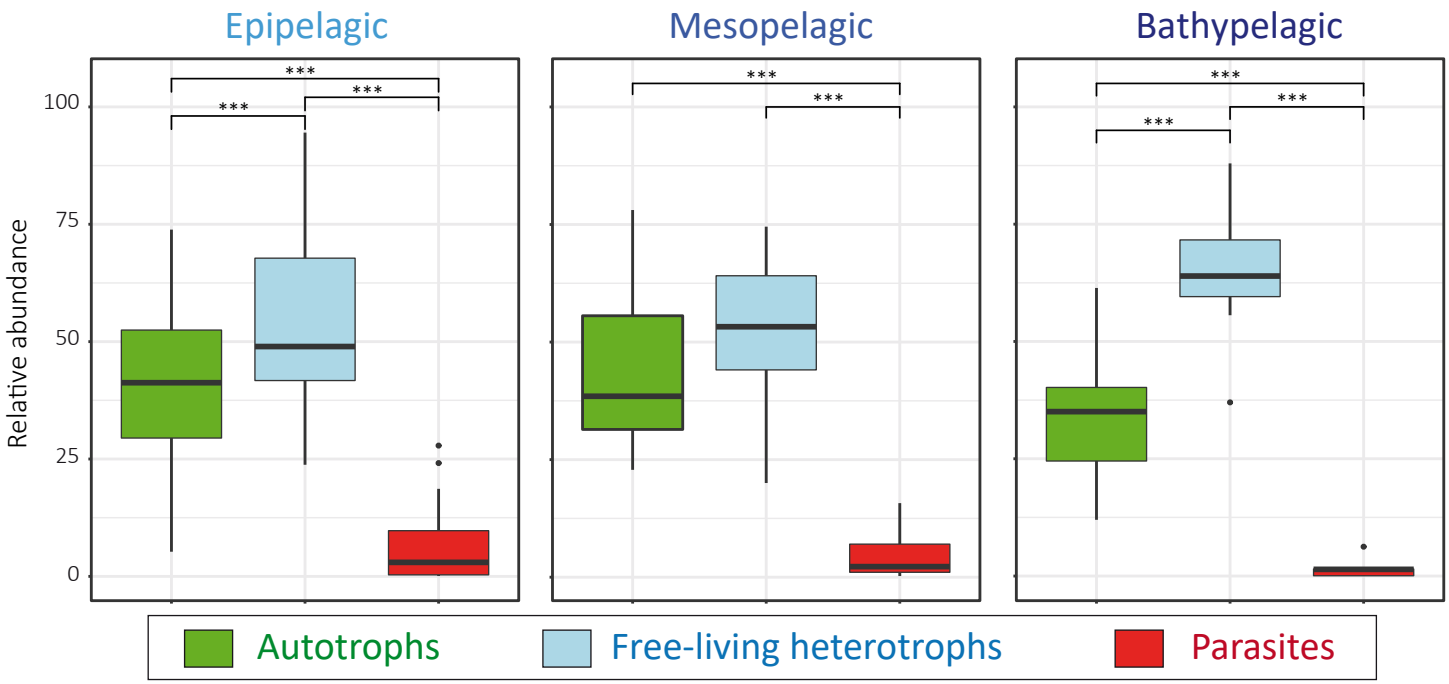


Figure 4. G. David et al.

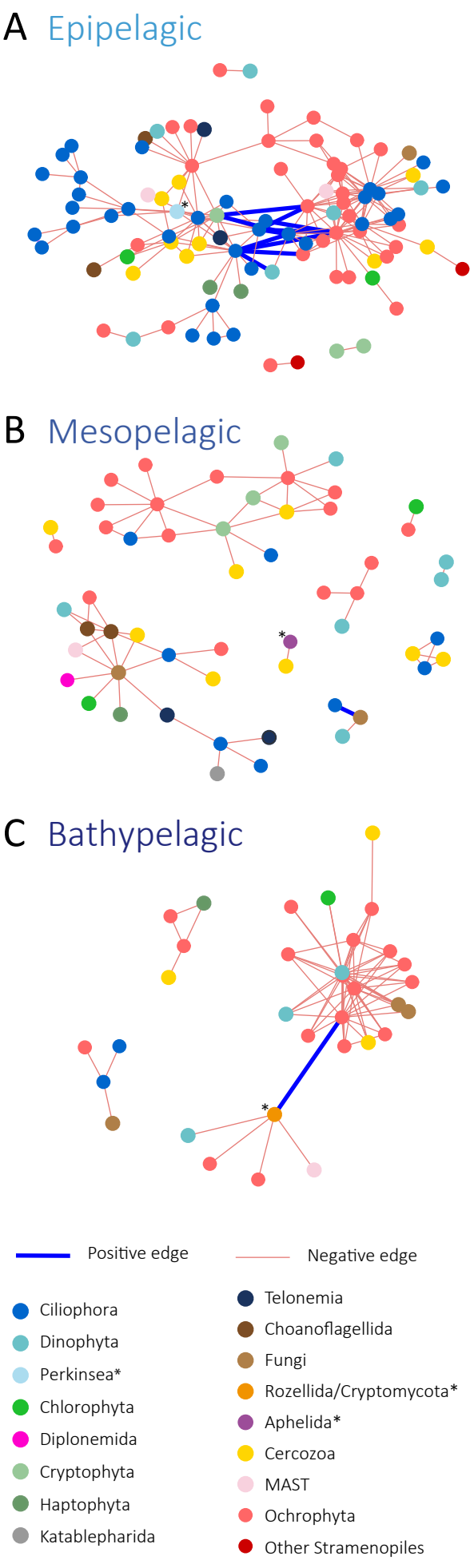


Figure 5. G. David et al.

APPENDIX

E

FIFTH APPENDIX

Ancient Adaptive Lateral Gene Transfers in the Symbiotic *Opalina* - *Blastocystis* Stramenopile Lineage

Naoji Yubuki[1], Luis J. Galindo[1], Guillaume Reboul[1], Purificación López-García[1], Matthew W. Brown[2,3], Nicolas Pollet[4], and David Moreira[1]*

[1]Unité d'Ecologie Systématique et Evolution, CNRS, Université Paris-Sud, AgroParisTech, Université Paris-Saclay, Orsay, France

[2]Department of Biological Sciences, Mississippi State University, MS Mississippi State, MS, USA

[3]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS, USA

[4]Laboratoire Evolution Génomes Comportement Ecologie. CNRS - Université Paris-Sud, Université Paris-Saclay, France

***Corresponding author:** E-mail: david.moreira@u-psud.fr.

Abstract

Lateral gene transfer (LGT) is a very common process in bacterial and archaeal evolution, playing an important role in the adaptation to new environments. In eukaryotes, its role and frequency remain highly debated, although recent research supports that gene transfer from bacteria to diverse eukaryotes may be much more common than previously appreciated. However, most of this research focused on animals and the true phylogenetic and functional impact of bacterial genes in less-studied microbial eukaryotic groups remains largely unknown. Here, we have analyzed transcriptome data from the deep-branching stramenopile Opalinidae, common members of frog gut microbiomes and distantly related to the well-known genus *Blastocystis*. Phylogenetic analyses suggest the early acquisition of several bacterial genes in a common ancestor of both lineages. Those LGTs most likely facilitated the adaptation of the free-living ancestor of the Opalinidae-*Blastocystis* symbiotic group to new niches in the oxygen-depleted animal gut environment.

Key words: Opalinids, *Blastocystis*, lateral gene transfer, gut microbiome.

Lateral gene transfer (LGT) plays an important role in prokaryotic evolution. LGT provides bacteria and archaea with the possibility to adapt, sometimes very rapidly, to new environments by obtaining genes from organisms already living in those environments. Although the significance of this phenomenon is widely recognized in prokaryotes, LGT-mediated gene acquisition from distant donors remains a contentious issue in eukaryotes (Martin 2017; Leger et al. 2018). Nevertheless, there is increasing evidence for LGT in eukaryotes from prokaryotes as well as from other eukaryotes (e.g., Keeling and Palmer 2008; Karnkowska et al. 2016; Eme et al. 2017; Husnik and McCutcheon 2017). A recent example concerns the stramenopile *Blastocystis,* which experienced LGTs from both eukaryotic and prokaryotic donors (Denoeud et al. 2011; Eme et al. 2017).

Blastocystis is recognized as the most widespread human gut eukaryotic parasite (Clark et al. 2013). This strict anaerobic and single-celled protist displays some unique and interesting biological features, such as the presence of unusual mitochondrion-related organelles (MRO) that display functions of mitochondria, hydrogenosomes and mitosomes (Stechmann et al. 2008). Some *Blastocystis* enzymes crucial for life in oxygen-depleted conditions were acquired by LGT from prokaryotes. For instance, the sulfur-mobilization (SUF) machinery involved in Fe-S protein maturation in the cytoplasm appears to have been acquired from archaeal Methanomicrobiales (Tsaousis et al. 2012). Furthermore, Eme et al. (2017) reported 74 purported cases of LGT mostly from prokaryotes to various subtypes of *Blastocystis* and suggested that several of the new LGT-acquired functions facilitated the metabolic adaptation of *Blastocystis* to the human gut in terms of metabolism but also to escape the immune defense mechanisms. The origins of those 74 gene families were very diverse. Although many of them were already present in the common ancestor of several *Blastocystis* subtypes, the time of their acquisition remained unclear due to the poor taxon sampling available for closely related stramenopile lineages.

Together with Alveolata and Rhizaria, Stramenopiles (or Heterokonta) constitute one of the main clades of the eukaryotic super-group SAR (Burki et al. 2007; Adl et al. 2019). Stramenopiles mostly encompass free-living phagotrophs or photosynthetic algae, but some are well-known parasites, such as the oomycetes and *Blastocystis*, or commensals, such as the Opalinidae (Patterson 1989; Andersen 2004). Ribosomal RNA phylogenetic analyses suggested a close relationship

between *Blastocystis* and Opalinidae, supporting the existence of a deep-branching symbiotic (parasitic/commensal) clade adapted to live in the gut of very diverse vertebrates (Silberman et al. 1996; Kostka et al. 2004; Li et al. 2018). However, despite the phylogenetic affinity of *Opalina* and *Blastocystis*, their morphological characteristics and lifestyles are very different. *Blastocystis* is characterized by a round unflagellated cell largely filled by a large vacuole. The cytoplasm and organelles are concentrated in the thin peripheral area between the vacuole and the cell membrane. Members of the genus *Blastocystis* live in the intestines of humans, birds, cows and pigs, most likely as parasites (Tan 2004). By contrast, members of the genus *Opalina* have a leaf-like cell shape with numerous nuclei and hundreds of short flagella on the cell surface, which is reminiscent of the cellular organization of ciliates. They live mainly in the intestine of anurans (frogs and toads) but seem to be innocuous to their hosts being therefore most often reported as commensal symbionts (Kostka 2016). Using the numerous flagella, *Opalina* members actively move in the intestine. All other known Opalinidae species are also commensal symbionts (Kostka 2016). Phylogenetic analyses have supported the monophyly of the Opalinidae-*Blastocystis* clade with the Placidida, a lineage of small free-living marine flagellates such as *Wobblia* and *Placidia* (Li et al. 2018; Shiratori et al. 2015, 2017; Derelle et al. 2016). Another free-living marine flagellate, *Cantina marsupialis*, is an anaerobic deep-branching relative that also possesses MROs (Yubuki et al. 2015). Since the closest relatives of Opalinidae and *Blastocystis* are all free-living, their ancestor was most likely free-living as well.

Here, we report the first transcriptome sequences from two Opalinidae strains, *Opalina* sp. OP10 and Opalinidae sp. Opal32, from two different continents (Europe and North America). OP10 and Opal32 cells were collected manually from the intestine of a *Xenopus tropicalis* frog and a *Lithobates sphenocephalus* tadpole, respectively. After transcriptome sequencing and assembly, we decontaminated the protein sequences inferred from the two transcriptomes to remove host and bacterial sequences (see Materials and Methods) and kept 7,232 and 18,765 proteins for OP10 and Opal32, respectively. Using BUSCO (Simão et al. 2015), we determined 33.3% transcriptome completeness for OP10 and 57.4% for Opal32. For comparison, we also applied BUSCO on the near-complete genome of *Blastocystis hominis* and determined a completeness of 75.2%, indicating a reduced genome as expected for a derived parasite. We found in our datasets 44.3% (OP10) and 76.3%

(Opal32) of the *Blastocystis* proteome (supplementary table S1, Supplementary Material online), suggesting a rather good coverage especially for Opal32.

After transcriptome decontamination we searched with BLAST (Camacho et al. 2009) Opalinidae homologues of the 74 gene families likely acquired by LGT in *Blastocystis* (Eme et al. 2017). We recovered 37 and 38 of those LGT candidates in OP10 and Opal32, respectively. Thirty genes were common in both OP10 and Opal32, and seven and eight genes were unique in OP10 and Opal32, respectively. In total, 45 different candidate LGT genes were found in the two Opalinidae species. To verify that they were not prokaryotic contaminants from the gut microbiome, we carried out two types of analyses. First, we investigated the codon usage of the coding sequences of both decontaminated transcriptomes and those of the LGT candidates and measured the frequency of optimal (F_{OP}) codons, which indicates the ratio of optimal (most frequent) codons to synonymous codons. The proportion of synonymous codons is unique to each genome and often results in a unimodal distribution of the F_{OP} score (Ikemura 1985), whereas the presence two F_{OP} peaks has been linked to contamination with bacterial sequences (Heinz et al. 2012). We obtained single-peak F_{OP} plots for our transcriptomes, indicating homogeneous codon usage and absence of contamination. All our LGT candidates fitted into these unimodal distributions supporting that they represent bona fide opalinid genes (supplementary figure S1, Supplementary Material online). Furthermore, their fit into the unimodal distribution supports an ancient integration of these LGT genes since they have adapted to the codon usage of the recipient genome. Second, we conducted phylogenetic analyses for all the LGT protein sequences. Phylogenetic trees showed that 29 of these proteins clustered robustly with their respective *Blastocystis* homologues (supplementary table S2 and supplementary figures S2-S30, Supplementary Material online). Those 29 proteins belonged to different functional families including carbohydrate metabolism, lipid metabolism, amino acid metabolism, and transporters. The phylogenetic analyses also allowed the identification of the donors of these sequences. Most of them had prokaryotic donors belonging to the Archaea, Proteobacteria and Actinobacteria, which are major components of frog gut microbiomes (Colombo et al. 2015). In some cases, the two Opalinidae species grouped with other eukaryotes belonging to the Amoebozoa, Excavata and Metazoa, suggesting eukaryote-to-eukaryote LGT, although it was impossible to infer from these trees whether the Opalinidae species were donors or recipients. Several of the LGT proteins most likely play important functions in the adaptation of Opalinidae to the anaerobic gut environment. One example is the mitochondrial iron-sulfur cluster (ISC) biogenesis system, essential for the assembly of iron-sulfur-containing proteins. These proteins are involved in a variety of metabolisms, including electron transport, nitrogen

fixation, and photosynthesis. In some protists living in low-oxygen environments, the canonical eukaryotic ISC machinery has been replaced by alternative bacterial machineries acquired via LGT, such as the nitrogen fixation (NIF) system and the bacterial sulfur mobilization (Suf) machinery. For instance, *Entamoeba histolytica* has a bacterial NIF system (van der Giezen et al. 2004), whereas *Monocercomonoides exilis*, which has completely lost mitochondria and the mitochondrial ISC pathway, contains a bacterial Suf system (Karnkowska et al. 2015). By contrast, *Blastocystis* has an archaeal-like SufC+SufB fused protein (Tsaousis et al. 2012). Similar fused *sufCB* genes related to Methanomicrobiales homologues were also identified in anaerobic flagellates such as the jakobid *Stygiella incarcerate* and the breviate *Pygsuia biforma* (Leger et al. 2016; Stairs et al. 2014). In prokaryotes, the *suf* operon is upregulated under oxidative stress (Outten et al. 2004), suggesting that the Suf machinery can be important for living in oxygen-depleted environments. We only identified an incomplete *sufB* gene in *Opalina*, which lacked a mitochondrial target signal. Similarly, SufCB is inferred to function in the cytosol in *Blastocystis*, *Pygsuia* and *Stygiella* (Tsaousis et al. 2012; Stairs et al. 2014; Leger et al. 2016). Our phylogenetic analysis showed that *Opalina* was closely related to these other anaerobic protists within a clade of Methanomicrobiales with robust support (fig. 1). These eukaryotes belong to three unrelated supergroups (*Opalina* and *Blastocystis* to SAR, *Pygsuia* to Breviatea, and *Stygiella* to Excavata). Therefore, one parsimonious explanation for this uneven distribution of SufCB is that one of these eukaryotic lineages first obtained the *sufC* and *sufB* genes from Methanomicrobiales, then both genes fused and, finally, the fused gene was transferred by eukaryote-to-eukaryote LGT to the other eukaryotic lineages. Since we only identified the *sufB* part in *Opalina*, it seems that it secondarily lost *sufC* after branching off from the lineages with fused *sufCB*. In fact, the well-supported separation of *Opalina* and *Blastocystis* in our tree (fig. 1) suggests that they have followed different evolutionary histories for the *sufCB* gene. Interestingly, the SufB and SufC proteins of *M. exilis* and *Paratrimastix pyriformis* are not related with the clade of *Opalina*, *Blastocystis*, *Pygsuia,* and *Stygiella*, indicating that they acquired these genes by independent LGT events from other prokaryotic donors. These genes were not identified in *C. marsupialis*.

In anoxic conditions, some eukaryotes use rhodoquinone instead of ubiquinone to receive electrons from NADH in the mitochondrial complex I of the electron

transport chain (ETC) and generate rhodoquinol (Castro-Guerrero et al. 2005; Sakai et al. 2012; Takamiya et al. 1999). Rhodoquinol is then reoxidized by the mitochondrial complex II catalyzing the reverse reaction as a fumarate reductase (van Hellemond and Tielens 1994; Tielens et al. 2002). This pathway helps to produce ATP and to reduce the respiratory chain without using the mitochondrial complexes III to V. The putative methyltransferase RquA is required for rhodoquinone biosynthesis (Lonjers et al. 2012) and its distribution among eukaryotes suggests that it is important for the adaptation of the mitochondrial metabolism to low-oxygen environments. In *Blastocystis,* RquA was suggested to be targeted to the MRO (Eme et al. 2017). We identified RquA homologues in both OP10 and Opal32 that also contained the predicted mitochondrial-targeting sequence. By contrast, this protein seemed to be absent in *C. marsupialis*. RquA is not very common in eukaryotes and previous phylogenetic analyses demonstrated that RquA-containing eukaryotes are scattered among prokaryotic lineages, mostly Proteobacteria. Stairs et al. (2018) proposed that LGT of *rquA* genes from bacteria to eukaryotes occurred at least twice before subsequent multiple independent LGTs among eukaryotes. Our updated RquA phylogeny (fig. 2) is consistent with this proposal. We retrieved two major clades, A and B: *Opalina* spp. branched together with *Proteromonas* and *Blastocystis* in clade A, composed mostly of alpha- and beta-proteobacteria, and several other eukaryotes (Breviata, Amoebozoa and Euglenida). Group B also contained some eukaryotes (choanoflagellates, diatoms, and ciliates) embedded among bacteria, again mostly alpha- and beta-proteobacteria. The presence of alphaproteobacteria close to the eukaryotic sequences opens the possibility of a mitochondrial origin by endosymbiotic gene transfer (EGT). Nevertheless, several observations argue against this hypothesis: (i) the eukaryotic sequences are not monophyletic, (ii) several eukaryotic sequences appear to be closer to betaproteobacteria than to alphaproteobacteria, and (iii) if *rquA* was present in the last eukaryotic common ancestor (which already had mitochondria), it must have been lost independently many times to result in its current patchy distribution. Thus, the available data so far rather support the origin of eukaryotic *rquA* by LGT from bacteria followed by subsequent LGTs among eukaryotes.

In most mitochondria, coenzyme A is transferred from acetyl-CoA to succinate by two types of acetate:succinate CoA-transferases (ASCT1B and ASCT1C). The resulting succinyl-CoA is used for ATP production by succinyl-CoA synthetase

(SCS). This ASCT/SCS system plays a crucial role in MROs of protists living in anoxic environments, such the human urogenital parasite *Trichomonas vaginalis*, for the production of ATP by substrate-level phosphorylation independent of the mitochondrial Krebs cycle (van Grinsven et al. 2008). In the case of the free-living amoeboflagellate *Naegleria gruberi*, which contains classical mitochondria and transiently experiences low-oxygen conditions, ASCT was predicted to function in mitochondria (Fritz-Laylin et al. 2010). We identified an ASCT/SCS system in our *Opalina* transcriptomes. In contrast with the *Blastocystis* ASCT, which has an MRO-targeting sequence, the *Opalina* *asct1C* and *asct1B* were incomplete ORFs and did not contain any recognizable mitochondrial targeting signal. The ASCT1C phylogenetic tree (fig. 3) recovered *Opalina* and *Blastocystis* grouped within a large clade also containing trichomonads, *Naegleria*, fungi, and dictyostelid cellular slime molds (Amoebozoa). This eukaryotic clade was closely related to Deltaproteobacteria and Firmicutes. As in the previous cases described above, this tree suggests a bacterial origin of the gene followed by eukaryote-to-eukaryote LGT.

To carry out a more comprehensive comparison of the mitochondrial metabolism of *Opalina* with that of other MRO-containing anaerobic stramenopiles (the parasitic *Blastocystis* and the free-living *C. marsupialis* (Stechmann et al. 2008; Noguchi et al. 2015)), we used BLAST to search for homologues of MRO proteins of these organisms in *Opalina*. We also manually annotated the *Opalina* mitochondrial proteins involved in major energy metabolism pathways. As shown above, *Opalina* obtained many genes for typical MRO anaerobic metabolism by LGT from either prokaryotes or other eukaryotes, but it also contains typical mitochondrial genes vertically inherited (supplementary tables S2 and S3, Supplementary Material online). *Blastocystis* spp. and *C. marsupialis* completely lack complexes III and IV, and F1Fo ATPase (complex V) (Gentekaki et al. 2017; Noguchi et al. 2015). *Opalina* possesses some genes of the tricarboxylic acid (TCA) cycle, complex I (NADH:ubiquinone oxidoreductase), and complex II (succinate dehydrogenase) of the ETC, but does not seem to encode any other recognizable canonical components such as complexes III and IV or the F1Fo ATPase (supplementary table S4, Supplementary Material online). This suggests that *Opalina* has a partial ETC that does not appear to function in energy generation. Data from *Blastocystis* and *Pygsuia* suggest that complex II functions in reverse as a fumarate reductase to regenerate the quinone pool under anaerobic conditions without using complex III, IV

and F1Fo ATPase to conduct oxidative phosphorylation. RquA, acquired by LGT in *Opalina* (see above), is the crucial enzyme for this alternative electron transport machinery. *Opalina* also possesses genes involved in classical mitochondrial activities, including transporters, fatty acid metabolism, amino acid metabolism, pyruvate metabolism, and [2Fe-2S] ferredoxin for FeS cluster assembly, some of which are lost in *Blastocystis*. (supplementary table S2, Supplementary Material online). By contrast, we did not identify in *Opalina* some essential mitochondrial proteins, such as those involved in the eukaryotic iron-sulfur cluster (ISC) synthesis system and several enzymes (pyruvate:ferredoxin oxidoreductase (PFO), [FeFe] hydrogenase (HydA), the HydA hydrogenase maturases HydE, HydF and HydG, and two subunits of the NADH:ubquinone oxidoreductase (NuoE and NuoF)) that are hallmarks of the MROs found in many anaerobic protists, including *Blastocystis and Cantina*. In those organisms, PFO oxidizes pyruvate to acetyl-CoA and CO_2. The reduced ferredoxin is reoxidized by HydA that reduces protons to H_2 gas. In *Opalina*, which lacks HydA, the pyruvate:NADP+ oxidoreductase (PNO), instead of PFO, presumably oxidizes pyruvate to acetyl-CoA and, then, acetyl-CoA can be utilized by the ASCT/SCS system to generate ATP by substrate-level phosphorylation. Since PFO and HydA are present in *Blastocystis*, we can propose two evolutionary scenarios: First, these two enzymes were present in the common ancestor of *Opalina* and *Blastocystis* and secondarily lost in the *Opalina* lineage or, second, they were obtained in *Blastocystis* independently after it diverged from the *Blastocystis-Opalina* common ancestor. As in the case of *Blastocystis* and *Cantina*, we did not identify a pyruvate carrier in *Opalina*. Glycolysis is described as a cytosolic process in eukaryotes and its product, pyruvate, is imported into the mitochondrion by the pyruvate carrier. However, the second half of glycolysis in some stramenopiles has been predicted to occur in both the cytosol and mitochondria/MRO (Abrahamian et al. 2017). Moreover, in *Blastocystis* this second half of the glycolysis is solely localized in the MRO (Rártulos et al. 2018). Similarly, we identified in *Opalina* several enzymes of the second half of the glycolsis (glyceraldehyde phosphate dehydrogenase (GAPDH), phosphoglycerate kinase (PGK), and enolase (ENO)) with mitochondria-targeting signals (supplementary table S3, Supplementary Material online). Despite these similarities and other shared key adaptations to the oxygen-depleted gut environment, *Opalina* appears to have kept a less derived version of

the mitochondrial metabolism than its sister lineage *Blastocystis* and the stramenopile relative *Cantina*.

Conclusion

Our examination of two *Opalina* transcriptomes based on sequence similarity searches and phylogenetic analyses identified 29 genes likely acquired by LGT by a common ancestor of both *Blastocystis* and Opalinidae (supplementary table S1, Supplementary Material online). Among these genes, those coding for the Suf, RquA and ASCT proteins play important roles in anaerobic metabolism in MROs. The LGTs investigated here were most likely already present in the last common ancestor of Opalinidae and *Blastocystis*. It is unclear when a common ancestor of these organisms entered the animal gut but some of the LGTs most likely facilitated the adaptation to this new oxygen-deprived environment before the divergence of these two lineages. *Blastocystis* MROs combine metabolic properties of both mitochondria and hydrogenosomes and contain PFO and [FeFe] hydrogenase as well as incomplete TCA cycle and the complexes I and II (Gentekaki et al. 2017; Stechmann et al. 2008). Although *Opalina* shares with *Blastocystis* many enzymes involved in anaerobic metabolisms acquired via LGT and both lineages have several metabolic modifications in common (incomplete TCA cycles and absence of complexes III and IV and F1Fo ATPase), our data suggest the absence of the typical hydrogenosomal enzymes PFO and [FeFe] hydrogenase. This important difference indicates that *Blastocystis* has achieved a more derived adaptation to hypoxic condition than Opalinidae. *Opalina* represents therefore an excellent model of intermediate adaptation between conventional aerobic mitochondria and derived anaerobic MROs and can help to understand the initial steps in the evolutionary path between both types of organelles.

Materials and Methods

Isolation of *Opalina* sp. Cells

For OP10 strains, the gut content of a *Xenopus tropicalis* frog was collected and resuspended in sterile PBS buffer. Eight *Opalina* cells were manually isolated under an inverted Leica DMI3000 microscope equipped with an Eppendorf TransferMan 4r micromanipulator. The cells were rinsed twice in sterile PBS and finally resuspended in 1.5 µl of sterile water. For Opal32 strain, a smear of ca. 100 µl of *Lithobates sphenocephalus* tadpole gut contents was placed onto a sterile Petri dish and 500 µl of sterile amphibian Ringer's solution (ARS: in 1 L distilled water, 6.6 g NaCl, 0.15 g KCl, 0.15 g CaCl2, and 0.2 g NaHCO3) was added to the drop of gut content. Roughly 10 µl of this solution was examined under a Zeiss AxioSkop Plus upright microscope, and cells were imaged. A single cell was manually isolated using a micropipetter and washed six times in 100 µl of fresh and sterile ARS. The cell was then transferred to a 0.5 µl to nuclease-free PCR tube and processed as below.

Opalina sp. Transcriptome Sequencing and Assembly

For Opalina sp. OP10, RNA extraction, cDNA synthesis and amplification were done using the REPLI-g WTA Single Cell kit following the manufacturer's protocol (Qiagen). The resulting cDNA was sequenced using Illumina HiSeq 2500 paired-end sequencing (2x125 bp). For Opal32, the cell was subjected to a modified version of SmartSeq-2 (Picelli et al. 2014, Kang et al. 2017) and full-length cDNA was constructed. This cDNA was then sheared using a Covaris focused-ultrasonicator (Duty% 10, Intesity 5, Burst Cycle 200, Time 30s, Frequency Sweeping Mode). This sheared cDNA was prepped using NEBnext Ultra DNA library kit for Illumina (New England Biolabs) and sequenced on an Illumina MiSeq paired-end (2x300 bp) sequencing run. For both datasets, Illumina adapters were removed using Trimmomatic v. 0.36 (Bolger et al. 2014) and paired-end sequences were assembled using Trinity v.2.2.0 (Haas et al. 2013) with default parameters. A total of 24,170 assembled transcripts were obtained from OP10 and 16,943 from Opal32.

Transcriptome Decontamination and Completeness

The decontamination of the two transcriptomes was carried out by a three-step process. First, the transcriptome sequences were subjected to two rounds of assembly, before and after bacterial sequence removal by BlobTools v0.9.19 (Laetsch et al. 2017). Second, open-reading frames were predicted and translated from the assembled transcripts using Transdecoder v2 (http:transdecoder.github.io) to produce protein sequences for OP10 and Opal32. Finally, to remove possible host sequences, the predicted protein sequences were searched by BLASTp (Camacho et al. 2009) against two predicted anuran proteomes. We used *Xenopus tropicalis* v9.1 for OP10 and, because of the lack of a proteome from the host species of Opal32 (*Lithobates sphenocephalus*) we used *Rana catesbeiana* RCv2.1, which is the closest member of the same Ranidae family with available sequence data. At the end, we obtained 8,432 and 11,480 protein sequences from OP10 and Opal32, respectively.

To assess transcriptome completeness, we used BUSCO v2.0.1 (Simão et al. 2015) on the decontaminated predicted proteins with the eukaryote_odb9 dataset of 303 near-universal single-copy orthologs. As an additional step of quality completeness comparison, we calculated the completeness value of the near-complete genome of *Blastocystis hominis* (ASM15166v1) and compared it with the opalinid data.

Codon usage for the coding sequences of both transcriptomes and their LGT candidates were measured using the index of frequency of optimal (F_{OP}) codons (Ikemura 1985). We calculated F_{OP} values using CodonW (Peden 2005) with default settings and generated F_{OP} plots using R (http://www.r-project.org).

Identification of LGT Candidates and Phylogenetic Analysis

We used the 74 LGT proteins of *Blastocystis* sp. ST1 Nand II (Eme et al. 2017) as queries to identify Opalinidae homologs using BLASTp searches (Camacho et al. 2009) with an e-value cutoff of 1e-05. 37 and 38 proteins yielded hits in the OP10 and Opal32 protein databases, respectively. Of these, 30 were found in both transcriptomes and 7 and 8 were unique to OP10 and Opal32, respectively. In total, 45 proteins were recovered from the two strains as LGT candidates. To reconstruct their phylogenies, we searched these proteins by BLASTp against the non-redundant GenBank database with an e-value cutoff of 1e-05 and maximum of 2,000

hits. To reduce the dataset size for subsequent phylogenetic analysis, hit sequences were clustered by CD-HIT (Limin et al. 2012) at 95% similarity. The resulting 45 protein sequence datasets were aligned using MAFFT v7.388 with default settings (Katoh and Stanley 2013). Ambiguously aligned sites were removed using trimAl v1.4.rev15 (Capella-Gutierrez et al. 2009) with -automated1 setting prior to phylogenetic analyses. Preliminary phylogenies were reconstructed using FastTree 2.1.7 (Price et al. 2010) and inspected manually to reduce the size of the data set by keeping only a few representatives for the prokaryotic clades distantly related to the eukaryotic sequences. We thus identified 29 proteins from the two Opalinidae strains as LGT candidates. The final datasets were aligned and trimmed as described above. Maximum likelihood phylogenetic trees for each dataset were constructed using IQ-TREE (Nguyen et al. 2015) with the best fitting model determined by applying the Bayesian Information Criterion (BIC) with the -m MFP (model selection) with default settings for each dataset. Branch supports were calculated with 1,000 ultrafast bootstrap replicates.

Protein cellular localization was predicted using TargetP 1.1 (Emanuelesson et al. 2000), MitoFates (Fukasawa et al. 2015) and TPpred 2.0 (Savojardo et al. 2014) with default settings. Homologs of mitochondrial proteins in *Opalina* sp. OP10 were searched with BLASTp using MRO sequences from two close relatives: *Blastocystis* (Stechmann et al. 2008) and *Cantina marsupialis* (Noguchi et al. 2015) (supplementary table S3, Supplementary Material online).

Data Availability

Protein sequence data sets used in this work, including complete and trimmed alignments and phylogenetic trees, are available for download at figshare (10.6084/m9.figshare.9746360). *Opalina* sequences have been submitted to GenBank (for accession numbers, see supplementary tables S2 and S3, Supplementary Material online).

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

This study was supported by European Research Council grants ProtistWorld (P.L.-G., agreement no. 322669) and Plast-Evol (D.M. agreement no. 787904), the Agence Nationale de la Recherche (D.M., project ANR-15-CE32-0003 “ANCESSTRAM”) and the Institut Diversité Ecologie et Evolution du Vivant (D.M. and N.P.).

References

Abrahamian M, Kagda M, Ah-Fong, AMV, Judelson HS 2017. Rethinking the evolution of eukaryotic metabolism: novel cellular partitioning of enzymes in stramenopiles links serine biosynthesis to glycolysis in mitochondria. *BMC Evol Biol*. 17:241.

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *J Eukaryot Microbiol*. 66(1):4-119.

Andersen RA. 2004. Biology and systematics of heterokont and haptophyte algae. *Am J Bot*. 91:1508–1522.

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

Burki F, Shalchian-Tabrizi K, Minge M, Skjæveland Å, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* 2:e790.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLASTþ: architecture and applications. *BMC Bioinformatics* 10:421.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-1973

Castro-Guerrero NA, Jasso-Chávez R, Moreno-Sánchez R. 2005. Physiological role of rhodoquinone in *Euglena gracilis* mitochondria. *Biochim Biophys Acta.* 1710:113-121.

Clark CG, van der Giezen M, Alfellani MA, Stensvold CR. 2013. Recent developments in *Blastocystis* research. *Adv Parasitol.* 82:1-32

Colombo BM, Scalvenzi T, Benlamara S, Pollet N. 2015. Microbiota and mucosal immunity in amphibians. *Front Immunol*. 6:111.

Denoeud F, Roussel M, Noel B, Wawrzyniak I, Da Silva C, Diogon M, Viscogliosi E, Brochier-Armanet C, Couloux A, Poulain J, et al. 2011. Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite. *Genome Biol*. 12:R29.

Derelle R, López-García P, Timpano H, Moreira D. 2016 A phylogenomic framework to study the diversity and evolution of stramenopiles (=heterokonts). *Mol Biol Evol*. 33:2890-2898.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*. 300:1005-1016.

Eme L, Gentekaki E, Curtis B, Archibald J, Roger A. 2017. Lateral gene transfer in the adaptation of the anaerobic parasite *Blastocystis* to the gut. *Curr Biol*. 27:807–820.

Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* 140:631–642.

Fukasawa Y, Tsuji J, Fu S, Tomii K, Horton P, Imai K. 2015. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol Cell Proteomics*.14:1113-1126.

Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Elias M, Salas-Leiva DE, Herman EK, Eme L, Arias MC, Henrissat B, et al. 2017. Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. *PLoS Biol*. 15:e2003769–42.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8:1494-1512.

Heinz E, Williams TA, Nakjang S, Noël CJ, Swan DC, Goldberg AV, Harris SR. 2012. The genome of the obligate intracellular parasite *Trachipleistophora hominis*: new insights into microsporidian genome dynamics and reductive evolution. *PLoS Pathog*. 8:e1002979.

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Micro.* 16:67–79.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2:13–34.

Kang S, Tice AK, Spiegel FW, Silberman JD, Pánek T, Čepička I, Kostka M, Kosakyan A, Alcântara DM, Roger AJ, Shadwick LL, Smirnov A, Kudryavstev A, Lahr DJG, Brown MW. 2017. Between a pod and a hard test: the deep evolution of amoebae. *Mol Biol Evol*. 34:2258-2270.

Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, Soukal P, Hroudová M, Dolezal P, Stairs CW, Roger AJ, Elias M, Dacks JB, Vlček C, and Hampl V. 2016. A eukaryote without a mitochondrial organelle. *Curr Biol* 26:1274–1284.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet.* 9:605–618.

Kostka M. 2016. Opalinidae. In: Archibald JM, Simpson AGB, Slamovits CH, Margulis L, editors. Handbook of the Protists. Springer, Charm, Switzerland.

Kostka M, Hampl V, Cepicka I, Flegr J. 2004. Phylogenetic position of *Protoopalina intestinalis* based on SSU rRNA gene sequence. *Mol Phylogenet Evol.* 33:220–224.

Laetsch DR, Blaxter ML, Leggett RM. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research.* 6:1287.

Leger MM, Eme L, Hug LA, Roger AJ. 2016. Novel hydrogenosomes in the microaerophilic jakobid *Stygiella incarcerata*. *Mol Biol Evol*. 33:2318–2336.

Leger MM, Eme L, Stairs CW, Roger AJ. 2018. Demystifying eukaryote lateral gene transfer. *Bioessays*. 40:e1700242.

Li M, Ponce-Gordo F, Grim JN, Li C, Zou H, Li W, Wu S, Wang G. 2018. Morphological redescription of *Opalina undulata* Nie 1932 from *Fejervarya limnocharis* with molecular phylogenetic study of Opalinids (Heterokonta, Opalinea). *J Eukaryot Microbiol.* 65:783-791

Limin F, Beifang N, Zhengwei Z, Sitao W, Weizhong L, 2012. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* 28:3150-3152.

Lonjers ZT, Dickson EL, Chu TPT, Kreutz JE, Neacsu FA, Anders KR, Shepherd JN. 2012. Identification of a new gene required for the biosynthesis of rhodoquinone in *Rhodospirillum rubrum*. *J Bacteriol*. 194:965–971.

Martin WF. 2017. Too much eukaryote LGT. *Bioessays*. 39:1700115.

Noguchi F, Shimamura S, Nakayama T, Yazaki E, Yabuki A, Hashimoto T, Inagaki Y, Fujikura K, Takishita K. 2015. Metabolic capacity of mitochondrion-related organelles in the free-living anaerobic stramenopile *Cantina marsupialis*. *Protist* 166:534–550.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32:268-274.

Outten FW, Djaman O, Storz G. 2004. A suf operon requirement for Fe-S cluster assembly during iron starvation in *Escherichia coli*. *Mol Microbiol.* 52:861–872.

Peden J 2005. CodonW version 1.4.2. http://codonw.sourceforge.net/

Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S,Sandberg R. 2014. Full-length RNA-Seq from single cells using Smart-seq2. *Nat. Protocols* 9:171-181.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

Patterson DJ. 1989. Stramenopiles: Chromophytes from a protistan perspective. In: Green JC, Leadbeater BSC, Diver WL, editors. The chromophyte algae: problems and perspectives. Clarendon Press, Oxford. p357–379.

Río Bártulos C, Rogers MB, Williams TA, Gentekaki E, Brinkmann H, Cerff R, Liaud M-F, Hehl AB, Yarlett NR, Gruber A, Kroth PG, van der Giezen M. 2018. Mitochondrial glycolysis in a major lineage of eukaryotes. *Genome Biol Evol*. 10:2310-2325.

Sakai C, Tomitsuka E, Esumi H, Harada S, Kita K. 2012. Mitochondrial fumarate reductase as a target of chemotherapy: from parasites to cancer cells. *Biochim Biophys Acta.* 1820:643-651.

Savojardo C, Martelli PL, Fariselli P, Casadio R. 2014. TPpred2: improving the prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs. *Bioinformatics* 30:2973-2974.

Shiratori T, Nakayama T, Ishida K-I. 2015. A new deep-branching stramenopile, *Platysulcus tardus* gen. nov., sp. nov. *Protist* 166:337–348.

Shiratori T, Thakur R, Ishida K-I. 2017. *Pseudophyllomitus vesiculosus* (Larsen and Patterson 1990) Lee, 2002, a poorly studied phagotrophic biflagellate is the first characterized member of stramenopile environmental clade MAST-6. *Protist* 168:439–451.

Silberman JD, Sogin ML, Leipe DD, Clark CG. 1996. Human parasite finds taxonomic home. *Nature* 380:398–398.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 31:3210–3212.

Stairs CW, Eme L, Brown MW, Mutsaers C, Susko E, Dellaire G, Soanes DM, van der Giezen M, Roger AJ. 2014. A SUF Fe-S cluster biogenesis system in the mitochondrion-related organelles of the anaerobic protist *Pygsuia*. *Curr Biol*. 24:1176-1186.

Stairs CW, Eme L, Muñoz-Gómez SA, Cohen A, Dellaire G, Shepherd JN, Fawcett JP, Roger AJ. 2018. Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *eLife* 7:e34292

Stechmann A, Hamblin K, Perez-Brocal V, Gaston D, Richmond GS, van der Giezen M, Clark CG, Roger AJ. 2008. Organelles in *Blastocystis* that blur the distinction between mitochondria and hydrogenosomes. *Curr Biol.* 18:580–585.

Takamiya S, Matsui T, Taka H, Murayama K, Matsuda M, Aoki T. 1999. Free-living nematodes *Caenorhabditis elegans* possess in their mitochondria an additional rhodoquinone, an essential component of the eukaryotic fumarate reductase system. *Arch Biochem Biophys*. 371:284-289.

Tan KSW. 2004. *Blastocystis* in humans and animals: new insights using modern methodologies. *Vet Parasitol*. 126:121–144.

Tielens AGM, Rotte C, van Hellemond JJ, Martin W. 2002. Mitochondria as we don't know them. *Trends Biochem Sci*. 27:564–572.

Tsaousis AD, de Choudens SO, Gentekaki E, Long S, Gaston D, Stechmann A, Vinella D, Py B, Fontecave M, Barras F. 2012. Evolution of Fe/S cluster biogenesis in the anaerobic parasite *Blastocystis*. *Proc Natl Acad Sci USA* 109:10426–10431.

van der Giezen M, Cox S, Tovar J. 2004. The iron-sulfur cluster assembly genes iscS and iscU of *Entamoeba histolytica* were acquired by horizontal gene transfer. *BMC Evol Biol.* 4:7.

van Grinsven KWA, Rosnowsky S, van Weelden SWH, Pütz S, van der Giezen M, Martin W, van Hellemond JJ, Tielens AGM, Henze K. 2008. Acetate:Succinate CoA-transferase in the hydrogenosomes of *Trichomonas vaginalis. J Biol Chem*. 283:1411–1418.

van Hellemond JJ, Tielens AGM 1994. Expression and functional properties of fumarate reductase. *Biochem J.* 304:321–331.

Yubuki N, Pánek T, Yabuki A, Cepicka I, Takishita K, Inagaki Y, Leander BS. 2015. Morphological identities of two different marine stramenopile environmental sequence clades: *Bicosoeca kenaiensis* (Hilliard, 1971) and *Cantina marsupialis* (Larsen and Patterson, 1990) gen. nov., comb. nov. *J Eukaryot Microbiol*. 62:532–542.

Figure Legends

FIG. 1. Maximum likelihood phylogenetic tree of SufCB (188 sequences). Bootstrap values <50% are not shown. For the complete tree see supplementary figure S2, Supplementary Material online.

FIG. 2. Maximum likelihood phylogenetic tree of RquA (102 sequences). Bootstrap values <50% are not shown. For the complete tree see supplementary figure S3, Supplementary Material online.

FIG. 3. Maximum likelihood phylogenetic tree of ASCT1C (96 sequences). Bootstrap values <50% are not shown. For the complete tree see supplementary figure S4, Supplementary Material online.

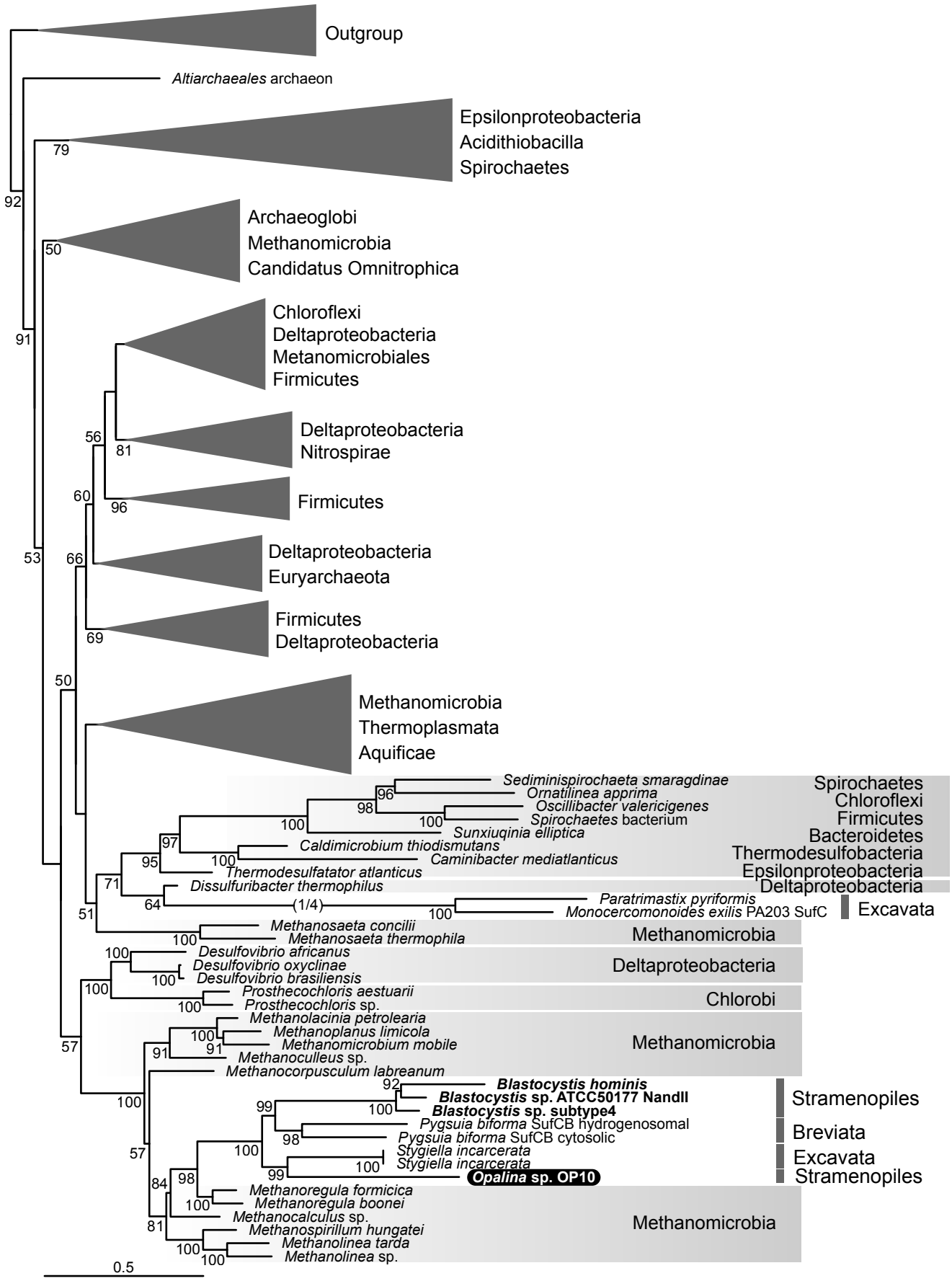

Outgroup
Altiarchaeales archaeon
Epsilonproteobacteria
Acidithiobacilla
Spirochaetes
Archaeoglobi
Methanomicrobia
Candidatus Omnitrophica
Chloroflexi
Deltaproteobacteria
Metanomicrobiales
Firmicutes
Deltaproteobacteria
Nitrospirae
Firmicutes
Deltaproteobacteria
Euryarchaeota
Firmicutes
Deltaproteobacteria
Methanomicrobia
Thermoplasmata
Aquificae
Sediminispirochaeta smaragdinae
Ornatilinea apprima
Oscillibacter valericigenes
Spirochaetes bacterium
Sunxiuqinia elliptica
Caldimicrobium thiodismutans
Caminibacter mediatlanticus
Thermodesulfatator atlanticus
Dissulfuribacter thermophilus
Paratrimastix pyriformis
Monocercomonoides exilis PA203 SufC
(1/4)
Spirochaetes
Chloroflexi
Firmicutes
Bacteroidetes
Thermodesulfobacteria
Epsilonproteobacteria
Deltaproteobacteria
Excavata
Methanosaeta concilii
Methanosaeta thermophila
Methanomicrobia
Desulfovibrio africanus
Desulfovibrio oxyclinae
Desulfovibrio brasiliensis
Deltaproteobacteria
Prosthecochloris aestuarii
Prosthecochloris sp.
Chlorobi
Methanolacinia petrolearia
Methanoplanus limicola
Methanomicrobium mobile
Methanoculleus sp.
Methanocorpusculum labreanum
Methanomicrobia
Blastocystis hominis
Blastocystis sp. ATCC50177 NandII
Blastocystis sp. subtype4
Pygsuia biforma SufCB hydrogenosomal
Pygsuia biforma SufCB cytosolic
Stygiella incarcerata
Stygiella incarcerata
Opalina sp. OP10
Stramenopiles
Breviata
Excavata
Stramenopiles
Methanoregula formicica
Methanoregula boonei
Methanocalculus sp.
Methanospirillum hungatei
Methanolinea tarda
Methanolinea sp.
Methanomicrobia
0.5

Figure 1

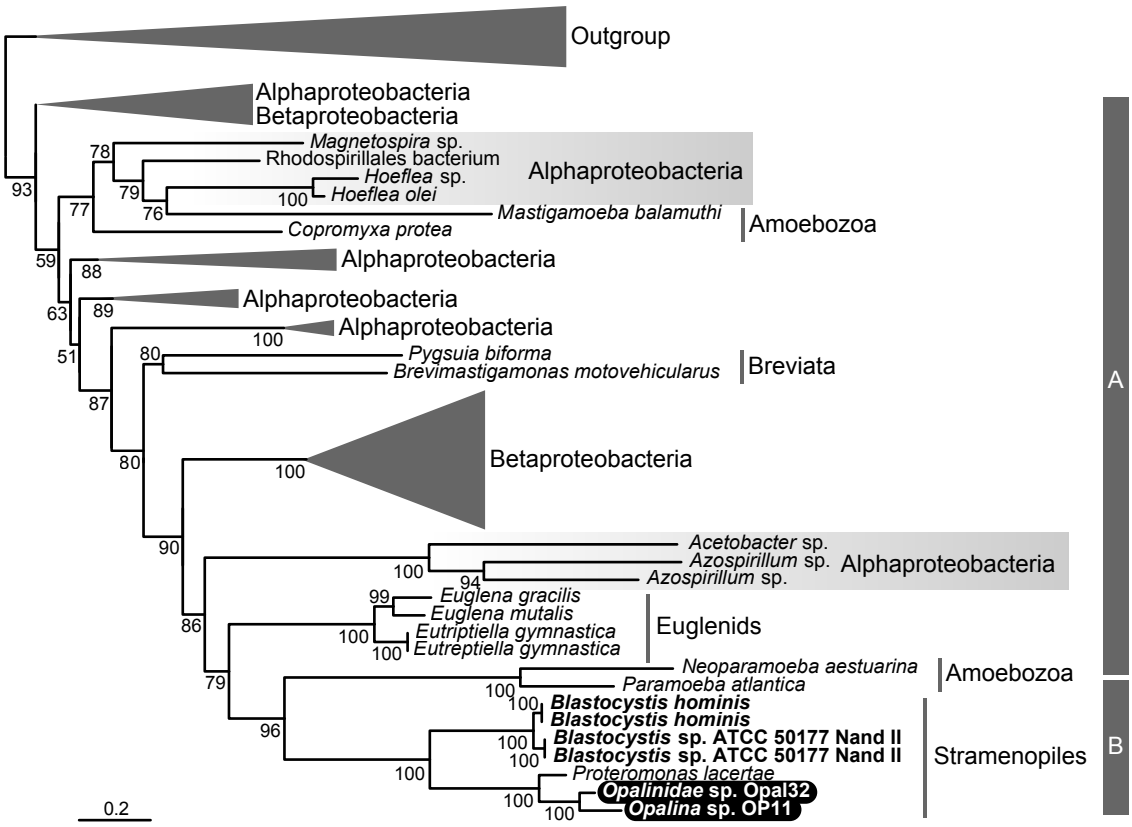

Outgroup
Alphaproteobacteria
Betaproteobacteria
Magnetospira sp.
Rhodospirillales bacterium
Hoeflea sp.
Hoeflea olei
Alphaproteobacteria
Mastigamoeba balamuthi
Copromyxa protea
Amoebozoa
Alphaproteobacteria
Alphaproteobacteria
Alphaproteobacteria
Pygsuia biforma
Brevimastigamonas motovehicularus
Breviata
Betaproteobacteria
Acetobacter sp.
Azospirillum sp.
Azospirillum sp.
Alphaproteobacteria
Euglena gracilis
Euglena mutalis
Eutriptiella gymnastica
Eutreptiella gymnastica
Euglenids
Neoparamoeba aestuarina
Paramoeba atlantica
Amoebozoa
Blastocystis hominis
Blastocystis hominis
Blastocystis sp. ATCC 50177 Nand II
Blastocystis sp. ATCC 50177 Nand II
Proteromonas lacertae
Opalinidae sp. Opal32
Opalina sp. OP11
Stramenopiles
0.2
A
B

Figure 2

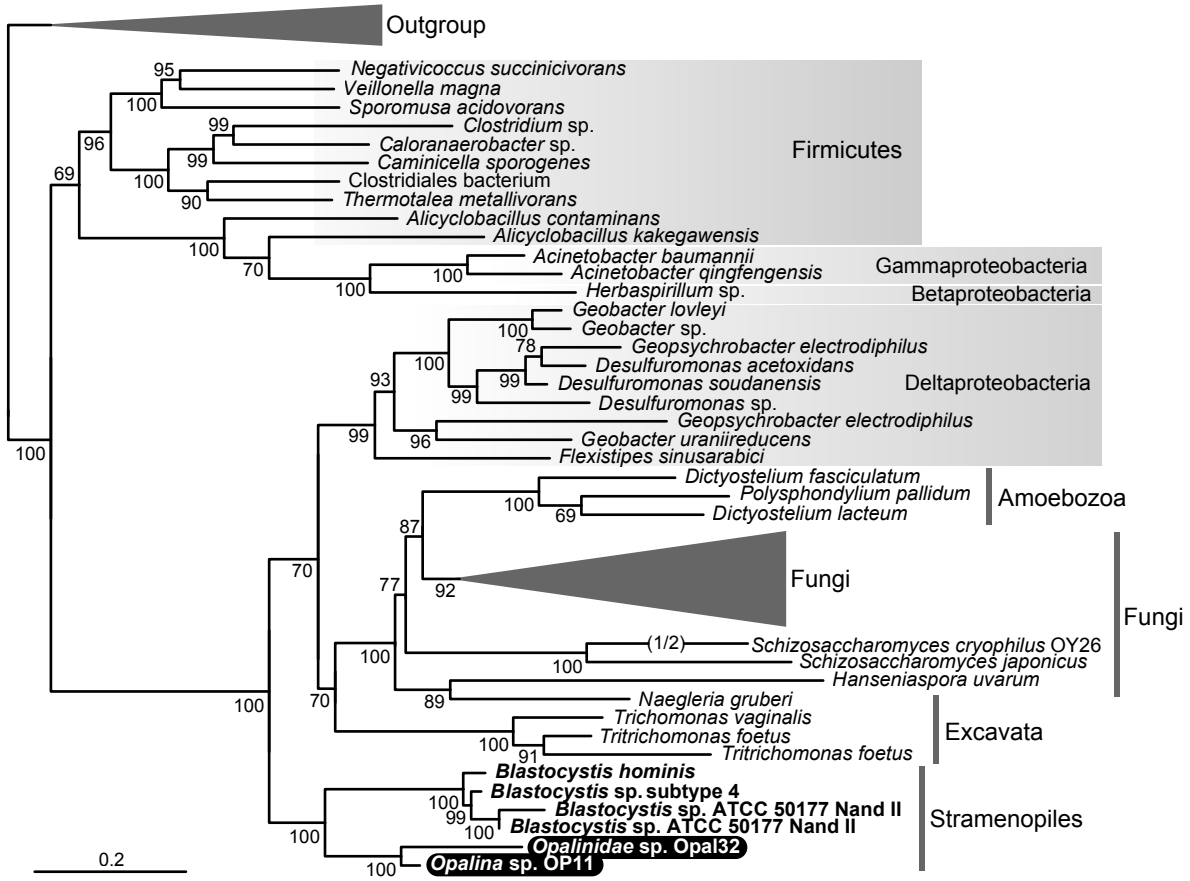

Outgroup
Negativicoccus succinicivorans
Veillonella magna
Sporomusa acidovorans
Clostridium sp.
Caloranaerobacter sp.
Caminicella sporogenes
Clostridiales bacterium
Thermotalea metallivorans
Alicyclobacillus contaminans
Alicyclobacillus kakegawensis
Firmicutes
Acinetobacter baumannii
Acinetobacter qingfengensis
Herbaspirillum sp.
Gammaproteobacteria
Betaproteobacteria
Geobacter lovleyi
Geobacter sp.
Geopsychrobacter electrodiphilus
Desulfuromonas acetoxidans
Desulfuromonas soudanensis
Desulfuromonas sp.
Geopsychrobacter electrodiphilus
Geobacter uraniireducens
Flexistipes sinusarabici
Deltaproteobacteria
Dictyostelium fasciculatum
Polysphondylium pallidum
Dictyostelium lacteum
Amoebozoa
Fungi
Fungi
(1/2)
Schizosaccharomyces cryophilus OY26
Schizosaccharomyces japonicus
Hanseniaspora uvarum
Naegleria gruberi
Trichomonas vaginalis
Tritrichomonas foetus
Tritrichomonas foetus
Excavata
Blastocystis hominis
Blastocystis sp. subtype 4
Blastocystis sp. ATCC 50177 Nand II
Blastocystis sp. ATCC 50177 Nand II
Opalinidae sp. Opal32
Opalina sp. OP11
Stramenopiles
0.2

Figure 3

APPENDIX

F

RÉSUMÉ EN FRANÇAIS

F.1 Introduction

F.1.1 Les microbes

Les micro-organismes sont des formes de vies invisibles à l'oeil nu qui ont été découvertes et étudiée grâce à l'invention et les améliorations du microscope par Robert Hooke (1635-1703), Antonie van Leeuwenhoek (1632–1723) (Caumette et al., 2015). En 1655, Robert Hooke a publié son livre *Micrographia* dans lequel il décrit ce qui peut être vu à travers la lentille de son prototype de microscope. Ses dessins de plantes et d'insectes vus au microscope étaient pionniers, éclipsés seulement par les premières esquisses de microorganismes jusqu'alors invisibles (champignons et probablement protozoaires). Hooke a également inventé le terme *cellule* en référence aux cellules d'un nid d'abeille, auxquelles les cellules végétales lui faisaient penser.

Le premier à avoir vu et décrit les organismes unicellulaires fut Antonie van Leeuwenhoek en 1674. En 1676, utilisant ingénieusement les capacités de grossissement de son prototype de microscope, il décrit pour la première fois des cellules bactériennes et estime que le volume de milliers de ces cellules serait égal à un petit grain de sable. Grâce à ces découvertes et à bien d'autres, Antonie van Leeuwenhoek est considéré comme le *père de la microbiologie*.

Jusqu'au milieu du XIXe siècle, la microbiologie n'avait pas subi de transformations majeures, principalement en raison du manque d'amélioration technique, notamment dans les outils pour observer et travailler avec le vivant microscopique. C'est grâce aux travaux de Ferdinand Cohn, Robert Koch et Luis Pasteur que les méthodes d'isolement et de culture ont été améliorées. Cela leur a permis de découvrir de nombreuses nouvelles lignées dans leurs recherches sur les bactéries pathogènes, réfutant également de manière décisive l'hypothèse de la génération spontanée qui était encore un sujet de discorde à l'époque.

F.1.2 Diversité microbienne

La vie sur Terre, sous la forme de cellules microbiennes, est apparue il y a 3,8 à 3,9 milliards d'années et les microbes ont été la forme de vie la plus abondante depuis lors. Les cellules microbiennes (protistes inclus) sont toujours des acteurs clés du fonctionnement de nos écosystèmes et cela n'est pas (plus) sujet à débat.

Les microorganismes sont principalement unicellulaires, omniprésents et s'étendent sur ce qu'on appelle l'arbre de vie. Au début des années 1900, le français Edouard Chatton decrira pour la première fois des différences majeures au sein des espèces microbienne et classifiera celles-ci en deux groupes qu'il nommera *procaryotes* et *eucaryotes* en fonction, respectivement, de l'absence ou présence d'un noyau dans la cellule observée. Ceci est confirmé plus tard par Stanier, Niel van (1962) qui décrit pour la première fois les différences moléculaires entre les virus, les bactéries et les protistes (eucaryotes microbiens).

F.1.3 Procaryotes

Les procaryotes sont des microorganismes dans lesquels toutes les réactions de la machinerie moléculaire telles que les processus de traduction et de transcription se produisent directement dans le cytoplasme, sans aucun organite comme le noyau. Les procaryotes, également appelés la "majorité invisible" (Whitman et al., 1998), sont divisés en deux domaines principaux de la vie - les archées et les bactéries, tous deux composés exclusivement d'organismes unicellulaires. Les bactéries sont connues depuis leur découverte par Antonie van Leeuwenhoek dans les années 1670, mais les Archées n'ont été découvertes que 300 ans plus tard par Carl Woese et George Fox en 1977 (Woese, Fox, 1977) grâce à une comparaison de gènes d'ARN ribosomique.

Bactéries Ces dernières années, le monde bactérien a également fait l'objet d'études de pointe. Hug et al. (2016) ont étudié les protéines conservées et ont retenu un ensemble de 16 gènes ribosomiques[1] afin de construire un alignement phylogénétique et d'inférer un arbre de la vie. Cela leur a permis de réviser considérablement l'arbre de la vie, en y ajoutant une vaste expansion soulignant la prédominance des lignées bactériennes par rapport aux lignées archées et eucaryotes (voir Figure F.1). Les séquences de génomes représentatifs isolés ou cultivés font encore défaut pour nombre de ces grandes lignées bactériennes, en particulier dans le groupe nouvellement décrit des radiations de phyla candidats (CPR (Brown et al., 2015)). Depuis, Parks et al. (2018, 2020) ont proposé une autre classification basée sur la seule phylogénie du génome, ce qui reste assez controversé à l'heure de l'écriture de cette thèse.

Archées Au cours des 5 à 10 dernières années, la phylogénie des domaines procaryotes a connu des changements importants qui ont eu un impact profond sur notre compréhension de ces domaines de la vie. Avant 2013, les archées étaient principalement divisées en 2 groupes, le superphylum TACK (Guy, Ettema, 2011) et Euryarchaeota. Puis, en 2013, Rinke et al. (2013) a créé le super-groupe DPANN regroupant de nombreuses lignées d'archées aux ramifications

[1] les gènes ribosomiques sont impliqués dans le mécanisme de traduction et sont donc de bons candidats marqueurs phylogénétiques à copie unique

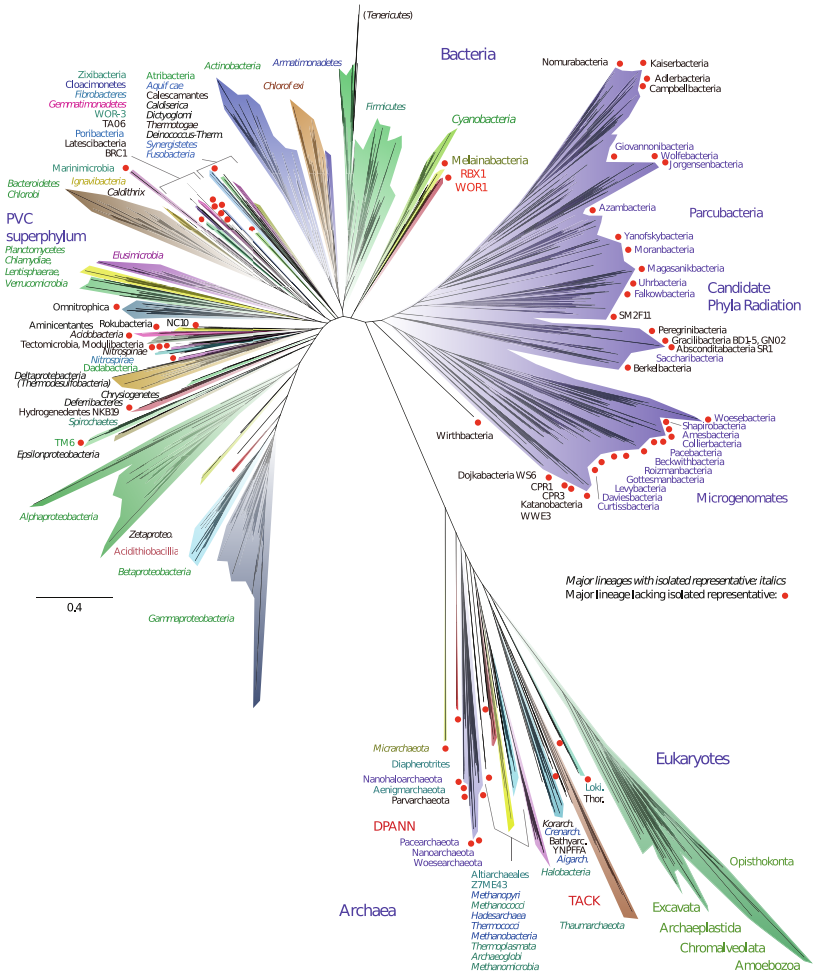
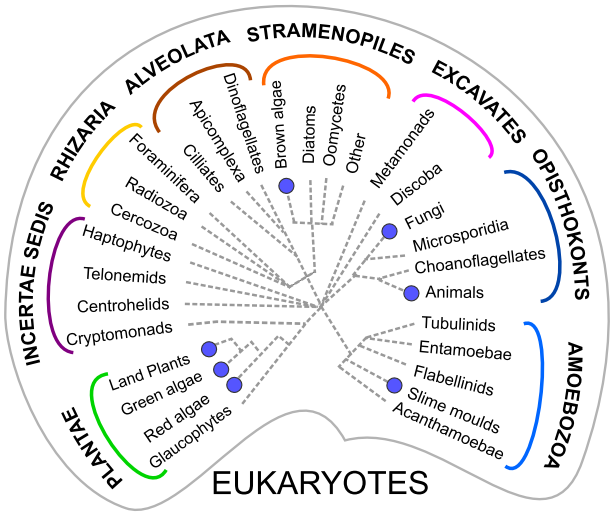


Figure F.1 – Arbre de la vie représentant 92 phyla de bactéries, 26 phyla d'archée et les 5 super-groupes eucaryotes. Cet arbre est basé sur des données de métagénomique et notamment un set de gènes ribosomaux manuellement préparé et sélectionné. (source: adapté de Hug et al. (2016))

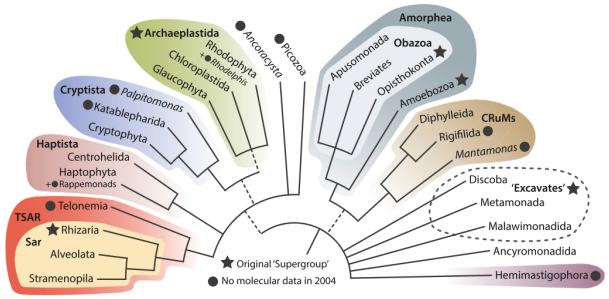
profondes qui n'appartenaient à aucun des premiers groupes primaires. En 2015, Spang et al. (2015) a décrit un nouveau phylum d'archée candidat : les Lokiarchaeota. Certaines protéines signature eucaryotes pourraient être trouvées dans les génomes de ce nouvel embranchement, le plaçant ainsi comme un groupe monophylétique à la base des eucaryotes dans l'arbre de vie. Depuis janvier 2020 et la publication par Imachi et al. (2020), il y a maintenant un représentant cultivé du super-groupe des archées Asgard, ouvrant de nouvelles possibilités pour étudier les hypothèses sur l'eucaryogenèse, l'origine des eucaryotes à partir des symbioses procaryotes comme décrites dans López-García, Moreira (2020).

F.1.4 Eucaryotes

Le domaine secondaire de la vie, Eucarya (cellules avec organelles comme le noyau) est aussi principalement composé de divers microorganismes unicellulaires appelés protistes (Figure F.2a ; Kazamia et al. (2016)) même si des espèces de Métazoaire (animaux), de plantes et de Champignons sont décrites plus en détail (Burki, 2014; Burki et al., 2019). L'arbre de vie eucaryote (eToL) reste débattu (voir Figure F.2b) car les données manquent pour certains taxons protistes sous-étudiés (Sibbald, Archibald, 2017), ce qui fait que les supergroupes de l'eToL ne bénéficient pas d'un support phylogénétique significatif (Burki et al., 2019)(voir Figure F.2b).



(a) Unicellularité chez les Eucaryotes



(b) Arbre de la vie eucaryote consensus

Figure F.2 – (a) Un diagramme schématique montrant les principaux groupes des eucaryotes et comment les organismes unicellulaires dominent cet arbre de la vie eucaryote. En 2016, la position des Haptophytes, des Télonémides, des Cryptomonades et des Centrohelides reste incertaine (Incertae sedis). Les groupes ayant de la multicellularité sont mis en évidence par des cercles remplis. (b) Consensus des arbres de la vie eucaryotes basés sur des analyses de phylogénomiques. Les couleurs représentent les *'super-groupes'* actuels. L'arbre montre les ordres de ramification non résolus et les incertitudes sur les monophylies de certains groupes par l'utilisation de multifurcations et de lignes pointillées. (source: (a) adapté de Kazamia et al. (2016); (b) adapté de Burki et al. (2019))

Les protistes sont tristement célèbres pour causer des maladies (Sibbald, Archibald, 2017) mais la grande majorité des protistes remplissent des rôles écologiques (Geisen et al., 2015). En effet, bien qu'ils reçoivent relativement peu d'attention dans le domaine de l'écologie microbienne par rapport aux procaryotes, comme le souligne le point de vue de Caron et al. (2009):

Protists are microbes too: a perspective, les protistes sont des acteurs clés dans les écosystèmes, soit en tant qu'autotrophes (producteurs primaires de molécules dans les écosystèmes ; (Field et al., 1998; Ynalvez et al., 2018)), soit en tant qu'hétérotrophes (consommateurs de molécules ou de cellules environnementales ; (Glücksman et al., 2010)), soit les deux (mixotrophes). Par exemple, les protistes aquatiques photosynthétiques *i.e.* des algues autotrophes sont responsables de la moitié du carbone fixé par photosynthèse chaque année sur Terre, comme estimé par Ynalvez et al. (2018). De plus, on a longtemps pensé que le transfert horizontal de gènes (HGT) ne jouait qu'un rôle d'adaptation des espèces procaryotes à leur environnement, mais des protistes ont été identifiés pour le faire également (Eme et al. (2017); Leger et al. (2018) et Yubuki et al. (2020), article dans **??**).

F.1.5 Ecologie microbienne

Toutes ces formes de vie n'évoluent et ne se développent pas de manière isolée : les microbes (protistes et procaryotes) coexistent dans des habitats complexes. Les premiers pas dans l'étude de ce phénomène et dans le domaine de l'écologie microbienne ont commencé avec les cinquante années de travail exceptionnel de Sergei Winogradsky à la fin du 19ème siècle (Dworkin, 2012; Caumette et al., 2015). Winogradsky a été un pionnier et, d'une certaine manière, le premier écologiste microbien en ce sens qu'il a essayé de comprendre le rôle des microorganismes dans leur environnement.

Comme nous l'avons introduit précédemment, l'écologie microbienne est la science qui étudie les microorganismes et leurs interactions avec l'environnement et entre eux. Les interactions entre les espèces microbiennes elles-mêmes ou avec les macro-organismes sont qualifiées de biotiques, tandis que les interactions avec les composants physiques et chimiques de l'habitat de la communauté microbienne sont qualifiées d'abiotiques. En termes généraux, les interactions abiotiques sont essentielles au métabolisme, à la structure cellulaire et à la physiologie d'un micro-organisme et, plus généralement, à sa survie dans l'environnement. Les interactions biotiques, d'autre part, sont les médiateurs du fonctionnement de la communauté et de l'écosystème dans son ensemble. La grande diversité des habitats microbiens se traduit par une

diversité métabolique et phylogénétique des microorganismes qui s'y trouvent. Cette diversité est l'objet d'étude de l'écologie microbienne.

F.1.6 Stratégies métaboliques

Chaque cellule vivante est constituée de 7 éléments majeurs qui sont essentiels : Le carbone (C, $\sim$50%), l'oxygène (O, $\sim$17%), l'azote (N, $\sim$13%), l'hydrogène (H, $\sim$8%), le phosphate (P, $\sim$3%), le soufre (S, $\sim$2%) et le sélénium (Se, $<$0. 01%)(pourcentages du poids sec de Madigan et al. (2015); Fagerbakke et al. (1996)). Ces éléments, ainsi que d'autres dont une cellule donnée peut avoir besoin, doivent être extraits de l'environnement, et le processus d'incorporation d'éléments extérieurs dans la cellule est appelé assimilation métabolique. En utilisant les éléments recueillis dans la nature, les cellules peuvent produire des molécules plus complexes.

Comme le carbone est la base même des molécules organiques, un écosystème dépend de ses sources de carbone (bien que dans les environnements où le carbone est omniprésent, les facteurs limitant la croissance cellulaire peuvent être d'autres nutriments tels que N et P (Elser et al., 2007)). Pour produire leur matériel cellulaire, les microorganismes peuvent obtenir du carbone à partir de sources inorganiques (autotrophes ou producteurs primaires) ou organiques (hétérotrophes) par un processus appelé assimilation du carbone.

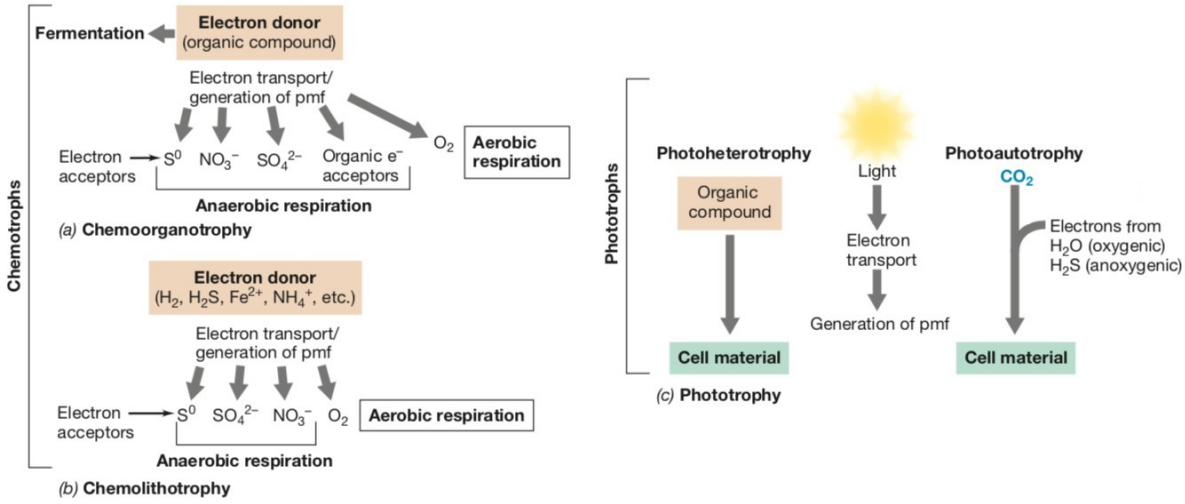


Figure F.3 – Un diagramme schématique montrant les différentes classes métaboliques que l'on trouve chez les microorganismes. (source : adapté de Madigan et al. (2015))

Quelle que soit la façon dont un micro-organisme assimile le carbone, il aura besoin d'énergie pour les processus de la machinerie cellulaire. Par conséquent, en plus de la stratégie de fixation du carbone, les organismes peuvent être phototrophes ou chimiotrophes selon la façon dont ils peuvent stocker de l'énergie. Les phototrophes sont des organismes qui convertissent la lumière en énergie stockée chimiquement. Au lieu de la lumière, les chimiotrophes tirent leur énergie de réactions chimiques par des voies cataboliques. Les chimiotrophes sont divisées en deux sous-classes selon le type de composés chimiques utilisés pour l'énergie : Les chimioorganotrophes oxydent les molécules organiques telles que les sucres, tandis que les chimiolithotrophes oxydent les substances inorganiques telles que NH_3 ou H_2S. Certaines réactions chimiques de ces cycles catabolique peuvent être spécifiques à une voie métabolique, les enzymes y participant étant exclusivement adaptées pour catalyser cette réaction exclusivement. Par conséquent, ces enzymes spécifiques peuvent être utilisées pour prédire la présence ou l'absence d'un trait métabolique spécifique dans la séquence du génome d'une espèce.

F.1.7 Le cas de la surface des sédiments

Tous les processus ci-dessus peuvent se dérouler dans des conditions aérobies (présence de O_2 dans l'environnement) ou anaérobies (absence de O_2 dans l'environnement). Les conditions aérobies sont les plus connues car les écosystèmes avec O_2 disponibles sont plus faciles d'accès et donc à étudier. En outre, la croissance dans des conditions aérobies est favorisée en présence de O_2 car le couple O_2/H_2O d'oxydation-réduction (redox) a le potentiel d'électrode standard le plus élevé et libère par conséquent une quantité considérable d'énergie. Cependant, en l'absence d'oxygène, la vie microbienne utilise d'autres couples redox pour réaliser la respiration anaérobie, même si ces derniers produisent moins d'énergie.

Les sédiments de la couche supérieure sont des écosystèmes intéressants et complexes. Dans les masses d'eau oligotrophes, les sédiments sont généralement composés de deux habitats, l'un au sommet du sédiment qui est encore aérobie ou micro-aérobie et le second, juste en dessous, dans des conditions anoxiques. De plus, les sédiments sont le réservoir de matières organiques qui s'enfoncent et le lieu de décomposition des matières organiques. Par conséquent,

le rôle des communautés microbiennes dans le cycle des nutriments implique de nombreuses classes différentes de transformation énergétique et une grande diversité de micro-organismes (Orsi, 2018).

F.1.8 La bio-informatique

Dans le cadre de cette thèse, je me suis donc intéressé à l'écologie microbienne d'écosystèmes peu explorés de manière générale que sont les écosystèmes pauvres en oxygène. Pour cela, j'ai appliqué des approches de bio-informatique afin de traiter des données de biologie moléculaire complexes obtenu par des nouvelles technologies de séquençage.

La bio-informatique est dans une certaine mesure l'héritière de la biologie computationnelle qui a débuté à la fin des années 1950 et au début des années 1960, principalement grâce aux travaux de Margaret Oakley Dayhoff et de ses collègues, pionniers dans l'alignement de séquence, les matrices de comparaisons, les bases de données de référence, etc..).

Alors que la biologie computationnelle était en plein essor, le terme "bio-informatique" a été inventé en 1970 par les scientifiques néerlandais Hesper et Hogeweg pour signifier "l'étude des processus informatiques dans les systèmes biotiques" (Hesper, Hogeweg, 1970; Hogeweg, 2011). Leur idée était d'étudier les processus de traitement, d'accumulation, de transmission et d'interprétation de l'information se produisant dans les systèmes vivants afin de mieux comprendre leur fonctionnement. Cependant, lorsque Frederick Sanger a réussi à séquencer l'ADN dans les années 70, toutes les techniques précédemment développées pour les comparaisons de peptides ont pu être appliquées aux acides nucléiques (Sanger et al., 1977). Avec ces capacités de séquençage et de comparaison d'ensembles de données en routine, ainsi que l'hypothèse de l'horloge moléculaire (Hagen, 2000, 2001), la bio-informatique a commencé à se référer au traitement de ces ensembles de données moléculaires en utilisant des outils modernes comme les ordinateurs et les bases de données de séquences disponibles pour l'alignement, la comparaison et la phylogénie des séquences.

Aujourd'hui, le terme bio-informatique désigne un domaine de recherche interdisciplinaire impliquant la biologie moléculaire théorique, le développement de méthodes et de logiciels, la biolo-

gie computationnelle, l'informatique ainsi que les mathématiques et les statistiques. En d'autres termes, la bio-informatique peut tout décrire, du calcul d'un nouveau logiciel à l'utilisation de ce logiciel pour l'interprétation biologique, en particulier sur des ensembles de données moléculaires. Même si la définition de la bio-informatique ou du bio-informaticien peut être débattue (Vincent, Charette, 2015; Smith, 2015, 2018), la plupart d'entre eux étudient des ensembles de données basés sur des macromolécules, les éléments constitutifs de la biologie moléculaire.

F.1.9 La Biologie moléculaire

Pendant 301 ans, depuis le premier signalement de cellules bactériennes au microscope par van Leeuwenhoek en 1676 et la première technique de séquençage de l'ADN par Sanger et al. (1977), les communautés microbiennes n'ont été étudiées qu'en référence à leurs caractéristiques morphologiques, aux milieux favorables à leur croissance et à d'autres caractéristiques visuelles. Grâce aux progrès de la biologie moléculaire, nous sommes maintenant en mesure de séquencer plus, plus rapidement et pour moins cher (Goodwin et al., 2016). Ces progrès nous permettent d'avancer et d'aborder de nouvelles hypothèses sur le monde microbien, son évolution, les divers rôles des microorganismes dans les écosystèmes de la Terre et leurs interactions avec les macroorganismes.

F.1.10 La Bio-informatique et les microbes

Après le premier génome entièrement séquencé (celui d'un bactériophage) (Sanger et al., 1977), le développement et la popularisation des méthodes de séquençage de l'ADN ont pris de l'ampleur, ouvrant de nouveaux points de vue sur l'écologie microbienne. Les hypothèses se sont multipliées, notamment sur le potentiel métabolique et la place écologique des microorganismes. En effet, nombre de découvertes et d'analyses en cours aujourd'hui concernent les interactions entre les microorganismes au sein des écosystèmes (transfert horizontal de gènes, symbiose, etc.) ou entre les communautés microbiennes et leur environnement ou la façon dont elles s'y adaptent (taille du génome, taille des introns, contenu en GC, etc.)

Trois générations de séquenceurs ont déjà vu le jour et sont déjà utilisé pour la recherche en Biologie en routine.

La première génération est celle du séquençage Sanger, le premier type de séquençage possible et le premier a avoir pu être modifié pour l'automatisation du processus de séquençage (Sanger et al., 1965, 1977; Hunkapiller et al., 1991). Cette grande avancée technique a permis de grande découverte et prouesse comme le séquençage des premiers génomes (Fleischmann et al., 1995; Craig Venter et al., 2001; Lander et al., 2001).

La deuxième génération de séquençage utilise la lumière dans son processus, soit par luminescence soit par fluorescence (Nyrén, Lundin, 1985) et la réaction en chaîne par polymérase (PCR) (Mullis et al., 1986; Saiki et al., 1988). D'abord il y a eu le pyroséquençage qui a eu un grand succès commercial en étant la première machine à faire du séquençage parallèle (Hyman, 1988). Depuis, la technologie Illumina a remplacé petit à petit le pyroséquençage grâce à sa capacité à produire de très grande quantité de données en peu de temps et à faible coût (Hall, 2007; Heydari et al., 2017; Escobar-Zepeda et al., 2015). Enfin, la troisème génération est celle du séquençage *d'unique molécules complètes*: *Pacific Bioscience* (Eid et al., 2009) et *Oxford Nanopore* (Mikheyev, Tin, 2014). Ces technologies ne produisent que peu de séquences mais ces séquences peuvent être très longues car elle correspondent à un morceau d'ADN entier de la cellule initial après extraction (Schadt et al., 2010; Garrido-Cardenas et al., 2017). La limitations de ces technologies reste le taux d'erreurs dans les séquences mais une nouvelle ère commence (Dijk van et al., 2018).

F.1.11 Le Métabarcoding

Le métabarcoding est une approche indépendante de la culture qui est aujourd'hui couramment utilisée pour étudier la diversité microbienne dans les écosystèmes. Elle consiste à utiliser une approche de code-barres à l'échelle d'un écosystème (**??**). Son principe est très bien résumé par la phrase suivante de Carl Woese : *To determine relationships covering the entire spectrum of extant living systems, one optimally needs a molecule of appropriately broad distribution* (Woese, Fox, 1977).

Bien sûr, la réalité est plus complexe que la théorie. Idéalement, une bonne molécule marqueur devrait être : **i)** ubiquitaire et référencée (une base de données de référence de ce fragment d'ADN avec les espèces classées identifiées) ; **ii)** peu susceptible d'être soumise à un transfert horizontal de gènes (HTG) ; **iii)** de longueur moyenne/courte pour être compatible avec la première et la deuxième génération de séquenceurs **iv)** une séquence d'ADN bien conservée en termes d'identité nucléotidique pour les espèces de niveau taxonomique choisies (afin de concevoir des amorces universelles ayant idéalement une affinité identique pour chaque espèce), mais en même temps avec une séquence d'ADN contenant une région variable afin de discriminer entre les groupes taxonomiques. La réalisation conjointe de ces exigences est une tâche difficile sur laquelle les biologistes et les phylogénéticiens travaillent depuis des décennies.

Les bases du métabarcoding ont été posées grâce au travail de Carl Woese et de ses collaborateurs. Tout d'abord, Sogin et al. (1972) a identifié les séquences d'ADN impliquées dans l'appareil de traduction comme des gènes marqueurs prometteurs. Trois ans plus tard, Woese et al. (1975) a publié une étude très importante sur l'ARNr 16S et a fait valoir que ces ARN sont directement impliqués dans la fonction ribosomique, ce qui expliquerait leur faible variabilité puisqu'ils conservent leurs fonctions primaires dans le mécanisme de traduction. L'année suivante, Woese et al. (1976) a adapté la méthode de séquençage Sanger (Sanger et al., 1965) et a séquencé jusqu'à 1500–3000 nucléotides. Grâce à cette prouesse technique, Fox et al. (1977) a fait valoir que les molécules d'ARNr 16S étaient les plus appropriées pour classer les procaryotes en comparaison avec les ARNr 5S et le 23S (Woese et al., 1976).

Un autre point important surmonté par Fox et al. (1977) a été de définir pour la première fois un pourcentage de similarité entre deux séquences d'ARNr 16S pour définir une espèce et répondre ainsi à la problématique question : quel est le seuil approprié à utiliser pour déduire une classification taxonomique lors d'une comparaison d'ARNr 16S? En comparant avec les résultats du standard de l'époque, la DDH (méthode d'hybridation ADN-ADN), Stackebrandt, Goebel (1994) a fixé le seuil d'identité de séquence pour la similarité de l'ARNr 16S à 97% pour définir les espèces puis affiné à 98,7-99 (Rosselló-Mora, Amann, 2001; Stackebrandt, Jonas, 2006; Chan et al., 2012; Kim et al., 2014; Edgar, 2018).

La première grande découverte utilisant une approche par *code-barres* a été sans conteste la découverte du troisième domaine de la vie, les archées par Woese, Fox (1977). Le métabarcoding appliqué comme technique exploratoire de l'écologie microbienne sur le terrain a été réalisé pour la première fois en 1990 (Giovannoni et al., 1990). Pour les études basées sur l'ARNr 18S, les premières applications ont suivi 10 ans plus tard (López-García et al., 2001; Moon-Van Der Staay et al., 2001) avec un impact majeur dans le domaine de la protistologie (voir Moreira, López-García (2002) pour un bilan).

F.1.12 La métagénomique: la stratégie du tout-en-un

L'approche métagénomique (également appelée "shotgun metagenomics") peut être résumée comme l'approche génomique ciblant une communauté microbienne entière, sans sélection préalable de gènes marqueurs. La première étude à démontrer le potentiel du séquençage de communautés entières sans l'utilisation de clones a été Breitbart et al. (2002). Deux ans plus tard, Venter et al. (2004) a été le premier à appliquer la métagénomique pour étudier la diversité d'un environnement. La même année, Tyson et al. (2004) a réussi à assembler deux génomes presque complets provenant d'un autre échantillon de faible complexité : les premiers génomes assemblés à partir de métagénome (MAG) d'une longe liste. En 2006, Poinar et al. (2006) a publié la première étude de métagénomique avec des séquences produites par la technologie NGS. L'apport des NGS et le développement d'outils bio-informatiques ont ouvert la voie à l'utilisation de la métagénomique pour étudier des échantillons de plus en plus complexes (Kowalchuk et al., 2007; Sleator et al., 2008; Simon, Daniel, 2009). Outre ces objectifs d'assemblage de MAGs et d'évaluation de la diversité, l'approche métagénomique peut être appliquée au profilage métabolique comme le montre Edwards et al. (2006). En outre, la métagénomique peut fournir des informations sur les principaux acteurs d'un écosystème, qui ne sont pas toujours les espèces prédominantes récupérées par des approches telles que le métabarcoding.

F.1.13 Les sites étudiés

La grotte Movile Située dans le sud-est de la Roumanie, Movile Cave a été découverte en 1986. La grotte est partiellement inondée et alimentée par de l'eau thermale sulfurée et son air est appauvri en oxygène, ce qui fait de la grotte Movile un écosystème unique (Sarbu, Lascu, 1997). De plus, la grotte a été isolée de la surface pendant près de 6 millions d'années (Lascu, 1989). Les conditions étaient propices au développement de la faune non photosynthétique et à l'adaptation des espèces, avec 30 des espèces d'invertébrés décrites (70 %) endémiques à la grotte (Sarbu et al., 1996; Fišer et al., 2015).

Le rôle des microorganismes dans la grotte a été étudié afin de démêler le système spécifique du réseau alimentaire et ses principaux acteurs. Sarbu et al. (1994) a identifié les bactéries chimioautotrophes oxydant les sulfures comme étant les principaux producteurs dans cet écosystème fermé et a ensuite montré la production autotrophe *in situ* soutenant la vie de nombreux (micro-)organismes différents (Sarbu et al., 1996). Rohwerder et al. (2003) a découvert que la majorité de l'activité métabolique a lieu sur des tapis microbiens flottant à la surface de l'eau et abritant le soufre élémentaire et les producteurs primaires (oxydants le soufre). Les auteurs ont également souligné l'importance des bactéries méthylotrophes. Les méthylotrophes et les méthanotrophes ont depuis été isolées et leur génome séquencé (Ganzert et al., 2014; Kumaresan et al., 2014). Chen et al. (2009) ont appliqué une approche de métabarcoding à leurs échantillons en utilisant de multiples gènes marqueurs : ARNr 16S bactérien et archée, RuBisCO, soxB et amoA. Leurs résultats ont confirmé la vie chimiolithoautotrophique dans la grotte et ont suggéré que l'oxydation de l'ammoniac et des nitrites pourrait jouer un rôle plus important qu'on ne le pensait initialement.

Le lac Baikal Le lac Baïkal (Figure F.4), formé il y a plus de 25 millions d'années, est le plus vieux lac de la planète. Les scientifiques étudient le lac Baïkal depuis le 18ème siècle, principalement parce que son originalité parmi les masses d'eau est captivante ; Mikhail Kozhov, Brooks (1965) l'a décrit comme : *a body of water which on the one hand can be considered as a marvellously old and complex lake, and on the other as a marvellously simplified ocean*.

(a) (b)

Figure F.4 – (a) Lac Baïkal pendant la période de glace et (b) Lac Baïkal en juillet. (source : (a) `https://news.algaeworld.org/2016/12/life-thrives-under-ice-covered-lakes/` et (b) moi-même)

Le lac Baïkal est situé en Sibérie en Russie (**??**). Attestant de son unicité géologique et biologique, le lac est inscrit au patrimoine mondial de l'UNESCO depuis 1966.

Par rapport aux autres lacs, le lac Baïkal est le plus profond ($\sim$1600m) et aussi le plus profond en moyenne ($\sim$750m) des lacs d'eau douce sur terre, suivi du lac Tanganyika en Afrique. Atteignant $\sim$1300m en dessous du niveau de la mer à son plus profond, le Baïkal est également la plus profonde dépression continentale sur terre juste avant la mer Caspienne Mikhail Kozhov, Brooks (1965); Zemskaya et al. (2020). Le lac Baïkal contient également plus de 200 km^3 d'eau, ce qui correspond à environ 20% du volume total d'eau douce non gelée de la planète (Shirstyankin et al., 2006). En outre, le lac Baïkal se classe deuxième de tous les lacs d'eau douce derrière le Tanganyika en termes de longueur (650 km) et sixième en termes de superficie totale (32 km), derrière les lacs d'Amérique du Nord et d'Afrique.

Géographiquement, le lac Baïkal est divisé en trois bassins de taille relativement similaire : les bassins nord, central et sud. Ceux-ci sont délimités par les deltas de la crête académique et de la Selenga (rivière à grand débit), respectivement (Mats, Perepelova, 2011; Touchart, 2012). En raison de sa haute latitude, le lac est gelé de janvier à mai malgré le récent changement climatique (Figure F.4a ; Hampton et al. (2008); Piccolroaz, Toffolon (2018)). Cette période de gel, associée à des vents forts, est un processus très important pour l'écosystème du lac. En effet, elle assure le renouvellement des eaux profondes et donc la stabilité de la température

froide de l'eau du lac autour de 4°C ainsi que la présence d'oxygène dans l'eau du lac à grande profondeur et la faible vitesse de sédimentation (eau oligotrophe) (Hohmann et al., 1997; Schmid et al., 2008; Moore et al., 2009; Shimaraev et al., 2011; Troitskaya et al., 2015; Klump et al., 2020). Il a été démontré que le processus de descente des eaux représente 50 % de l'intrusion d'eau froide oxygénée au fond du lac, tandis que l'autre moitié peut être expliquée par des événements printaniers (30 %) et des événements sous glace (20 %) (Tsimitri et al., 2015). La basse température de l'eau (ici due à la plongée) et la haute pression (ici due à la profondeur) sont connues pour favoriser le méthane en phase solide. Par conséquent, le lac Baïkal est jusqu'à présent le seul lac qui abrite des sites de décharge de méthane (De Batist et al., 2002; Granin et al., 2019).

Isolé depuis un certain temps, le lac Baïkal est un grand réservoir d'espèces endémiques ; en effet, 1455 des 2595 espèces décrites ($\sim$60%) sont endémiques au lac Baïkal (Yu Sherbakov, 1999). L'intérêt pour les microbes et la l'écologie microbienne s'est également manifesté très tôt, en particulier les protistes qui ont fait l'objet d'études dès le début des années 1900 (Mikhail Kozhov, Brooks, 1965).

F.2 Objectifs de la thèse

Cette thèse avait trois objectifs majeurs : i) mettre en place et tester un pipeline interne de métabarcoding, ii) appliquer ce pipeline à une étude plus complexe, les sédiments du lac Baïkal et, iii) utiliser la métagénomique pour explorer le potentiel métabolique et récupérer les MAGs dans ces environnements uniques et précédemment décrit en ii).

F.3 Résultats

F.3.1 Pipeline et applications

Tout d'abord, j'ai donc développé de toutes pièces un pipeline de métabarcoding que j'ai d'abord testé sur une étude de cas afin de caractériser les communautés protistes dans la grotte kars-

tique suboxique de Movile, en Roumanie. Cette application a fait l'objet d'un article publié (Reboul et al., 2019) qui est aussi disponible dans cette thèse (Chapter 2). Ensuite, j'ai mis le pipeline à la disposition de tous les membres de l'équipe DEEM et de ces collaborateurs, ce qui m'a permis de participer à plusieurs études publiées disponibles dans les annexes de ce manuscrit de thèse.

F.3.2 Métabarcoding des sédiments

Le second objectif était d'appliquer ce pipeline à une analyse plus complexe : les sédiments du lac Baïkal (Sibérie, Russie). Le métabarcoding a été utilisé pour décrire les microorganismes qui se développent dans la surface de ces sédiments et comparer les communautés en fonction de la profondeur (influence hydrothermale possible) et sur le transept latitudinal N-S du lac (différents apports fluviaux et géologiques). Nous avons aussi cherché des traces de micro-organismes "marins" typiques pour confirmer les découvertes récentes. Les résultats de cette études sont en cours de publication à l'écriture de ce manuscrit mais la version envoyée aux éditeurs est disponible ici Chapter 3.

F.3.3 Métagénomique des sédiments

Troisièmement, sur une sélection des sédiments les plus profonds du Baïkal, nous avons appliqué une approche métagénomique pour dépeindre les stratégies métaboliques des communautés. En effet, cette approche pourrait nous permettre de faire la lumière sur les acteurs clés de ces communautés et d'en déduire leurs fonctions métaboliques en termes de fixation du carbone et de métabolisme énergétique. En outre, cette approche pourrait permettre la reconstruction de MAGs et ainsi se concentrer sur certains acteurs clef de ces sédiments de surface.

Les données de métagénomique ont permis ici de confirmer les résultats obtenus avec le métabarcoding sur la diversité des micro-organismes présents dans nos échantillons. Aussi, grâce à des prédictions métaboliques, j'ai réussi à mettre en évidence les métabolismes présents dans ces échantillons et leurs acteurs majeurs. Ces résultats sont disponible dans une version

presque finalisée d'un manuscrit, voir Chapter 4.

De plus, j'ai pu isoler et reconstruire près de 300 génomes de plus ou moins bonnes qualités représentant les acteurs majeurs de ces sédiments. Les données disponibles de ces génomes et notamment ceux des acteurs majeurs que sont les Thaumarchaeota vont constituer un autre pan des analyses qu'il reste à conduire sur ce projet.

F.4 Discussion, conclusion et perspectives

F.4.1 Le pipeline de métabarcoding

Ce dernier, même s'il reste encore à améliorer, à néanmoins prouvé qu'il était fonctionnel et a pu être testé et vérifié par la publication de nombreux travaux scientifiques dans des journaux peer-reviewed. Quelques améliorations pourraient grandement faciliter l'utilisation par les biologistes comme l'ajout de la fonction de préparation des tableaux et des fichiers pour la soumission de séquences au format SRA du NCBI. De plus, le pipeline offre actuellement des bases de données de références relativement anciennes qu'il serait bien de mettre à jour rapidement. Enfin, l'ajout ou la prise en compte d'une technique utilisant les ASV plutôt que les OTU serait un plus.

F.4.2 Diversité microbienne des sédiments du Baïkal

Malgré les limites qu'offre une approche de métabarcoding, j'ai réalisé pendant ma thèse la première analyse de sédiments du lac Baïkal prenant en compte des facteurs géographiques comme la profondeur et la latitude. Les résultats de ces analyses sont que les sédiments du lac Baïkal arborent une communauté très diverse comprenant une large proportion d'archée en ce qui concerne les procaryotes. Ces communautés microbiennes ont été montrée stables tout au long du lac malgré les différents basins échantillonnés (gradient latitudinal) et les différentes profondeurs auxquelles les sédiments ont été prélevés. On retrouve aussi dans ces sédiments les traces d'organismes typiquement marin ce qui vient confirmer des résultats précédemment

publiés. Enfin, je montre aussi dans cette thèse que les sédiments du lac Baïkal sont différents des autres écosystèmes aquatiques retrouvés dans les bases de données publiques indépendamment de leur salinité ou de leur profondeur.

F.4.3 Métabolisme des sédiments du Baïkal

Les données métagénomiques ont permis de mettre en évidence les acteurs majeurs dans ces échantillons uniques. Ce sont les Thaumarchaeota qui semblent être les plus importantes dans ces écosystèmes notamment *via* leur rôle dans les cycles du souffre et de l'azote ainsi que dans la fixation du carbone. La reconstruction de MAGs a permis d'isoler une dizaine de génomes de bonne qualité et d'identifier potentiellement un nouveau genre de Thaumarchaeota. Une étude plus approfondie de ces génomes est en cours de réalisation afin de clarifier leurs "vrais" rôles métabolique et décrire ce nouveau genre.

APPENDIX

G

BIBLIOGRAPHY

Method of the Year 2013. jan 2014. 1.

Alawi Mashal, Schneider Beate, Kallmeyer Jens. A procedure for separate recovery of extra- and intracellular DNA from a single marine sediment sample // Journal of Microbiological Methods. sep 2014. 104. 36–42.

Allen F W. The Biochemistry of the Nucleic Acids, Purines, and Pyrimidines // Annual Review of Biochemistry. jun 1941. 10, 1. 221–244.

Amann Rudolf I, Ludwig Wolfgang, Schleifer Karl Heinz. Phylogenetic identification and in situ detection of individual microbial cells without cultivation // Microbiological Reviews. mar 1995. 59, 1. 143–169.

Anderson Marti J. A new method for non-parametric multivariate analysis of variance // Austral Ecology. feb 2001. 26, 1. 32–46.

Annenkova N. V., Belykh O. I., Denikina N. N., Belikov S. I. Identification of dinoflagellates from the lake baikal on the basis of molecular genetic data // Doklady Biological Sciences. jun 2009. 426, 1. 253–256.

Annenkova Natalia V, Lavrov Dennis V, Belikov Sergey I. Dinoflagellates Associated with Freshwater Sponges from the Ancient Lake Baikal // Protist. 2011. 162, 2. 222–236.

Annenkova Nataliia V., Giner Caterina R., Logares Ramiro. Tracing the Origin of Planktonic Protists in an Ancient Lake // Microorganisms. apr 2020. 8, 4. 543.

Assié Adrien, Leisch Nikolaus, Meier Dimitri V., Gruber-Vodicka Harald, Tegetmeyer Halina E., Meyerdierks Anke, Kleiner Manuel, Hinzke Tjorven, Joye Samantha, Saxton Matthew, Dubilier Nicole, Petersen Jillian M. Horizontal acquisition of a patchwork Calvin cycle by symbiotic and free-living Campylobacterota (formerly Epsilonproteobacteria) // ISME Journal. jan 2020. 14, 1. 104–122.

Bahram Mohammad, Anslan Sten, Hildebrand Falk, Bork Peer, Tedersoo Leho. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment // Environmental Microbiology Reports. aug 2019. 11, 4. 487–494.

Baker Brett J, Lazar Cassandre Sara, Teske Andreas P, Dick Gregory J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria // Microbiome. dec 2015. 3, 1. 14.

Baker Brett J., Saw Jimmy H., Lind Anders E., Lazar Cassandre Sara, Hinrichs Kai Uwe, Teske Andreas P., Ettema Thijs J.G. Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea // Nature Microbiology. 2016. 1, 3. 1–7.

Baltar Federico, Herndl Gerhard J. Ideas and perspectives: Is dark carbon fixation relevant for oceanic primary production estimates? // Biogeosciences. oct 2019. 16, 19. 3793–3799.

Barker Winona C, George David G, Hunt Lois T, Garavelli John S. The pir protein sequence database // Nucleic Acids Research. apr 1991. 19, Suppl. 2231–2236.

Bashenkhaeva Maria V., Galachyants Yuri P., Khanaev Igor V., Sakirko Maria V., Petrova Darya P., Likhoshway Yelena V., Zakharova Yulia R. Comparative analysis of free-living and particle-associated bacterial communities of Lake Baikal during the ice-covered period // Journal of Great Lakes Research. apr 2020.

Becraft Eric D., Woyke Tanja, Jarett Jessica, Ivanova Natalia, Godoy-Vitorino Filipa, Poulton Nicole, Brown Julia M., Brown Joseph, Lau M. C.Y., Onstott Tullis, Eisen Jonathan A., Moser Duane, Stepanauskas Ramunas. Rokubacteria: Genomic giants among the uncultured bacterial phyla // Frontiers in Microbiology. nov 2017. 8, NOV. 2264.

Belilla Jodie, Moreira David, Jardillier Ludwig, Reboul Guillaume, Benzerara Karim, López-García José M., Bertolino Paola, López-Archilla Ana I., López-García Purificación. Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area // Nature Ecology and Evolution. nov 2019. 3, 11. 1552–1561.

Bel'kova N L, Denisova L Ya, Manakova E N, Zaichikov E. F., Grachev M A. Species diversity of deep-water baikalian bacteria as revealed by 16S rRNA sequencing // Doklady Akademii Nauk. jun 1996. 348, 5. 692–695.

Bel'kova N. L., Parfenova V. V., Kostornova T. Ya., Denisova L. Ya., Zaichikov E. F. Microbial biodiversity in the water of lake baikal // Microbiology. 2003. 72, 2. 203–213.

Benítez-Páez Alfonso, Portune Kevin J., Sanz Yolanda. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION^{TM} portable nanopore sequencer // GigaScience. dec 2016. 5, 1. 4.

Berg Ivan A. Ecological aspects of the distribution of different autotrophic CO2 fixation pathways. mar 2011. 1925–1936.

Bernfield Merton R., Nirenberg Marshall W. RNA codewords and protein synthesis // Science. jan 1965. 147, 3657. 479–484.

Biddle Jennifer F., Fitz-Gibbon Sorel, Schuster Stephan C., Brenchley Jean E., House Christopher H. Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment // Proceedings of the National Academy of Sciences of the United States of America. jul 2008. 105, 30. 10583–10588.

Binga Erik K, Lasken Roger S, Neufeld Josh D. Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology // ISME Journal. mar 2008. 2, 3. 233–241.

Blazewicz Steven J, Barnard Romain L, Daly Rebecca A, Firestone Mary K. Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses // The ISME Journal. nov 2013. 7, 11. 2061–2068.

Bolyen Evan, Rideout Jai Ram, Dillon Matthew R, Bokulich Nicholas A, Abnet Christian C, Al-Ghalith Gabriel A, Alexander Harriet, Alm Eric J, Arumugam Manimozhiyan, Asnicar Francesco, Bai Yang, Bisanz Jordan E, Bittinger Kyle, Brejnrod Asker, Brislawn Colin J, Brown C Titus, Callahan Benjamin J, Caraballo-Rodríguez Andrés Mauricio, Chase John, Cope Emily K, Da Silva Ricardo, Diener Christian, Dorrestein Pieter C, Douglas Gavin M, Durall Daniel M, Duvallet Claire, Edwardson Christian F, Ernst Madeleine, Estaki Mehrbod, Fouquier Jennifer, Gauglitz Julia M, Gibbons Sean M, Gibson Deanna L, Gonzalez Antonio, Gorlick Kestrel, Guo Jiarong, Hillmann Benjamin, Holmes Susan, Holste Hannes, Huttenhower Curtis, Huttley Gavin A, Janssen Stefan, Jarmusch Alan K, Jiang Lingjing, Kaehler Benjamin D, Kang Kyo Bin, Keefe Christopher R, Keim Paul, Kelley Scott T, Knights Dan, Koester Irina, Kosciolek Tomasz, Kreps Jorden, Langille Morgan G.I., Lee Joslynn, Ley Ruth, Liu Yong Xin, Loftfield Erikka, Lozupone Catherine, Maher Massoud, Marotz Clarisse, Martin Bryan D, McDonald Daniel, McIver Lauren J, Melnik Alexey V, Metcalf Jessica L, Morgan Sydney C, Morton Jamie T, Naimey Ahmad Turan, Navas-Molina Jose A, Nothias Louis Felix, Orchanian Stephanie B, Pearson Talima, Peoples Samuel L, Petras Daniel, Preuss Mary Lai, Pruesse Elmar, Rasmussen Lasse Buur, Rivers Adam, Robeson Michael S, Rosenthal Patrick, Segata Nicola, Shaffer Michael, Shiffer Arron, Sinha Rashmi, Song Se Jin, Spear John R,

Swafford Austin D, Thompson Luke R, Torres Pedro J, Trinh Pauline, Tripathi Anupriya, Turnbaugh Peter J, Ul-Hasan Sabah, Hooft Justin J.J. van der, Vargas Fernando, Vázquez-Baeza Yoshiki, Vogtmann Emily, Hippel Max von, Walters William, Wan Yunhu, Wang Mingxun, Warren Jonathan, Weber Kyle C, Williamson Charles H.D., Willis Amy D, Xu Zhenjiang Zech, Zaneveld Jesse R, Zhang Yilong, Zhu Qiyun, Knight Rob, Caporaso J. Gregory. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. 2019. 852–857.

Borin Sara, Crotti Elena, Mapelli Francesca, Tamagnini Isabella, Corselli Cesare, Daffonchio Daniele. DNA is preserved and maintains transforming potential after contact with brines of the deep anoxic hypersaline lakes of the Eastern Mediterranean Sea // Saline Systems. aug 2008. 4, 1. 10.

Boughner Lisa A., Singh Pallavi. Microbial Ecology: Where are we now? // Postdoc Journal. 2016. 4, 11. 3.

Bower Susan M., Carnegie Ryan B., Goh Benjamin, Jones Sevion R.M., Lowe Geoffrey J., Mak Michelle W.S. Preferential PCR amplification of parasitic protistan small subunit rDNA from metazoan tissues // Journal of Eukaryotic Microbiology. may 2004. 51, 3. 325–332.

Bray J. Roger, Curtis J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin // Ecological Monographs. feb 1957. 27, 4. 325–349.

Brehm-Stecher Byron F, Johnson Eric A. Single-Cell Microbiology: Tools, Technologies, and Applications // Microbiology and Molecular Biology Reviews. sep 2004. 68, 3. 538–559.

Breitbart Mya, Salamon Peter, Andresen Bjarne, Mahaffy Joseph M, Segall Anca M, Mead David, Azam Farooq, Rohwer Forest. Genomic analysis of uncultured marine viral communities. // Proceedings of the National Academy of Sciences of the United States of America. oct 2002. 99, 22. 14250–5.

Breitwieser Florian P, Lu Jennifer, Salzberg Steven L. A review of methods and databases for metagenomic classification and assembly // Briefings in Bioinformatics. 2018. 20, 4. 1125–1139.

Brown Christopher T., Hug Laura A., Thomas Brian C., Sharon Itai, Castelle Cindy J., Singh Andrea, Wilkins Michael J., Wrighton Kelly C., Williams Kenneth H., Banfield Jillian F. Unusual biology across a group comprising more than 15% of domain Bacteria // Nature. jul 2015. 523, 7559. 208–211.

Browne Patrick Denis, Nielsen Tue Kjærgaard, Kot Witold, Aggerholm Anni, Gilbert M Thomas P, Puetz Lara, Rasmussen Morten, Zervas Athanasios, Hansen Lars Hestbjerg. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms // GigaScience. feb 2020. 9, 2.

Brownlee G. G., Sanger F., Barrell B. G. Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli [16]. aug 1967. 735–736.

Bukin S. V., Pavlova O. N., Kalmychkov G. V., Ivanov V. G., Pogodaeva T. V., Galach'yants Yu. P., Bukin Yu. S., Khabuev A. V., Zemskaya T. I. Substrate Specificity of Methanogenic Communities from Lake Baikal Bottom Sediments Associated with Hydrocarbon Gas Discharge // Microbiology (Russian Federation). jul 2018. 87, 4. 549–558.

Bukin Sergei V, Pavlova Olga N, Manakov Andrei Y, Kostyreva Elena A, Chernitsyna Svetlana M, Mamaeva Elena V, Pogodaeva Tatyana V, Zemskaya Tamara I. The ability of microbial community of Lake Baikal bottom sediments associated with gas discharge to carry out the transformation of organic matter under thermobaric conditions // Frontiers in Microbiology. 2016. 7, MAY. 690.

Burki Fabien. The eukaryotic tree of life from a global phylogenomic perspective // Cold Spring Harbor Perspectives in Biology. 2014. 6, 5.

Burki Fabien, Roger Andrew, Brown Matthew W., Simpson Alastair G. B. The New Tree of Eukaryotes // Trends in Ecology & Evolution2. oct 2019. In press, 0.

Butina Tatyana V, Bukin Yurij S, Krasnopeev Andrey S, Belykh Olga I, Tupikin Aleksey E, Kabilov Marsel R, Sakirko V., Belikov Sergey I. Estimate of the diversity of viral and bacterial

assemblage in the coastal water of Lake Baikal // FEMS Microbiology Letters. may 2019. 366, 9.

Butina Tatyana V., Khanaev Igor V., Kravtsova Lyubov S., Maikova Olga O., Bukin Yurij S. Metavirome datasets from two endemic Baikal sponges Baikalospongia bacillifera // Data in Brief. apr 2020. 29. 105260.

Butina Tatyana Vladimirovna, Belykh Olga I., Maksimenko Svetlana Yu., Belikov Sergey I. Phylogenetic diversity of T4-like bacteriophages in Lake Baikal, East Siberia // FEMS Microbiology Letters. jun 2010. 309, 2. 122–129.

Buttigieg Pier Luigi, Ramette Alban. A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses // FEMS Microbiology Ecology. dec 2014. 90, 3. 543–550.

Cabello-Yeves Pedro J., Rodriguez-Valera Francisco. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome // Microbiome. dec 2019. 7, 1. 117.

Cabello-Yeves Pedro J., Zemskaya Tamara I., Rosselli Riccardo, Coutinho Felipe H., Zakharenko Alexandra S., Blinov Vadim V., Rodriguez-Valera Francisco, Zemskay Tamara I., Rosselli Riccardo, Coutinho Felipe H., Zakharenko Alexandra S., Blinov Vadim V., Rodriguez-Valera Francisco. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal // Applied and Environmental Microbiology. jan 2017. 84, 1. AEM.02132–17.

Cabello-Yeves Pedro J., Zemskaya Tamara I., Zakharenko Alexandra S., Sakirko Mariya V., Ivanov Vyacheslav G., Ghai Rohit, Rodriguez-Valera Francisco. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat // Limnology and Oceanography. dec 2019. lno.11401.

Callahan Benjamin J, McMurdie Paul J, Rosen Michael J, Han Andrew W, Johnson Amy Jo A, Holmes Susan P. DADA2: High-resolution sample inference from Illumina amplicon data // Nature Methods. 2016. 13, 7. 581–583.

Callahan Benjamin J, Wong Joan, Heiner Cheryl, Oh Steve, Theriot Casey M, Gulati Ajay S, McGill Sarah K, Dougherty Michael K. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution // Nucleic acids research. oct 2019. 47, 18. e103.

Campo Javier del, Kolisko Martin, Boscaro Vittorio, Santoferrara Luciana F., Nenarokov Serafim, Massana Ramon, Guillou Laure, Simpson Alastair, Berney Cedric, Vargas Colomban de, Brown Matthew W., Keeling Patrick J., Wegener Parfrey Laura. EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution // PLoS Biology. sep 2018. 16, 9. e2005849.

Cangelosi Gerard A, Meschke John S. Dead or alive: Molecular assessment of microbial viability. oct 2014. 5884–5891.

Caporaso J Gregory, Kuczynski Justin, Stombaugh Jesse, Bittinger Kyle, Bushman Frederic D, Costello Elizabeth K, Fierer Noah, Pea Antonio Gonzalez, Goodrich Julia K, Gordon Jeffrey I, Huttley Gavin A, Kelley Scott T, Knights Dan, Koenig Jeremy E, Ley Ruth E, Lozupone Catherine A, McDonald Daniel, Muegge Brian D, Pirrung Meg, Reeder Jens, Sevinsky Joel R, Turnbaugh Peter J, Walters William A, Widmann Jeremy, Yatsunenko Tanya, Zaneveld Jesse, Knight Rob. QIIME allows analysis of high-throughput community sequencing data. may 2010. 335–336.

Caron David A, Worden Alexandra Z, Countway Peter D, Demir Elif, Heidelberg Karla B. Protists are microbes too: A perspective // ISME Journal. jan 2009. 3, 1. 4–12.

Castelle Cindy J, Wrighton Kelly C, Thomas Brian C, Hug Laura A, Brown Christopher T, Wilkins Michael J, Frischkorn Kyle R, Tringe Susannah G, Singh Andrea, Markillie Lye Meng, Taylor Ronald C, Williams Kenneth H, Banfield Jillian F. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling // Current Biology. mar 2015. 25, 6. 690–701.

Caumette Pierre, Bertrand Jean Claude, Normand Philippe. Some historical elements of microbial ecology // Environmental Microbiology: Fundamentals and Applications. Dordrecht: Springer Netherlands, 2015. 9–24.

Chan Jacqueline Z-M, Halachev Mihail R, Loman Nicholas J, Constantinidou Chrystala, Pallen Mark J. Defining bacterial species in the genomic era: Insights from the genus Acinetobacter // BMC Microbiology. dec 2012. 12. 302.

Chen Yin, Wu Liqin, Boden Rich, Hillebrand Alexandra, Kumaresan Deepak, Moussard Hélène, Baciu Mihai, Lu Yahai, Murrell J. Colin. Life without light: Microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile Cave // ISME Journal. 2009. 3, 9. 1093–1104.

Chernitsyna C. M., Shubenkova O. V., Zemskaya T. I., Grachev M. A., Vereshchagin A. L., Kostornova T. Ya. Isolation of total bacterial DNA for ecological characterization of bottom sediments of Lake Baikal // Contemporary Problems of Ecology. 2008. 1, 1. 115–119.

Chernitsyna S. M., Mamaeva E. V., Lomakina A. V., Pogodaeva T. V., Galach'yants Yu. P., Bukin S. V., Pimenov N. V., Khlystov O. M., Zemskaya T. I. Phylogenetic diversity of microbial communities of the Posolsk Bank bottom sediments, Lake Baikal // Microbiology (Russian Federation). nov 2016. 85, 6. 672–680.

Clarke K. R. Non-parametric multivariate analyses of changes in community structure // Australian Journal of Ecology. mar 1993. 18, 1. 117–143.

Collins R. Eric, Deming Jody W. Abundant dissolved genetic material in Arctic sea ice Part I: Extracellular DNA // Polar Biology. dec 2011. 34, 12. 1819–1830.

Coutinho Felipe Hernandes, Cabello-Yeves Pedro J, Gonzalez-Serrano Rafael, Rosselli Riccardo, Lopez-Perez Mario, Zemskaya Tamara, Zakharenko Alexandra, Ivanov Vyacheslav, Rodriguez-Valera Francisco. New Viral Biogeochemical Roles Revealed Through Metagenomic Analysis of Lake Baikal // bioRxiv. apr 2020. 2020.04.02.019802.

Craig Venter J., Adams M D, Myers E W, Li P. W., Mural R J, Sutton G G, Smith H. O., Yandell M, Evans C A, Holt R A, Gocayne J D, Amanatides P, Ballew R M, Huson D H, Wortman J R, Zhang Q., Kodira C D, Zheng X. H., Chen L, Skupski M, Subramanian G, Thomas P. D., Zhang J., Gabor Miklos G L, Nelson C., Broder S, Clark A G, Nadeau J, McKusick V A, Zinder N, Levine A J, Roberts R J, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco V. di, Dunn P, Eilbeck K, Evangelista C, Gabrielian A E, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman T J, Higgins M E, Ji R R, Ke Z, Ketchum K A, Lai Z, Lei Y, Li Z., Li J, Liang Y, Lin X, Lu F, Merkulov G V, Milshina N, Moore H M, Naik A K, Narayan V A, Neelam B, Nusskern D, Rusch D B, Salzberg S, Shao W, Shue B, Sun J, Yuan Wang Z., Wang A., Wang X., Wang J., Wei M. H., Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W., Zhang H, Zhao Q., Zheng L, Zhong F., Zhong W, Zhu S. C., Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Lai Cheng M., Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J., Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M., Moy L, Murphy B., Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Yu H, Romblad D, Ruhfel B, Scott R., Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Ni Tint N., Tse S, Vech C, Wang G., Wetter J, Williams S., Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J., Zaveri K, Abril J F, Guigo R., Campbell M J, Sjolander K V, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang Y H, Coyne M, Dahlke C, Deslattes Mays A., Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover

J, Jennings D, Jordan C., Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T., Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D., Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome // Science. feb 2001. 291, 5507. 1304–1351.

Crick F. H.C., Barnett Leslie, Brenner S., Watts-Tobin R. J. General nature of the genetic code for proteins // Nature. dec 1961. 192, 4809. 1227–1232.

Crick Francis. Central dogma of molecular biology // Nature. aug 1970. 227, 5258. 561–563.

D'Amore Rosalinda, Ijaz Umer Zeeshan, Schirmer Melanie, Kenny John G., Gregory Richard, Darby Alistair C., Shakya Migun, Podar Mircea, Quince Christopher, Hall Neil. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling // BMC Genomics. dec 2016. 17, 1. 55.

Da Cunha Violette, Gaia Morgan, Gadelle Daniele, Nasir Arshan, Forterre Patrick. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes // PLoS Genetics. jun 2017. 13, 6. e1006810.

Da Cunha Violette, Gaia Morgan, Nasir Arshan, Forterre Patrick. Asgard archaea do not close the debate about the universal tree of life topology. mar 2018. e1007215.

Dahm Ralf. Friedrich Miescher and the discovery of DNA. feb 2005. 274–288.

Daly Aisling J., Baetens Jan M., De Baets Bernard. Ecological diversity: Measuring the unmeasurable. jul 2018. 119.

David Gwendoline M., Moreira David, Reboul Guillaume, Annenkova Nataliia V., Galindo Luis J., Bertolino Paola, López-Archilla Ana I., Jardillier Ludwig, López-García Purificación. Environmental drivers of plankton protist communities along latitudinal and vertical gradients in the oldest and deepest freshwater lake // bioRxiv. sep 2020. 2020.09.26.308536.

Dayhoff M. O. Computer aids to protein sequence determination // Journal of Theoretical Biology. jan 1965. 8, 1. 97–112.

Dayhoff M. O., Schwartz R. M., Chen H. R., Barker W. C., Hunt L. T., Orcutt B. C. Nucleic Acid Sequence Database // DNA. jan 1981. 1, 1. 51–58.

Dayhoff Margaret Oakley, Ledley Robert S. Comprotein: A computer program to aid primary protein structure determination // AFIPS Conference Proceedings - 1962 Fall Joint Computer Conference, AFIPS 1962. New York, New York, USA: ACM Press, 1962. 262–274.

De Batist Marc, Klerkx Jan, Van Rensbergen Pieter, Vanneste Maarten, Poort Jeffrey, Golmshtok Alexander Y., Kremlev Andrei A., Khlystov Oleg M., Krinitsky Petr. Active hydrate destabilization in Lake Baikal, Siberia? // Terra Nova. 2002. 14, 6. 436–442.

DeSantis T Z, Hugenholtz P, Larsen N, Rojas M, Brodie E L, Keller K, Huber T, Dalevi D, Hu P, Andersen G L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB // Applied and Environmental Microbiology. jul 2006. 72, 7. 5069–5072.

Dell'Anno Antonio, Danovaro Roberto. Ecology: Extracellular DNA plays a key role in deep-sea ecosystem functioning // Science. sep 2005. 309, 5744. 2179.

Denisova L Ya, Bel'kova N L, Tuiokhonov I. I., Zaichikov E. F. Bacterial diversity at various depths in the southern part of lake baikal as revealed by 16S rRNA sequencing // Mikrobiologiya. 1999. 68, 4. 547–556.

Dijk Erwin L van, Jaszczyszyn Yan, Naquin Delphine, Thermes Claude. The Third Revolution in Sequencing Technology // Trends in Genetics. sep 2018. 34, 9. 666–681.

Dombrowski Nina, Lee Jun Hoe, Williams Tom A, Offre Pierre, Spang Anja. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea // FEMS Microbiology Letters. jan 2019. 366, 2.

Dominiak Dominik Marek, Nielsen Jeppe Lund, Nielsen Per Halkjær. Extracellular DNA is abundant and important for microcolony strength in mixed microbial biofilms // Environmental Microbiology. mar 2011. 13, 3. 710–721.

Dopheide Andrew, Xie Dong, Buckley Thomas R., Drummond Alexei J., Newcomb Richard D. Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity // Methods in Ecology and Evolution. jan 2019. 10, 1. 120–133.

Dorrell Richard G, Azuma Tomonori, Nomura Mami, Kerdrel Guillemette Audren de, Paoli Lucas, Yang Shanshan, Bowler Chris, Ishii Ken ichiro, Miyashita Hideaki, Gile Gillian H, Kamikawa Ryoma. Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes // Proceedings of the National Academy of Sciences of the United States of America. apr 2019. 116, 14. 6914–6923.

DuBuy B., Weissman S. M. Nucleotide sequence of Pseudomonas fluorescens 5 S ribonucleic acid. // Journal of Biological Chemistry. feb 1971. 246, 3. 747–761.

Dworkin Martin. Sergei Winogradsky: A founder of modern microbiology and the first microbial ecologist. mar 2012. 364–379.

Earl Joshua P., Adappa Nithin D., Krol Jaroslaw, Bhat Archana S., Balashov Sergey, Ehrlich Rachel L., Palmer James N., Workman Alan D., Blasetti Mariel, Sen Bhaswati, Hammond Jocelyn, Cohen Noam A., Ehrlich Garth D., Mell Joshua Chang. Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0605 Microbiology // Microbiome. dec 2018. 6, 1. 190.

Eberwine James, Sul Jai Yoon, Bartfai Tamas, Kim Junhyong. The promise of single-cell sequencing. jan 2014. 25–27.

Edgar Robert C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs // Bioinformatics. jul 2018. 34, 14. 2371–2375.

Edwards Robert A, Rodriguez-Brito Beltran, Wegley Linda, Haynes Matthew, Breitbart Mya, Peterson Dean M, Saar Martin O, Alexander Scott, Alexander E Calvin, Rohwer Forest. Using pyrosequencing to shed light on deep mine microbial ecology // BMC Genomics. mar 2006. 7. 57.

Eichorst Stephanie A, Trojan Daniela, Roux Simon, Herbold Craig, Rattei Thomas, Woebken Dagmar. Genomic insights into the Acidobacteria reveal strategies for their success in terrestrial environments // Environmental Microbiology. 2018. 20, 3. 1041–1063.

Eid John, Fehr Adrian, Gray Jeremy, Luong Khai, Lyle John, Otto Geoff, Peluso Paul, Rank David, Baybayan Primo, Bettman Brad, Bibillo Arkadiusz, Bjornson Keith, Chaudhuri Bidhan, Christians Frederick, Cicero Ronald, Clark Sonya, Dalal Ravindra, DeWinter Alex, Dixon John, Foquet Mathieu, Gaertner Alfred, Hardenbol Paul, Heiner Cheryl, Hester Kevin, Holden David, Kearns Gregory, Kong Xiangxu, Kuse Ronald, Lacroix Yves, Lin Steven, Lundquist Paul, Ma Congcong, Marks Patrick, Maxham Mark, Murphy Devon, Park Insil, Pham Thang, Phillips Michael, Roy Joy, Sebra Robert, Shen Gene, Sorenson Jon, Tomaney Austin, Travers Kevin, Trulson Mark, Vieceli John, Wegener Jeffrey, Wu Dawn, Yang Alicia, Zaccarin Denis, Zhao Peter, Zhong Frank, Korlach Jonas, Turner Stephen. Real-time DNA sequencing from single polymerase molecules // Science. jan 2009. 323, 5910. 133–138.

Eisen Jonathan A. The RecA protein as a model molecule for molecular systematic studies of bacteria: Comparison of trees of RecAs and 16S rRNAs from the same species // Journal of Molecular Evolution. dec 1995. 41, 6. 1105–1123.

Elser James J., Bracken Matthew E.S., Cleland Elsa E., Gruner Daniel S., Harpole W. Stanley, Hillebrand Helmut, Ngai Jacqueline T., Seabloom Eric W., Shurin Jonathan B., Smith Jennifer E. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems // Ecology Letters. dec 2007. 10, 12. 1135–1142.

Eme Laura, Gentekaki Eleni, Curtis Bruce, Archibald John M., Roger Andrew J. Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut // Current Biology.

mar 2017. 27, 6. 807–820.

Eren A. Murat, Maignien Loïs, Sul Woo Jun, Murphy Leslie G., Grim Sharon L., Morrison Hilary G., Sogin Mitchell L. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data // Methods in Ecology and Evolution. dec 2013. 4, 12. 1111–1119.

Escobar-Zepeda Alejandra, De León Arturo Vera Ponce, Sanchez-Flores Alejandro. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. dec 2015. 348.

Fagerbakke Kjell Magne, Heldal Mikal, Norland Svein. Content of carbon, nitrogen, oxygen, sulfur and phosphorus in native aquatic and cultured bacteria // Aquatic Microbial Ecology. mar 1996. 10, 1. 15–27.

Fernanda Sánchez-Soto Jiménez Ma, Cerqueda-García Daniel, Montero-Muñoz Jorge L., Leopoldina Aguirre-Macedo Ma, García-Maldonado José Q. Assessment of the bacterial community structure in shallow and deep sediments of the Perdido Fold Belt region in the Gulf of Mexico // PeerJ. sep 2018. 2018, 9. e5583.

Field Christopher B., Behrenfeld Michael J., Randerson James T., Falkowski Paul. Primary production of the biosphere: Integrating terrestrial and oceanic components // Science. jul 1998. 281, 5374. 237–240.

Field Dawn, Amaral-Zettler Linda, Cochrane Guy, Cole James R., Dawyndt Peter, Garrity George M., Gilbert Jack, Glöckner Frank Oliver, Hirschman Lynette, Karsch-Mizrachi Ilene, Klenk Hans Peter, Knight Rob, Kottmann Renzo, Kyrpides Nikos, Meyer Folker, Gil Inigo San, Sansone Susanna Assunta, Schriml Lynn M., Sterk Peter, Tatusova Tatiana, Ussery David W., White Owen, Wooley John. The Genomic Standards Consortium // PLoS Biology. jun 2011. 9, 6. e1001088.

Field Dawn, Garrity George, Gray Tanya, Morrison Norman, Selengut Jeremy, Sterk Peter, Tatusova Tatiana, Thomson Nicholas, Allen Michael J, Angiuoli Samuel V, Ashburner Michael,

Axelrod Nelson, Baldauf Sandra, Ballard Stuart, Boore Jeffrey, Cochrane Guy, Cole James, Dawyndt Peter, De Vos Paul, Depamphilis Claude, Edwards Robert, Faruque Nadeem, Feldman Robert, Gilbert Jack, Gilna Paul, Glöckner Frank Oliver, Goldstein Philip, Guralnick Robert, Haft Dan, Hancock David, Hermjakob Henning, Hertz-Fowler Christiane, Hugenholtz Phil, Joint Ian, Kagan Leonid, Kane Matthew, Kennedy Jessie, Kowalchuk George, Kottmann Renzo, Kolker Eugene, Kravitz Saul, Kyrpides Nikos, Leebens-Mack Jim, Lewis Suzanna E, Li Kelvin, Lister Allyson L, Lord Phillip, Maltsev Natalia, Markowitz Victor, Martiny Jennifer, Methe Barbara, Mizrachi Ilene, Moxon Richard, Nelson Karen, Parkhill Julian, Proctor Lita, White Owen, Sansone Susanna Assunta, Spiers Andrew, Stevens Robert, Swift Paul, Taylor Chris, Tateno Yoshio, Tett Adrian, Turner Sarah, Ussery David, Vaughan Bob, Ward Naomi, Whetzel Trish, San Gil Ingio, Wilson Gareth, Wipat Anil. The minimum information about a genome sequence (MIGS) specification. may 2008. 541–547.

Figueras María José, Beaz-Hidalgo Roxana, Hossain Mohammad J, Liles Mark R. Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis // Genome Announcements. dec 2014. 2, 6.

Fišer Cene, Luštrik Roman, Sarbu Serban, Flot Jean François, Trontelj Peter. Morphological evolution of coexisting amphipod species pairs from sulfidic caves suggests competitive interactions and character displacement, but no environmental filtering and convergence // PLoS ONE. 2015. 10, 4.

Fleischmann Robert D., Adams Mark D., White Owen, Clayton Rebecca A., Kirkness Ewen F., Kerlavage Anthony R., Bult Carol J., Tomb Jean Francois, Dougherty Brian A., Merrick Joseph M., McKenney Keith, Sutton Granger, FitzHugh Will, Fields Chris, Gocayne Jeannine D., Scott John, Shirley Robert, Liu Li Ing, Glodek Anna, Kelley Jenny M., Weidman Janice F., Phillips Cheryl A., Spriggs Tracy, Hedblom Eva, Cotton Matthew D., Utterback Teresa R., Hanna Michael C., Nguyen David T., Saudek Deborah M., Brandon Rhonda C., Fine Leah D., Fritchman Janice L., Fuhrmann Joyce L., Geoghagen N. S.M., Gnehm Cheryl L., McDonald Lisa A., Small Keith V., Fraser Claire M., Smith Hamilton O., Venter J. Craig. Whole-genome

random sequencing and assembly of Haemophilus influenzae Rd // Science. jul 1995. 269, 5223. 496–512.

Flot Jean François, Bauermeister Jan, Brad Traian, Hillebrand-Voiculescu Alexandra, Sarbu Serban M., Dattagupta Sharmishtha. Niphargus-Thiothrix associations may be widespread in sulphidic groundwater ecosystems: Evidence from southeastern Romania // Molecular Ecology. 2014. 23, 6. 1405–1417.

Forget Bernard G, Weissman Sherman M. Nucleotide sequence of KB cell 5S RNA // Science. dec 1967. 158, 3809. 1695–1699.

Fouhy Fiona, Clooney Adam G., Stanton Catherine, Claesson Marcus J., Cotter Paul D. 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform // BMC Microbiology. dec 2016. 16, 1. 123.

Fox G. E., Pechman K. R., Woese C. R. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics // International Journal of Systematic Bacteriology. jan 1977. 27, 1. 44–57.

Fox G E, Wisotzkey J D, Jurtshuk P. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity // International Journal of Systematic Bacteriology. jan 1992. 42, 1. 166–170.

Frias-Lopez Jorge, Shi Yanmei, Tyson Gene W, Coleman Maureen L, Schuster Stephan C, Chisholm Sallie W, DeLong Edward F. Microbial community gene expression in ocean surface waters // Proceedings of the National Academy of Sciences of the United States of America. mar 2008. 105, 10. 3805–3810.

Fu Limin, Niu Beifang, Zhu Zhengwei, Wu Sitao, Li Weizhong. CD-HIT: Accelerated for clustering the next-generation sequencing data // Bioinformatics. dec 2012. 28, 23. 3150–3152.

Ganzert Lars, Schirmack Janosch, Alawi Mashal, Mangelsdorf Kai, Sand Wolfgang, Hillebrand-Voiculescu Alexandra, Wagner Dirk. Methanosarcina spelaei sp. nov., a methanogenic ar-

chaeon isolated from a floating biofilm of a subsurface sulphurous lake // International journal of systematic and evolutionary microbiology. 2014. 64. 3478–3484.

Garrido-Cardenas Jose Antonio, Garcia-Maroto Federico, Alvarez-Bermejo Jose Antonio, Manzano-Agugliaro Francisco. DNA sequencing sensors: An overview. mar 2017.

Geisen Stefan, Tveit Alexander T, Clark Ian M, Richter Andreas, Svenning Mette M, Bonkowski Michael, Urich Tim. Metatranscriptomic census of active protists in soils // ISME Journal. oct 2015. 9, 10. 2178–2190.

Ghurye Jay S, Cepeda-Espinoza Victoria, Pop Mihai. Metagenomic assembly: Overview, challenges and applications. 2016. 353–362.

Gilbert Jack A., Dupont Christopher L. Microbial metagenomics: Beyond the genome // Annual Review of Marine Science. jan 2011. 3, 1. 347–371.

Giovannoni Stephen J., Britschgi Theresa B., Moyer Craig L., Field Katharine G. Genetic diversity in Sargasso Sea bacterioplankton // Nature. may 1990. 345, 6270. 60–63.

Glockner F. O., Zaichikov E, Belkova N, Denissova L, Pernthaler J., Pernthaler A, Amann R. Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria // Applied and Environmental Microbiology. nov 2000. 66, 11. 5053–5065.

Gloor Gregory B., Macklaim Jean M., Pawlowsky-Glahn Vera, Egozcue Juan J. Microbiome datasets are compositional: And this is not optional. nov 2017. 2224.

Glücksman Edvard, Bell Thomas, Griffiths Robert I, Bass David. Closely related protist strains have different grazing impacts on natural bacterial communities // Environmental Microbiology. dec 2010. 12, 12. 3105–3113.

Gong Weida, Marchetti Adrian. Estimation of 18S gene copy number in marine eukaryotic plankton using a next-generation sequencing approach // Frontiers in Marine Science. apr 2019. 6, APR. 219.

Goodwin Sara, McPherson John D, McCombie W. Richard. Coming of age: Ten years of next-generation sequencing technologies. 2016. 333–351.

Goris Johan, Konstantinidis Konstantinos T, Klappenbach Joel A, Coenye Tom, Vandamme Peter, Tiedje James M. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities // International Journal of Systematic and Evolutionary Microbiology. jan 2007. 57, 1. 81–91.

Granin N. G., Aslamov I. A., Kozlov V. V., Makarov M. M., Kirillin G., McGinnis D. F., Kucher K. M., Blinov V. V., Ivanov V. G., Mizandrontsev I. B., Zhdanov A. A., Anikin A. S., Granin M. N., Gnatovsky R. Yu. Methane hydrate emergence from Lake Baikal: direct observations, modelling, and hydrate footprints in seasonal ice cover // Scientific Reports. dec 2019. 9, 1. 19361.

Guillou Laure, Bachar Dipankar, Audic St??phane, Bass David, Berney C??dric, Bittner Lucie, Boutte Christophe, Burgaud Ga??tan, De Vargas Colomban, Decelle Johan, Del Campo Javier, Dolan John R., Dunthorn Micah, Edvardsen Bente, Holzmann Maria, Kooistra Wiebe H C F, Lara Enrique, Le Bescot Noan, Logares Ramiro, Mah?? Fr??d??ric, Massana Ramon, Montresor Marina, Morard Raphael, Not Fabrice, Pawlowski Jan, Probert Ian, Sauvadet Anne Laure, Siano Raffaele, Stoeck Thorsten, Vaulot Daniel, Zimmermann Pascal, Christen Richard. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy // Nucleic Acids Research. jan 2013. 41, D1. D597–604.

Gutiérrez-Preciado Ana, Saghaï Aurélien, Moreira David, Zivanovic Yvan, Deschamps Philippe, López-García Purificación. Functional shifts in microbial mats recapitulate early Earth metabolic transitions // Nature Ecology & Evolution. nov 2018. 2, 11. 1700–1708.

Guy Lionel, Ettema Thijs J G. The archaeal 'TACK' superphylum and the origin of eukaryotes. dec 2011. 580–587.

Haas Brian J, Gevers Dirk, Earl Ashlee M, Feldgarden Mike, Ward Doyle V, Giannoukos Georgia, Ciulla Dawn, Tabbaa Diana, Highlander Sarah K, Sodergren Erica, Methé Barbara, DeSantis Todd Z, Petrosino Joseph F, Knight Rob, Birren Bruce W. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons // Genome Research. mar 2011. 21, 3. 494–504.

Hagen Joel B. The origins of bioinformatics. dec 2000. 231–236.

Hagen Joel B. 1The introduction of computers into systematic research in the United States during the 1960s // Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences. jun 2001. 32, 2. 291–314.

Hall Neil. Advanced sequencing technologies and their wider impact in microbiology. may 2007. 1518–1525.

Hampton Stephanie E., Izmest'eva Lyubov R., Moore Marianne V., Katz Stephen L., Dennis Brian, Silow Eugene A. Sixty years of environmental change in the world's largest freshwater lake - Lake Baikal, Siberia // Global Change Biology. aug 2008. 14, 8. 1947–1958.

Harris Timothy D, Buzby Phillip R, Babcock Hazen, Beer Eric, Bowers Jayson, Braslavsky Ido, Causey Marie, Colonell Jennifer, DiMeo James, Efcavitch J William, Giladi Eldar, Gill Jaime, Healy John, Jarosz Mirna, Lapen Dan, Moulton Keith, Quake Stephen R, Steinmann Kathleen, Thayer Edward, Tyurina Anastasia, Ward Rebecca, Weiss Howard, Xie Zheng. Single-molecule DNA sequencing of a viral genome // Science. apr 2008. 320, 5872. 106–109.

Heather James M, Chain Benjamin. The sequence of sequencers: The history of sequencing DNA. jan 2016. 1–8.

Hesper B, Hogeweg P. Bioinformatica: een werkconcept // Kameleon. 1970. 1, 6. 28–29.

Hewson Ian, Jacobson-Meyers Myrna E., Fuhrman Jed A. Diversity and biogeography of bacterial assemblages in surface sediments across the San Pedro Basin, Southern California Borderlands // Environmental Microbiology. apr 2007. 9, 4. 923–933.

Heydari Mahdi, Miclotte Giles, Demeester Piet, Van de Peer Yves, Fostier Jan. Evaluation of the impact of Illumina error correction tools on de novo genome assembly // BMC Bioinformatics. dec 2017. 18, 1. 374.

Hogeweg Paulien. The roots of bioinformatics in theoretical biology // PLoS Computational Biology. mar 2011. 7, 3. e1002021.

Hohmann R., Kipfer R., Peeters F., Piepke G., Imboden D. M., Shimaraev M. N. Processes of deep-water renewal in Lake Baikal // Limnology and Oceanography. jul 1997. 42, 5. 841–855.

Holley Robert W., Apgar Jean, Everett George A., Madison James T., Marquisee Mark, Merrill Susan H., Penswick John Robert, Zamir Ada. Structure of a ribonucleic acid // Science. mar 1965. 147, 3664. 1462–1465.

Hu Yu, Fang Li, Nicholson Christopher, Wang Kai. Implications of Error-Prone Long-Read Whole-Genome Shotgun Sequencing on Characterizing Reference Microbiomes // iScience. jun 2020. 23, 6. 101223.

Hug Laura A., Baker Brett J., Anantharaman Karthik, Brown Christopher T., Probst Alexander J., Castelle Cindy J., Butterfield Cristina N., Hernsdorf Alex W., Amano Yuki, Ise Kotaro, Suzuki Yohey, Dudek Natasha, Relman David A., Finstad Kari M., Amundson Ronald, Thomas Brian C., Banfield Jillian F. A new view of the tree of life // Nature Microbiology. may 2016. 1, 5. 16048.

Hunkapiller T, Kaiser R J, Koop B F, Hood L. Large-scale and automated DNA sequence determination. oct 1991. 59–67.

Hutchens Elena, Radajewski Stefan, Dumont Marc G., McDonald Ian R., Murrell J. Colin. Analysis of methanotrophic bacteria in Movile Cave by stable isotope probing // Environmental Microbiology. 2004. 6, 2. 111–120.

Hyman Edward David. A new method of sequencing DNA // Analytical Biochemistry. nov 1988. 174, 2. 423–436.

Imachi Hiroyuki, Nobu Masaru K., Nakahara Nozomi, Morono Yuki, Ogawara Miyuki, Takaki Yoshihiro, Takano Yoshinori, Uematsu Katsuyuki, Ikuta Tetsuro, Ito Motoo, Matsui Yohei, Miyazaki Masayuki, Murata Kazuyoshi, Saito Yumi, Sakai Sanae, Song Chihong, Tasumi Eiji, Yamanaka Yuko, Yamaguchi Takashi, Kamagata Yoichi, Tamaki Hideyuki, Takai Ken. Isolation of an archaeon at the prokaryote–eukaryote interface // Nature. jan 2020. 577, 7791. 519–525.

Itskovich Valeria, Gontcharov Andrey, Masuda Yoshiki, Nohno Tsutomu, Belikov Sergey, Efremova Sofia, Meixner Martin, Janussen Dorte. Ribosomal ITS sequences allow resolution of freshwater sponge phylogeny with alignments guided by secondary structure prediction // Journal of Molecular Evolution. dec 2008. 67, 6. 608–620.

Johnson Jethro S., Spakowicz Daniel J., Hong Bo Young, Petersen Lauren M., Demkowicz Patrick, Chen Lei, Leopold Shana R., Hanson Blake M., Agresta Hanako O., Gerstein Mark, Sodergren Erica, Weinstock George M. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis // Nature Communications. dec 2019. 10, 1. 5029.

Jünemann Sebastian, Kleinbölting Nils, Jaenicke Sebastian, Henke Christian, Hassa Julia, Nelkner Johanna, Stolze Yvonne, Albaum Stefan P., Schlüter Andreas, Goesmann Alexander, Sczyrba Alexander, Stoye Jens. Bioinformatics for NGS-based metagenomics and the application to biogas research. nov 2017. 10–23.

Kadnikov V. V., Lomakina A. V., Likhoshvai A. V., Gorshkov A. G., Pogodaeva T. V., Beletsky A. V., Mardanov A. V., Zemskaya T. I., Ravin N. V. Composition of the microbial communities of bituminous constructions at natural oil seeps at the bottom of Lake Baikal // Microbiology (Russian Federation). may 2013. 82, 3. 373–382.

Kadnikov Vitaly V, Mardanov Andrey V, Beletsky Alexey V, Shubenkova Olga V, Pogodaeva Tatiana V, Zemskaya Tamara I, Ravin Nikolai V, Skryabin Konstantin G. Microbial community structure in methane hydrate-bearing sediments of freshwater Lake Baikal // FEMS Microbiology Ecology. 2012. 79, 2. 348–358.

Kamble Asmita, Singh Harinder. Different Methods of Soil DNA Extraction // BIO-PROTOCOL. 2020. 10, 2.

Kazamia Elena, Helliwell Katherine Emma, Purton Saul, Smith Alison Gail. How mutualisms arise in phytoplankton communities: building eco-evolutionary principles for aquatic microbes. jul 2016. 810–822.

Kendrew J. C., Bodo G., Dintzis H. M., Parrish R. G., Wyckoff H., Phillips D. C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis // Nature. mar 1958. 181, 4610. 662–666.

Kim Mincheol, Oh Hyun Seok, Park Sang Cheol, Chun Jongsik. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes // International Journal of Systematic and Evolutionary Microbiology. feb 2014. 64, PART 2. 346–351.

Kiss Antal, Sain Béla, Venetianer Pál. The number of rRNA genes in Escherichia coli // FEBS Letters. jul 1977. 79, 1. 77–79.

Klappenbach Joel A, Saxman Paul R, Cole James R, Schmidt Thomas M. Rrndb: The ribosomal RNA operon copy number database // Nucleic Acids Research. jan 2001. 29, 1. 181–184.

Klump J. Val, Edgington David N., Granina Liba, Remsen Charles C. Estimates of the remineralization and burial of organic carbon in Lake Baikal sediments // Journal of Great Lakes Research. feb 2020. 46, 1. 102–114.

Knight Rob, Vrbanac Alison, Taylor Bryn C., Aksenov Alexander, Callewaert Chris, Debelius Justine, Gonzalez Antonio, Kosciolek Tomasz, McCall Laura Isobel, McDonald Daniel, Melnik Alexey V., Morton James T., Navas Jose, Quinn Robert A., Sanders Jon G., Swafford Austin D., Thompson Luke R., Tripathi Anupriya, Xu Zhenjiang Z., Zaneveld Jesse R., Zhu Qiyun, Caporaso J. Gregory, Dorrestein Pieter C. Best practices for analysing microbiomes. jul 2018. 410–422.

Konstantinidis Konstantinos T, Tiedje James M. Genomic insights that advance the species definition for prokaryotes // Proceedings of the National Academy of Sciences of the United States of America. feb 2005a. 102, 7. 2567–2572.

Konstantinidis Konstantinos T, Tiedje James M. Towards a genome-based taxonomy for prokaryotes // Journal of Bacteriology. sep 2005b. 187, 18. 6258–6264.

Kowalchuk George A, Speksnijder Arjen G.C.L., Zhang Kun, Goodman Robert M, Van Veen Johannes A. Finding the needles in the metagenome haystack // Microbial Ecology. 53, 3. apr 2007. 475–485.

Kruskal J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis // Psychometrika. mar 1964. 29, 1. 1–27.

Ku Chuan, Sebé-Pedrós Arnau. Using single-cell transcriptomics to understand functional states and interactions in microbial eukaryotes // Philosophical Transactions of the Royal Society B: Biological Sciences. nov 2019. 374, 1786. 20190098.

Kulikovskiy Maxim, Lange-Bertalot Horst, Witkowski Andrzej, Khursevich Galina. Achnanthidium sibiricum (Bacillariophyceae), a new species from bottom sediments in Lake Baikal // Algological Studies. mar 2011. 136-137. 77–87.

Kumaresan Deepak, Wischer Daniela, Stephenson Jason, Hillebrand-Voiculescu Alexandra, Murrell J. Colin. Microbiology of Movile Cave-A Chemolithoautotrophic Ecosystem // Geomicrobiology Journal. 2014. 31, 3. 186–193.

Kurilkina Maria I, Zakharova Yulia R, Galachyants Yuri P, Petrova Darya P, Bukin Yuri S, Domysheva Valentina M, Blinov Vadim V, Likhoshway Yelena V. Bacterial community composition in the water column of the deepest freshwater Lake Baikal as determined by next-generation sequencing // FEMS Microbiology Ecology. 2016. 92, 7.

Kwok Shirley, Chang Sheng Yung, Sninsky John J, Wang Alice. A guide to the design and use of mismatched and degenerate primers // Genome Research. feb 1994. 3, 4. S39–47.

Lagesen Karin, Hallin Peter, Rødland Einar Andreas, Stærfeldt Hans Henrik, Rognes Torbjørn, Ussery David W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes // Nucleic Acids Research. 2007. 35, 9. 3100–3108.

Lambert Christophe, Campenhout Jean-Marc, DeBolle Xavier, Depiereux Eric. Review of Common Sequence Alignment Methods: Clues to Enhance Reliability // Current Genomics. feb 2005. 4, 2. 131–146.

Lander Eric S., Linton Lauren M., Birren Bruce, Nusbaum Chad, Zody Michael C., Baldwin Jennifer, Devon Keri, Dewar Ken, Doyle Michael, Fitzhugh William, Funke Roel, Gage Diane, Harris Katrina, Heaford Andrew, Howland John, Kann Lisa, Lehoczky Jessica, Levine Rosie, McEwan Paul, McKernan Kevin, Meldrim James, Mesirov Jill P., Miranda Cher, Morris William, Naylor Jerome, Raymond Christina, Rosetti Mark, Santos Ralph, Sheridan Andrew, Sougnez Carrie, Stange-Thomann Nicole, Stojanovic Nikola, Subramanian Aravind, Wyman Dudley, Rogers Jane, Sulston John, Ainscough Rachael, Beck Stephan, Bentley David, Burton John, Clee Christopher, Carter Nigel, Coulson Alan, Deadman Rebecca, Deloukas Panos, Dunham Andrew, Dunham Ian, Durbin Richard, French Lisa, Grafham Darren, Gregory Simon, Hubbard Tim, Humphray Sean, Hunt Adrienne, Jones Matthew, Lloyd Christine, McMurray Amanda, Matthews Lucy, Mercer Simon, Milne Sarah, Mullikin James C., Mungall Andrew, Plumb Robert, Ross Mark, Shownkeen Ratna, Sims Sarah, Waterston Robert H., Wilson Richard K., Hillier Ladeana W., McPherson John D., Marra Marco A., Mardis Elaine R., Fulton Lucinda A., Chinwalla Asif T., Pepin Kymberlie H., Gish Warren R., Chissoe Stephanie L., Wendl Michael C., Delehaunty Kim D., Miner Tracie L., Delehaunty Andrew, Kramer Jason B., Cook Lisa L., Fulton Robert S., Johnson Douglas L., Minx Patrick J., Clifton Sandra W., Hawkins Trevor, Branscomb Elbert, Predki Paul, Richardson Paul, Wenning Sarah, Slezak Tom, Doggett Norman, Cheng Jan Fang, Olsen Anne, Lucas Susan, Elkin Christopher, Uberbacher Edward, Frazier Marvin, Gibbs Richard A., Muzny Donna M., Scherer Steven E., Bouck John B., Sodergren Erica J., Worley Kim C., Rives Catherine M., Gorrell James H., Metzker Michael L., Naylor Susan L., Kucherlapati Raju S., Nelson David L., Weinstock George M.,

Sakaki Yoshiyuki, Fujiyama Asao, Hattori Masahira, Yada Tetsushi, Toyoda Atsushi, Itoh Takehiko, Kawagoe Chiharu, Watanabe Hidemi, Totoki Yasushi, Taylor Todd, Weissenbach Jean, Heilig Roland, Saurin William, Artiguenave Francois, Brottier Philippe, Bruls Thomas, Pelletier Eric, Robert Catherine, Wincker Patrick, Rosenthal André, Platzer Matthias, Nyakatura Gerald, Taudien Stefan, Rump Andreas, Smith Douglas R., Doucette-Stamm Lynn, Rubenfield Marc, Weinstock Keith, Hong Mei Lee, Dubois Joann, Yang Huanming, Yu Jun, Wang Jian, Huang Guyang, Gu Jun, Hood Leroy, Rowen Lee, Madan Anup, Qin Shizen, Davis Ronald W., Federspiel Nancy A., Abola A. Pia, Proctor Michael J., Roe Bruce A., Chen Feng, Pan Huaqin, Ramser Juliane, Lehrach Hans, Reinhardt Richard, McCombie W. Richard, De La Bastide Melissa, Dedhia Neilay, Blöcker Helmut, Hornischer Klaus, Nordsiek Gabriele, Agarwala Richa, Aravind L., Bailey Jeffrey A., Bateman Alex, Batzoglou Serafim, Birney Ewan, Bork Peer, Brown Daniel G., Burge Christopher B., Cerutti Lorenzo, Chen Hsiu Chuan, Church Deanna, Clamp Michele, Copley Richard R., Doerks Tobias, Eddy Sean R., Eichler Evan E., Furey Terrence S., Galagan James, Gilbert James G.R., Harmon Cyrus, Hayashizaki Yoshihide, Haussler David, Hermjakob Henning, Hokamp Karsten, Jang Wonhee, Johnson L. Steven, Jones Thomas A., Kasif Simon, Kaspryzk Arek, Kennedy Scot, Kent W. James, Kitts Paul, Koonin Eugene V., Korf Ian, Kulp David, Lancet Doron, Lowe Todd M., McLysaght Aoife, Mikkelsen Tarjei, Moran John V., Mulder Nicola, Pollara Victor J., Ponting Chris P., Schuler Greg, Schultz Jörg, Slater Guy, Smit Arian F.A., Stupka Elia, Szustakowki Joseph, Thierry-Mieg Danielle, Thierry-Mieg Jean, Wagner Lukas, Wallis John, Wheeler Raymond, Williams Alan, Wolf Yuri I., Wolfe Kenneth H., Yang Shiaw Pyng, Yeh Ru Fang, Collins Francis, Guyer Mark S., Peterson Jane, Felsenfeld Adam, Wetterstrand Kris A., Myers Richard M., Schmutz Jeremy, Dickson Mark, Grimwood Jane, Cox David R., Olson Maynard V., Kaul Rajinder, Raymond Christopher, Shimizu Nobuyoshi, Kawasaki Kazuhiko, Minoshima Shinsei, Evans Glen A., Athanasiou Maria, Schultz Roger, Patrinos Aristides, Morgan Michael J. Initial sequencing and analysis of the human genome // Nature. feb 2001. 409, 6822. 860–921.

Lang Jenna Morgan, Darling Aaron E., Eisen Jonathan A. Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices // PLoS ONE. apr 2013.

8, 4. e62510.

Lapteva N. A., Bel'Kova N. L., Parfenova V. V. Spatial distribution and species composition of prosthecate bacteria in Lake Baikal // Microbiology. aug 2007. 76, 4. 480–486.

Laroche Olivier, Wood Susanna A., Tremblay Louis A., Lear Gavin, Ellis Joanne I., Pochon Xavier. Metabarcoding monitoring analysis: the pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities // PeerJ. may 2017. 5. e3347.

Larsen Niels, Olsen Gary J, Maidak Bonnie L, Mccaughey Michael J., Overbeek Ross, Macke Thomas J, Marsh Terry L, Woese Carl R. The ribosomal database project // Nucleic Acids Research. jul 1993. 21, 13. 3021–3023.

Lascu C. Paleogeographical and hydrogeological hypothesis regarding the origin of a peculiar cave fauna // Misc speol Rom Bucharest. 1989. 1. 13–18.

Lechevalier H. Louis Joblot and his microscopes // Bacteriological Reviews. mar 1976. 40, 1. 241–258.

Legendre Piere, Andersson Marti J. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments // Ecological Monographs. feb 1999. 69, 1. 1–24.

Legendre Pierre, Borcard Daniel, Peres-Neto Pedro R. Analyzing beta diversity: Partitioning the spatial variation of community composition data // Ecological Monographs. nov 2005. 75, 4. 435–450.

Leger Michelle M., Eme Laura, Stairs Courtney W., Roger Andrew J. Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). may 2018. 1700242.

Leininger S., Urich T., Schloter M., Schwark L., Qi J., Nicol G. W., Prosser J. I., Schuster S. C., Schleper C. Archaea predominate among ammonia-oxidizing prokaryotes in soils // Nature. aug 2006. 442, 7104. 806–809.

Lennon J T, Muscarella M E, Placella S A, Lehmkuhl B K. How, When, and Where Relic DNA Affects Microbial Diversity. // mBio. jul 2018. 9, 3.

Lennon Jay T, Locey Kenneth J. More support for Earth's massive microbiome // Biology Direct. 2020. 15, 1. 5.

Lever Mark A., Torti Andrea, Eickenbusch Philip, Michaud Alexander B., Å antl-Temkiv Tina, JÃ¸rgensen Bo Barker. A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types // Frontiers in Microbiology. may 2015. 6. 476.

Lomakina A. V., Pogodaeva T. V., Morozov I. V., Zemskaya T. I. Microbial communities of the discharge zone of oil- and gas-bearing fluids in low-mineral Lake Baikal // Microbiology (Russian Federation). may 2014. 83, 3. 278–287.

Lomakina Anna, Pogodaeva Tatyana, Kalmychkov Gennady, Chernitsyna Svetlana, Zemskaya Tamara. Diversity of NC10 bacteria and ANME-2d archaea in sediments of fault zones at Lake Baikal // Diversity. dec 2020. 12, 1. 10.

Lomakina Anna V., Mamaeva Elena V., Galachyants Yuri P., Petrova Darya P., Pogodaeva Tatyana V., Shubenkova Olga V., Khabuev Andrey V., Morozov Igor V., Zemskaya Tamara I. Diversity of Archaea in Bottom Sediments of the Discharge Areas With Oil- and Gas-Bearing Fluids in Lake Baikal // Geomicrobiology Journal. jan 2018. 35, 1. 50–63.

López-García Purificación, Moreira David. Cultured Asgard Archaea Shed Light on Eukaryogenesis // Cell. apr 2020. 181, 2. 232–235.

López-García Purificación, Rodríguez-Valera Francisco, Pedrós-Alió Carlos, Moreira David. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton // Nature. feb 2001. 409, 6820. 603–607.

López-García Purificación, Vereshchaka Alexander, Moreira David. Eukaryotic diversity associated with carbonates and fluid-seawater interface in Lost City hydrothermal field // Environmental Microbiology. feb 2007. 9, 2. 546–554.

Louca Stilianos, Doebeli Michael, Parfrey Laura Wegener. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem // Microbiome. dec 2018. 6, 1. 41.

Louca Stilianos, Mazel Florent, Doebeli Michael, Parfrey Laura Wegener. A census-based estimate of earth's bacterial and archaeal diversity // PLoS Biology. 2019. 17, 2. e3000106.

Lozupone Catherine, Knight Rob. UniFrac: A new phylogenetic method for comparing microbial communities // Applied and Environmental Microbiology. dec 2005. 71, 12. 8228–8235.

Madigan Mickael T. (Southern Illinois University Carbondale), Martinko John M. (Southern Illinois University Carbondale), Bender Kelly S. (Southern Illinois University Carbondale), Buckley Daniel H. (Cornell University), Stahl David A. (University of Seatle). Brock biology of microorganisms 14th edition. 2015.

Magoč Tanja, Salzberg Steven L. FLASH: Fast length adjustment of short reads to improve genome assemblies // Bioinformatics. 2011. 27, 21. 2957–2963.

Mahé Frédéric, Rognes Torbjørn, Quince Christopher, De Vargas Colomban, Dunthorn Micah. Swarm v2: highly-scalable and high-resolution amplicon clustering // PeerJ. 2015. 3. e1420.

Mahé Frédéric, Rognes Torbjørn, Quince Christopher, Vargas Colomban de, Dunthorn Micah. Swarm: Robust and fast clustering method for amplicon-based studies // PeerJ. sep 2014. 2014, 1. e593.

Maidak Bonnie L., Olsen Gary J., Larsen Niels, Overbeek Ross, McCaughey Michael J., Woese Carl R. The Ribosomal Database Project (RDP). jan 1996. 82–85.

Mall Achim, Sobotta Jessica, Huber Claudia, Tschirner Carolin, Kowarschik Stefanie, Bačnik Katarina, Mergelsberg Mario, Boll Matthias, Hügler Michael, Eisenreich Wolfgang, Berg Ivan A. Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium // Science. feb 2018. 359, 6375. 563–567.

Mande Sharmila S., Mohammed Monzoorul Haque, Ghosh Tarini Shankar. Classification of metagenomic sequences: Methods and challenges // Briefings in Bioinformatics. nov 2012. 13, 6. 669–681.

Martin Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads // EMBnet.journal. 2011. 17, 1. pp. 10–12.

Mats Victor D., Perepelova Tatiana I. A new perspective on evolution of the Baikal Rift // Geoscience Frontiers. jul 2011. 2, 3. 349–365.

McLaren Michael R, Willis Amy D, Callahan Benjamin J. Consistent and correctable bias in metagenomic sequencing experiments // eLife. sep 2019. 8.

Menzel Peter, Ng Kim Lee, Krogh Anders. Fast and sensitive taxonomic classification for metagenomics with Kaiju // Nature Communications. sep 2016. 7, 1. 11257.

Metzker Michael L. Sequencing technologies the next generation // Nature Reviews Genetics. jan 2010. 11, 1. 31–46.

Mikhail Kozhov , Brooks John Langdon. Lake Baikal and its Life. Mikhail Kozhov // The Quarterly Review of Biology. mar 1965. 40, 1. 74–76.

Mikhailov I S, Zakharova Yu R, Galachyants Yu P, Usoltseva M V, Petrova D P, Sakirko M V, Likhoshway Ye V, Grachev M A. Similarity of structure of taxonomic bacterial communities in the photic layer of Lake Baikal's three basins differing in spring phytoplankton composition and abundance // Doklady Biochemistry and Biophysics. 2015. 465, 1. 413–419.

Mikhailov Ivan S., Zakharova Yulia R., Bukin Yuri S., Galachyants Yuri P., Petrova Darya P., Sakirko Maria V., Likhoshway Yelena V. Co-occurrence Networks Among Bacteria and Microbial Eukaryotes of Lake Baikal During a Spring Phytoplankton Bloom // Microbial Ecology. jan 2019. 77, 1. 96–109.

Mikheyev Alexander S, Tin Mandy M.Y. A first look at the Oxford Nanopore MinION sequencer // Molecular Ecology Resources. nov 2014. 14, 6. 1097–1102.

Miller Christopher S, Baker Brett J, Thomas Brian C, Singer Steven W, Banfield Jillian F. EMIRGE: Reconstruction of full-length ribosomal genes from microbial community short read sequencing data // Genome Biology. may 2011. 12, 5. R44.

Moon-Van Der Staay Seung Yeo, De Wachter Rupert, Vaulot Daniel. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity // Nature. feb 2001. 409, 6820. 607–610.

Moore Marianne V., Hampton Stephanie E., Izmest'eva Lyubov R., Silow Eugene A., Peshkova Ekaterina V., Pavlov Boris K. Climate Change and the World's "Sacred Sea"—Lake Baikal, Siberia // BioScience. 2009. 59, 5. 405–417.

Moreira David, López-García Purificación. The molecular ecology of microbial eukaryotes unveils a hidden world. jan 2002. 31–38.

Mosher Jennifer J., Bowman Brett, Bernberg Erin L., Shevchenko Olga, Kan Jinjun, Korlach Jonas, Kaplan Louis A. Improved performance of the PacBio SMRT technology for 16S rDNA sequencing // Journal of Microbiological Methods. sep 2014. 104. 59–60.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction // Cold Spring Harbor Symposia on Quantitative Biology. 1986. 51, 1. 263–273.

Nagler Magdalena, Insam Heribert, Pietramellara Giacomo, Ascher-Jenull Judith. Extracellular DNA in natural environments: features, relevance and applications. aug 2018. 6343–6356.

Needleman Saul B., Wunsch Christian D. A general method applicable to the search for similarities in the amino acid sequence of two proteins // Journal of Molecular Biology. mar 1970. 48, 3. 443–453.

Nguyen Lam Tung, Schmidt Heiko A., Von Haeseler Arndt, Minh Bui Quang. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies // Molecular Biology and Evolution. jan 2015. 32, 1. 268–274.

Nguyen Nam Phuong, Warnow Tandy, Pop Mihai, White Bryan. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. nov 2016. 16004.

Nicholls Samuel M, Quick Joshua C, Tang Shuiquan, Loman Nicholas J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards // GigaScience. may 2019. 8, 5.

Nirenberg Marshall, Leder Philip. RNA codewords and protein synthesis // Science. sep 1964. 145, 3639. 1399–1407.

Nocker Andreas, Cheung Ching-Ying, Camper Anne K. Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells // Journal of Microbiological Methods. nov 2006. 67, 2. 310–320.

Nováková Alena, Hubka Vít, Valinová Šárka, Kolařík Miroslav, Hillebrand-Voiculescu Alexandra Maria. Cultivable microscopic fungi from an underground chemosynthesis-based ecosystem: a preliminary study // Folia Microbiologica. jan 2018. 63, 1. 43–55.

Nunoura Takuro, Chikaraishi Yoshito, Izaki Rikihisa, Suwa Takashi, Sato Takaaki, Harada Takeshi, Mori Koji, Kato Yumiko, Miyazaki Masayuki, Shimamura Shigeru, Yanagawa Katsunori, Shuto Aya, Ohkouchi Naohiko, Fujita Nobuyuki, Takaki Yoshihiro, Atomi Haruyuki, Takai Ken. A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile // Science. feb 2018. 359, 6375. 559–563.

Nyrén Pål. Enzymatic method for continuous monitoring of DNA polymerase activity // Analytical Biochemistry. dec 1987. 167, 2. 235–238.

Nyrén Pål, Lundin Arne. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis // Analytical Biochemistry. dec 1985. 151, 2. 504–509.

Orcutt B C, George D G, Dayhoff M O. PROTEIN AND NUCLEIC ACID SEQUENCE DATABASE SYSTEMS. // Annual Review of Biophysics and Bioengineering. jun 1983. 12, 1. 419–441.

Orsi William D. Ecology and evolution of seafloor and subseafloor microbial communities. nov 2018. 671–683.

The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. // . 1986. 1–55.

Pajares Silvia, Ramos Ramiro. Processes and Microorganisms Involved in the Marine Nitrogen Cycle: Knowledge and Gaps. nov 2019. 739.

Parks Donovan H., Chuvochina Maria, Chaumeil Pierre Alain, Rinke Christian, Mussig Aaron J., Hugenholtz Philip. A complete domain-to-species taxonomy for Bacteria and Archaea // Nature Biotechnology. apr 2020. 1–8.

Parks Donovan H, Chuvochina Maria, Waite David W, Rinke Christian, Skarshewski Adam, Chaumeil Pierre Alain, Hugenholtz Philip. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life // Nature Biotechnology. nov 2018. 36, 10. 996.

Parks Donovan H, Imelfort Michael, Skennerton Connor T, Hugenholtz Philip, Tyson Gene W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes // Genome Research. 2015. 25, 7. 1043–1055.

Pavlova O. N., Bukin S. V., Kostyreva E., Moskvin V. I., Manakov A. Yu, Morozov I. V., Galachyants Yu P., Khabuev A. V., Zemskaya T. I. Experimental transformation of organic matter by the microbial community from the bottom sediments of Akademichesky Ridge (Lake Baikal) // Russian Geology and Geophysics. 2019. 60, 8. 926–937.

Peretolchina T E, Sitnikova T Ya, Sherbakov D Yu. The complete mitochondrial genomes of four Baikal molluscs from the endemic family Baicaliidae (Caenogastropoda: Truncatelloida) // Journal of Molluscan Studies. apr 2020.

Pérez-Cobas Ana Elena, Gomez-Valero Laura, Buchrieser Carmen. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses // Microbial genomics. aug 2020. 6, 8. e000409.

Pericard Pierre, Dufresne Yoann, Couderc Loïc, Blanquart Samuel, Touzet Hélène. MATAM: Reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes // Bioinformatics. feb 2018. 34, 4. 585–591.

Perutz Max Ferdinand. The structure of haemoglobin - VI. Fourier projections on the 010 plane // Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences. sep 1954. 225, 1162. 315–329.

Piccolroaz Sebastiano, Toffolon Marco. The fate of Lake Baikal: how climate change may alter deep ventilation in the largest lake on Earth // Climatic Change. oct 2018. 150, 3-4. 181–194.

Pimenov N. V., Kalmychkov G. V., Veryasov M. B., Sigalevich P. A., Zemskaya T. I. Microbial oxidation of methane in the sediments of central and southern Baikal // Microbiology (Russian Federation). nov 2014. 83, 6. 773–781.

Poinar Hendrik N., Schwarz Carsten, Qi Ji, Shapiro Beth, MacPhee Ross D.E., Buigues Bernard, Tikhonov Alexei, Huson Daniel M., Tomsho Lynn P., Auch Alexander, Rampp Markus, Miller Webb, Schuster Stephan C. Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA // Science. jan 2006. 311, 5759. 392–394.

Pollock Jolinda, Glendinning Laura, Wisedchanwet Trong, Watson Mick. The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. apr 2018.

Poretsky Rachel S, Bano Nasreen, Buchan Alison, LeCleir Gary, Kleikemper Jutta, Pickering Maria, Pate Whitney M, Moran Mary Ann, Hollibaugh James T. Analysis of microbial gene

transcripts in environmental samples // Applied and Environmental Microbiology. jul 2005. 71, 7. 4121–4126.

Potapov Sergey, Belykh Olga, Krasnopeev Andrey, Gladkikh Anna, Kabilov Marsel, Tupikin Aleksey, Butina Tatyana. Assessing the diversity of the g23 gene of T4-like bacteriophages from Lake Baikal with high-throughput sequencing // FEMS Microbiology Letters. feb 2018. 365, 3.

Potapov Sergey A., Tikhonova Irina V., Krasnopeev Andrey Yu., Kabilov Marsel R., Tupikin Aleksey E., Chebunina Nadezhda S., Zhuchenko Natalia A., Belykh Olga I. Metagenomic analysis of virioplankton from the pelagic zone of lake baikal // Viruses. oct 2019. 11, 11. 991.

Preheim Sarah P, Perrott Allison R., Martin-Platero Antonio M, Gupta Anika, Alm Eric J. Distribution-based clustering: Using ecology to refine the operational taxonomic unit // Applied and Environmental Microbiology. nov 2013. 79, 21. 6593–6603.

Quast Christian, Pruesse Elmar, Yilmaz Pelin, Gerken Jan, Schweer Timmy, Yarza Pablo, Peplies Jörg, Glöckner Frank Oliver. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools // Nucleic Acids Research. jan 2013. 41, D1. D590–D596.

Quinn Thomas P, Erb Ionas, Richardson Mark F, Crowley Tamsyn M. Understanding sequencing data as compositions: An outlook and review // Bioinformatics. aug 2018. 34, 16. 2870–2878.

R Core Team . R: A Language and Environment for Statistical Computing. Vienna, Austria, 2017.

Reboul Guillaume, Moreira David, Bertolino Paola, Hillebrand-Voiculescu Alexandra Maria, López-García Purificación. Microbial eukaryotes in the suboxic chemosynthetic ecosystem of Movile Cave, Romania // Environmental Microbiology Reports. 2019. 11, 3. 464–473.

Reñé Albert, Auladell Adrià, Reboul Guillaume, Moreira David, López-García Purificación. Performance of the melting seawater-ice elution method on the metabarcoding characterization of benthic protist communities // Environmental Microbiology Reports. 2020. 00.

Reysenbach A L, Giver L J, Wickham G S, Pace N R. Differential amplification of rRNA genes by polymerase chain reaction. oct 1992. 3417–3418.

Riesenfeld Christian S., Schloss Patrick D., Handelsman Jo. Metagenomics: Genomic analysis of microbial communities. dec 2004. 525–552.

Rinke Christian, Schwientek Patrick, Sczyrba Alexander, Ivanova Natalia N., Anderson Iain J., Cheng Jan Fang, Darling Aaron, Malfatti Stephanie, Swan Brandon K., Gies Esther A., Dodsworth Jeremy A., Hedlund Brian P., Tsiamis George, Sievert Stefan M., Liu Wen Tso, Eisen Jonathan A., Hallam Steven J., Kyrpides Nikos C., Stepanauskas Ramunas, Rubin Edward M., Hugenholtz Philip, Woyke Tanja. Insights into the phylogeny and coding potential of microbial dark matter // Nature. jul 2013. 499, 7459. 431–437.

Roberts Sarah L., Swann George E.A., McGowan Suzanne, Panizzo Virginia N., Vologina Elena G., Sturm Michael, Mackay Anson W. Diatom evidence of 20th century ecosystem change in Lake Baikal, Siberia // PLoS ONE. dec 2018. 13, 12. e0208765.

Rognes Torbjørn, Flouri Tomáš, Nichols Ben, Quince Christopher, Mahé Frédéric. VSEARCH: a versatile open source tool for metagenomics // PeerJ Preprints. 2016. 4. e2409v1.

Rohwerder T., Sand W., Lascu C. Preliminary evidence for a sulphur cycle in Movile Cave, Romania. 2003. 101–107.

Rosselló-Mora Ramon, Amann Rudolf. The species concept for prokaryotes // FEMS Microbiology Reviews. jan 2001. 25, 1. 39–67.

Rothberg Jonathan M., Hinz Wolfgang, Rearick Todd M., Schultz Jonathan, Mileski William, Davey Mel, Leamon John H., Johnson Kim, Milgrew Mark J., Edwards Matthew, Hoon Jeremy, Simons Jan F., Marran David, Myers Jason W., Davidson John F., Branting Annika, Nobile John R., Puc Bernard P., Light David, Clark Travis A., Huber Martin, Branciforte Jeffrey T., Stoner Isaac B., Cawley Simon E., Lyons Michael, Fu Yutao, Homer Nils, Sedova Marina, Miao Xin, Reed Brian, Sabina Jeffrey, Feierstein Erika, Schorn Michelle, Alanjary Mohammad, Dimalanta Eileen, Dressman Devin, Kasinskas Rachel, Sokolsky Tanya, Fidanza Jacqueline A., Namsaraev Eugeni, McKernan Kevin J., Williams Alan, Roth G. Thomas, Bustillo James. An

integrated semiconductor device enabling non-optical genome sequencing // Nature. jul 2011. 475, 7356. 348–352.

Rubin-Blum Maxim, Dubilier Nicole, Kleiner Manuel. Genetic Evidence for Two Carbon Fixation Pathways (the Calvin-Benson-Bassham Cycle and the Reverse Tricarboxylic Acid Cycle) in Symbiotic and Free-Living Bacteria // mSphere. feb 2019. 4, 1.

Saary Paul, Mitchell Alex L., Finn Robert D. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC // Genome Biology. dec 2020. 21, 1. 244.

Saiki Randall K, Gelfand David H, Stoffel Susanne, Scharf Stephen J, Higuchi Russell, Horn Glenn T, Mullis Kary B, Erlich Henry A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase // Science. jan 1988. 239, 4839. 487–491.

Sanger F., Brownlee G. G., Barrell B. G. A two-dimensional fractionation procedure for radioactive nucleotides // Journal of Molecular Biology. sep 1965. 13, 2. 373–398.

Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. // Proceedings of the National Academy of Sciences of the United States of America. dec 1977. 74, 12. 5463–5467.

Santoro A. E., Kellom M., Laperriere S. M. Contributions of single-cell genomics to our understanding of planktonic marine archaea // Philosophical Transactions of the Royal Society B: Biological Sciences. nov 2019. 374, 1786. 20190096.

Santoro Ana Lúcia, Bastviken David, Gudasz Cristian, Tranvik Lars, Enrich-Prast Alex. Dark Carbon Fixation: An Important Process in Lake Sediments // PLoS ONE. jun 2013. 8, 6. e65813.

Sarbu S. M., Kinkle B. K., Vlasceanu L., Kane T. C., Popa R. Microbiological characterization of a sulfide-rich groundwater ecosystem // Geomicrobiology Journal. jul 1994. 12, 3. 175–182.

Sarbu Serban M., Kane Thomas C., Kinkle Brian K. A chemoautotrophically based cave ecosystem // Science. 1996. 272, 5270. 1953–1954.

Sarbu Serban M., Lascu Cristian. Condensation corrosion in Movile cave, Romania // Journal of Cave and Karst Studies. 1997. 59, 3. 99–102.

Sauvadet Anne Laure, Gobet Angélique, Guillou Laure. Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats // Environmental Microbiology. nov 2010. 12, 11. 2946–2964.

Schadt Eric E., Turner Steve, Kasarskis Andrew. A window into third-generation sequencing // Human Molecular Genetics. oct 2010. 19, R2. R227–R240.

Schena Mark, Shalon Dari, Davis Ronald W, Brown Patrick O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray // Science. oct 1995. 270, 5235. 467–470.

Schirmer Melanie, Ijaz Umer Z., D'Amore Rosalinda, Hall Neil, Sloan William T., Quince Christopher. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform // Nucleic Acids Research. mar 2015. 43, 6. e37–e37.

Schmid Martin, Budnev Nikolay M., Granin Nick G., Sturm Michael, Schurter Michael, Wüest Alfred. Lake Baikal deepwater renewal mystery solved // Geophysical Research Letters. may 2008. 35, 9. L09605.

Schoenle Alexandra, Nitsche Frank, Werner Jennifer, Arndt Hartmut. Deep-sea ciliates: Recorded diversity and experimental studies on pressure tolerance // Deep-Sea Research Part I: Oceanographic Research Papers. oct 2017. 128. 55–66.

Sczyrba Alexander, Hofmann Peter, Belmann Peter, Koslicki David, Janssen Stefan, Dröge Johannes, Gregor Ivan, Majda Stephan, Fiedler Jessika, Dahms Eik, Bremges Andreas, Fritz Adrian, Garrido-Oter Ruben, Jørgensen Tue Sparholt, Shapiro Nicole, Blood Philip D, Gurevich Alexey, Bai Yang, Turaev Dmitrij, Demaere Matthew Z., Chikhi Rayan, Nagarajan Niranjan, Quince Christopher, Meyer Fernando, Balvočiutė Monika, Hansen Lars Hestbjerg, Sørensen Søren J, Chia Burton K.H., Denis Bertrand, Froula Jeff L, Wang Zhong, Egan Robert, Don

Kang Dongwan, Cook Jeffrey J, Deltel Charles, Beckstette Michael, Lemaitre Claire, Peterlongo Pierre, Rizk Guillaume, Lavenier Dominique, Wu Yu Wei, Singer Steven W, Jain Chirag, Strous Marc, Klingenberg Heiner, Meinicke Peter, Barton Michael D, Lingner Thomas, Lin Hsin Hung, Liao Yu Chieh, Silva Genivaldo Gueiros Z, Cuevas Daniel A, Edwards Robert A, Saha Surya, Piro Vitor C, Renard Bernhard Y, Pop Mihai, Klenk Hans Peter, Göker Markus, Kyrpides Nikos C, Woyke Tanja, Vorholt Julia A, Schulze-Lefert Paul, Rubin Edward M, Darling Aaron E, Rattei Thomas, McHardy Alice C. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software // Nature Methods. nov 2017. 14, 11. 1063–1071.

Sedlar Karel, Kupkova Kristyna, Provaznik Ivo. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics // Computational and Structural Biotechnology Journal. 2017. 15. 48–55.

Seitz Kiley W., Dombrowski Nina, Eme Laura, Spang Anja, Lombard Jonathan, Sieber Jessica R., Teske Andreas P., Ettema Thijs J.G., Baker Brett J. Asgard archaea capable of anaerobic hydrocarbon cycling // Nature Communications. dec 2019. 10, 1. 1822.

Semenova E A, Kuznedelov K D. Species diversity of picoplankton from Lake Baikal by comparative analysis of 5'-terminal segments of 16S rRNA genes // Molekuliarnaia biologiia. 1998. 32, 5. 895–901.

Shade Ashley. Diversity is the question, not the answer. jan 2017. 1–6.

Shagin Dmitriy A., Shagina Irina A., Zaretsky Andrew R., Barsova Ekaterina V., Kelmanson Ilya V., Lukyanov Sergey, Chudakov Dmitriy M., Shugay Mikhail. A high-throughput assay for quantitative measurement of PCR errors // Scientific Reports. dec 2017. 7, 1. 2718.

Sharpton Thomas J. An introduction to the analysis of shotgun metagenomic data. jun 2014. 209.

Shendure Jay, Ji Hanlee. Next-generation DNA sequencing. oct 2008. 1135–1145.

Shendure Jay, Porreca Gregory J, Reppas Nikos B, Lin Xiaoxia, McCutcheon John P, Rosenbaum Abraham M, Wang Michael D, Zhang Kun, Mitra Robi D, Church George M. Molecular biology: Accurate multiplex polony sequencing of an evolved bacterial genome // Science. sep 2005. 309, 5741. 1728–1732.

Sherstyankin P. P., Alekseev S. P., Abramov A. M., Stavrov K. G., De Batist M., Hus R., Canals M., Casamor J. L. Computer-based bathymetric map of Lake Baikal // Doklady Earth Sciences. may 2006. 408, 4. 564–569.

Shimaraev M. N., Gnatovskii R. Yu., Blinov V. V., Ivanov V. G. Renewal of deep waters of Lake Baikal revisited // Doklady Earth Sciences. may 2011. 438, 1. 652–655.

Shiratori Takashi, Suzuki Shigekatsu, Kakizawa Yukako, Ishida Ken ichiro. Phagocytosis-like cell engulfment by a planctomycete bacterium // Nature Communications. dec 2019. 10, 1. 5529.

Shubenkova O. V., Zemskaya T. I., Chernitsyna S. M., Khlystov O. M., Triboi T. I. The first results of an investigation into the phylogenetic diversity of microorganisms in southern Baikal sediments in the region of subsurface depositions of methane hydrates // Mikrobiologiya. 2005. 74, 3. 370–377.

Sibbald Shannon J., Archibald John M. More protist genomes needed // Nature Ecology & Evolution. apr 2017. 1, 5. 0145.

Simon Carola, Daniel Rolf. Achievements and new knowledge unraveled by metagenomic approaches. nov 2009. 265–276.

Singer Esther, Bushnell Brian, Coleman-Derr Devin, Bowman Brett, Bowers Robert M, Levy Asaf, Gies Esther A, Cheng Jan Fang, Copeland Alex, Klenk Hans Peter, Hallam Steven J, Hugenholtz Philip, Tringe Susannah G, Woyke Tanja. High-resolution phylogenetic microbial community profiling // ISME Journal. 2016. 10, 8. 2020–2032.

Sleator R D, Shortall C, Hill C. Metagenomics. nov 2008. 361–366.

Smith David R. Bringing bioinformatics to the scientific masses // EMBO reports. 2018. 19, 6.

Smith David Roy. Broadening the definition of a bioinformatician. aug 2015. 258.

Sogin S. J., Sogin M. L., Woese C. R. Phylogenetic measurement in procaryotes by primary structural characterization // Journal of Molecular Evolution. jun 1972. 1, 2. 173–184.

Solden Lindsey, Lloyd Karen, Wrighton Kelly. The bright side of microbial dark matter: Lessons learned from the uncultivated majority. jun 2016. 217–226.

Spang Anja, Saw Jimmy H., Jørgensen Steffen L., Zaremba-Niedzwiedzka Katarzyna, Martijn Joran, Lind Anders E., Van Eijk Roel, Schleper Christa, Guy Lionel, Ettema Thijs J.G. Complex archaea that bridge the gap between prokaryotes and eukaryotes // Nature. may 2015. 521, 7551. 173–179.

Stackebrandt E., Goebel B. M. Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology // International Journal of Systematic Bacteriology. oct 1994. 44, 4. 846–849.

Stackebrandt Erko, Jonas Ebers. Taxonomic parameters revisited: tarnished gold standards // Microbiology Today. 2006. 33. 152–155.

Stanier R. Y., Niel C. B. van. The concept of a bacterium // Archiv für Mikrobiologie. 1962. 42, 1. 17–35.

Stelbrink Björn, Shirokaya Alena A., Clewing Catharina, Sitnikova Tatiana Y., Prozorova Larisa A., Albrecht Christian. Conquest of the deep, old and cold: An exceptional limpet radiation in Lake Baikal // Biology Letters. jul 2015. 11, 7. 20150321.

Stieglmeier Michaela, Alves Ricardo J.E., Schleper Christa. The phylum thaumarchaeota // The Prokaryotes: Other Major Lineages of Bacteria and The Archaea. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. 347–362.

Stoddard Steven F., Smith Byron J., Hein Robert, Roller Benjamin R.K., Schmidt Thomas M. rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a

new foundation for future development // Nucleic Acids Research. jan 2015. 43, D1. D593–D598.

Strasser Bruno J. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965 // Journal of the History of Biology. dec 2010. 43, 4. 623–660.

Thauer Rudolf K. A fifth pathway of carbon fixation. dec 2007. 1732–1733.

Tikhonov Mikhail, Leach Robert W, Wingreen Ned S. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution // ISME Journal. jan 2015. 9, 1. 68–80.

Torti Andrea, Jørgensen Bo Barker, Lever Mark Alexander. Preservation of microbial DNA in marine sediments: insights from extracellular DNA pools // Environmental Microbiology. dec 2018. 20, 12. 4526–4542.

Torti Andrea, Lever Mark Alexander, Jørgensen Bo Barker. Origin, dynamics, and implications of extracellular DNA pools in marine sediments. dec 2015. 185–196.

Touchart Laurent. Baikal, lake // Encyclopedia of Earth Sciences Series. 2012. 83–91.

Troitskaya Elena, Blinov Vadim, Ivanov Vyacheslav, Zhdanov Andrey, Gnatovsky Ruslan, Sutyrina Ekaterina, Shimaraev Mikhail. Cyclonic circulation and upwelling in Lake Baikal // Aquatic Sciences. apr 2015. 77, 2. 171–182.

Tsimitri Chrysanthi, Rockel Burkhardt, Wüest Alfred, Budnev Nikolay M., Sturm Michael, Schmid Martin. Drivers of deep-water renewal events observed over 13 years in the South Basin of Lake Baikal // Journal of Geophysical Research: Oceans. mar 2015. 120, 3. 1508–1526.

Tyson Gene W., Chapman Jarrod, Hugenholtz Philip, Allen Eric E., Ram Rachna J., Richardson Paul M., Solovyev Victor V., Rubin Edward M., Rokhsar Daniel S., Banfield Jillian F. Community structure and metabolism through reconstruction of microbial genomes from the environment // Nature. mar 2004. 428, 6978. 37–43.

Venter J Craig, Remington Karin, Heidelberg John F, Halpern Aaron L, Rusch Doug, Eisen Jonathan A, Wu Dongying, Paulsen Ian, Nelson Karen E., Nelson William, Fouts Derrick E, Levy Samuel, Knap Anthony H, Lomas Michael W, Nealson Ken, White Owen, Peterson Jeremy, Hoffman Jeff, Parsons Rachel, Baden-Tillson Holly, Pfannkoch Cynthia, Rogers Yu Hui, Smith Hamilton O. Environmental Genome Shotgun Sequencing of the Sargasso Sea // Science. apr 2004. 304, 5667. 66–74.

Vieites José M., Guazzaroni María Eugenia, Beloqui Ana, Golyshin Peter N., Ferrer Manuel. Metagenomics approaches in systems microbiology. jan 2009. 236–255.

Vincent Antony T., Charette Steve J. Who qualifies to be a bioinformatician? apr 2015. 164.

Vollmers John, Wiegand Sandra, Kaster Anne Kristin. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! 2017.

Vuillemin Aurèle, Friese André, Alawi Mashal, Henny Cynthia, Nomosatryo Sulung, Wagner Dirk, Crowe Sean A., Kallmeyer Jens. Geomicrobiological Features of Ferruginous Sediments from Lake Towuti, Indonesia // Frontiers in Microbiology. jun 2016. 7. 1007.

Wang Yu, Sheng Hua Fang, He Yan, Wu Jin Ya, Jiang Yun Xia, Tam Nora Fung Yee, Zhou Hong Wei. Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags // Applied and Environmental Microbiology. dec 2012. 78, 23. 8264–8271.

Ribosomal RNA Analysis of Microorganisms as They Occur in Nature. // . 1992. 219–286.

Watson J. D., Crick F. H.C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid // Nature. apr 1953. 171, 4356. 737–738.

Wayne L. G., Brenner D. J., Colwell R. R., Grimont P. A. D., Kandler O., Krichevsky M. I., Moore L. H., Moore W. E. C., Murray R. G. E., Stackebrandt E., Starr M. P., Truper H. G. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics // International Journal of Systematic and Evolutionary Microbiology. oct 1987. 37, 4. 463–464.

Weibel Douglas B., DiLuzio Willow R., Whitesides George M. Microfabrication meets microbiology. mar 2007. 209–218.

Weiss Sophie, Xu Zhenjiang Zech, Peddada Shyamal, Amir Amnon, Bittinger Kyle, Gonzalez Antonio, Lozupone Catherine, Zaneveld Jesse R, Vázquez-Baeza Yoshiki, Birmingham Amanda, Hyde Embriette R, Knight Rob. Normalization and microbial differential abundance strategies depend upon data characteristics // Microbiome. 2017. 5, 1.

West Patrick T., Probst Alexander J., Grigoriev Igor V., Thomas Brian C., Banfield Jillian F. Genome-reconstruction for eukaryotes from complex natural microbial communities // Genome Research. apr 2018. 28, 4. 569–580.

Wetlaufer D B. Nucleation, rapid folding, and globular intrachain regions in proteins. // Proceedings of the National Academy of Sciences of the United States of America. mar 1973. 70, 3. 697–701.

Whitman William B., Coleman David C., Wiebe William J. Prokaryotes: The unseen majority. jun 1998. 6578–6583.

Whittaker R. H. EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY // TAXON. may 1972. 21, 2-3. 213–251.

Williams Tom A., Cox Cymon J., Foster Peter G., Szöllősi Gergely J., Embley T. Martin. Phylogenomics provides robust support for a two-domains tree of life // Nature Ecology and Evolution. jan 2020. 4, 1. 138–147.

Willis Amy D. Rarefaction, alpha diversity, and statistics // Frontiers in Microbiology. oct 2019. 10, OCT. 2407.

Wischer Daniela, Kumaresan Deepak, Johnston Antonia, El Khawand Myriam, Stephenson Jason, Hillebrand-Voiculescu Alexandra M., Chen Yin, Murrell J. Colin. Bacterial metabolism of methylated amines and identification of novel methylotrophs in Movile Cave // ISME Journal. 2015. 9, 1. 195–206.

Woese C R, Fox G E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms // Proceedings of the National Academy of Sciences of the United States of America. nov 1977. 74, 11. 5088–5090.

Woese Carl, Sogin Mitchell, Stahl David, Lewis Bobby Joe, Bonen Linda. A comparison of the 16S ribosomal RNAs from mesophilic and thermophilic bacilli: Some modifications in the sanger method for RNA sequencing // Journal of Molecular Evolution. apr 1976. 7, 3. 197–213.

Woese Carl R., Fox GEORGE E., Zablen Lawrence, Uchida Tsuneko, Bonen Linda, Pechman Kenneth, Lewis Bobby J., Stahl David. Conservation of primary structure in 16S ribosomal RNA // Nature. mar 1975. 254, 5495. 83–86.

Wrighton Kelly C, Thomas Brian C, Sharon Itai, Miller Christopher S, Castelle Cindy J, VerBerkmoes Nathan C, Wilkins Michael J, Hettich Robert L, Lipton Mary S, Williams Kenneth H, Long Philip E, Banfield Jillian F. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla // Science. sep 2012. 337, 6102. 1661–1665.

Wu Dongying, Wu Martin, Halpern Aaron, Rusch Douglas B., Yooseph Shibu, Frazier Marvin, Venter J. Craig, Eisen Jonathan A. Stalking the fourth domain in metagenomic data: Searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees // PLoS ONE. mar 2011. 6, 3. e18011.

Wu Kaiyuan, Zhao Wenqian, Wang Qian, Yang Xiangdong, Zhu Lifeng, Shen Ji, Cheng Xiaoying, Wang Jianjun. The relative abundance of benthic bacterial phyla along a water-depth gradient in a plateau lake: Physical, chemical, and biotic drivers // Frontiers in Microbiology. jul 2019. 10, JULY. 1521.

Yi Zhenzhen, Berney Cedric, Hartikainen Hanna, Mahamdallie Shazia, Gardner Michelle, Boenigk Jens, Cavalier-Smith Thomas, Bass David. High-throughput sequencing of microbial eukaryotes in Lake Baikal reveals ecologically differentiated communities and novel evolutionary radiations // FEMS microbiology ecology. aug 2017. 93, 8.

Ynalvez Ruby A, Dinamarca Jorge, Moroney James V. Algal Photosynthesis // eLS. Chichester, UK: John Wiley & Sons, Ltd, nov 2018. 1–9.

Yoon Hwan Su, Price Dana C, Stepanauskas Ramunas, Rajah Veeran D, Sieracki Michael E, Wilson William H, Yang Eun Chan, Duffy Siobain, Bhattacharya Debashish. Single-cell genomics reveals organismal interactions in uncultivated marine protists // Science. may 2011. 332, 6030. 714–717.

Yu Sherbakov Dmitrii. Molecular phylogenetic studies on the origin of biodiversity in Lake Baikal // Trends in Ecology and Evolution. mar 1999. 14, 3. 92–95.

Yubuki Naoji, Galindo Luis Javier, Reboul Guillaume, López-García Purificación, Brown Matthew W., Pollet Nicolas, Moreira David. Ancient Adaptive Lateral Gene Transfers in the Symbiotic Opalina-Blastocystis Stramenopile Lineage // Molecular Biology and Evolution. nov 2020. 37, 3. 651–659.

Zaremba-Niedzwiedzka Katarzyna, Caceres Eva F., Saw Jimmy H., Bäckström Disa, Juzokaite Lina, Vancaester Emmelien, Seitz Kiley W., Anantharaman Karthik, Starnawski Piotr, Kjeldsen Kasper U., Stott Matthew B., Nunoura Takuro, Banfield Jillian F., Schramm Andreas, Baker Brett J., Spang Anja, Ettema Thijs J.G. Asgard archaea illuminate the origin of eukaryotic cellular complexity // Nature. jan 2017. 541, 7637. 353–358.

Zemskaya Tamara I., Cabello-Yeves Pedro J., Pavlova Olga N., Rodriguez-Valera Francisco. Microorganisms of Lake Baikal—the deepest and most ancient lake on Earth // Applied Microbiology and Biotechnology. may 2020. 1–12.

Zemskaya Tamara I., Lomakina Anna V., Mamaeva Elena V., Zakharenko Alexandra S., Pogodaeva Tatyana V., Petrova Darya P., Galachyants Yuri P. Bacterial communities in sediments of Lake Baikal from areas with oil and gas discharge // Aquatic Microbial Ecology. oct 2015. 76, 2. 95–109.

Zemskaya Tamara I., Pogodaeva Tatiayna V., Shubenkova Olga V., Chernitsina Svetlana M., Dagurova Olga P., Buryukhaev Savelii P., Namsaraev Bair B., Khlystov Oleg M., Egorov Alek-

sandr V., Krylov Aleksei A., Kalmychkov Gennadii V. Geochemical and microbiological characteristics of sediments near the Malenky mud volcano (Lake Baikal, Russia), with evidence of Archaea intermediate between the marine anaerobic methanotrophs ANME-2 and ANME-3 // Geo-Marine Letters. jun 2010. 30, 3-4. 411–425.

Zinger Lucie, Bonin Aurélie, Alsos Inger G., Bálint Miklós, Bik Holly, Boyer Frédéric, Chariton Anthony A., Creer Simon, Coissac Eric, Deagle Bruce E., De Barba Marta, Dickie Ian A., Dumbrell Alex J., Ficetola Gentile Francesco, Fierer Noah, Fumagalli Luca, Gilbert M. Thomas P., Jarman Simon, Jumpponen Ari, Kauserud Håvard, Orlando Ludovic, Pansu Johan, Pawlowski Jan, Tedersoo Leho, Thomsen Philip Francis, Willerslev Eske, Taberlet Pierre. DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. apr 2019. 1857–1862.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre: Utilisation d'approches de métabarcoding et de métagénomique pour l'analyse de communautés microbiennes suboxiques

Mots clés: métabarcoding, métagénomique, environement suboxiques, lac Baikal, cave Movile, écologie microbienne

Résumé: L'écologie microbienne concerne l'étude des microorganismes et de leurs interactions biotiques et abiotiques dans un écosystème donné. Ces vingt dernières années, l'avancement des techniques moléculaires pour analyser la diversité microbienne et, notamment, les nouvelles technologies de séquençages (NGS) ont permis de surmonter les limitations associées aux approches traditionnelles basées sur la culture et la microscopie. Ces approches moléculaires ont conduit à une accumulation des données de diversité microbienne et de potentiel métabolique dans des communautés microbiennes des écosystèmes variés. Cependant, ces efforts ont été principalement appliqués sur des environnements facilement accessibles ou liés à l'humain, comme le plancton (marin principalement) et la flore intestinale. Néanmoins, ceci a conduit à une très forte augmentation de données environnementales et au développement de la bioinformatique par le biais de nombreux outils. Parmi les environnements délaissés des études, les environnements faibles en oxygène sont probablement également porteurs de nouveautés phylogénique ou métaboliques. Afin de palier à cela, nous avons choisi d'explorer deux environnements suboxiques relativement peu étudiés : la cave Movile (Roumanie) et les sédiments du lac Baikal (Sibérie, Russie). Notre but étant de montrer les diversités phylogénétiques et fonctionnelles des microbes de ces biotopes. Pour cela, j'ai d'abord développé un pipeline d'analyse de données métabarcoding (petite sous-unités ribosomique). Ensuite, j'ai appliqué cet outil sur des données de métabarcoding de protistes provenant d'échantillons d'eau et de tapis microbiens de la cave de Movile, un écosystème chemosynthétique pratiquement fermé. Nous avons montré que la diversité des protistes de la cave s'étendait à quasiment tous les grands groupes eucaryotes et provenait à la fois d'origine d'eaux douces et marines. De plus, la plupart ont été affiliées à des groupes d'organismes typiquement anaérobies, ce qui est concordant avec les paramètres abiotiques de la cave. Écologiquement, ces protistes sont des prédateurs mais aussi vraisemblablement des partenaires symbiotiques avec des espèces procaryotes de la cave. Dans une deuxième étude, j'ai eu l'opportunité d'appliquer ce pipeline de métabarcoding sur des données procaryotes et eucaryotes provenant des couches superficielles des sédiments du lac d'eau douce Baikal. Comme attendu, les communautés microbiennes dans ces sédiments sont particulièrement diverses et relativement enrichis en archées. Nous avons aussi pu mettre en évidence des lignées que l'on pensait exclusivement marines dans ces sédiments. Ces lignées sont probablement planctoniques mais s'accumulent au fond par sédimentation. Enfin, les échantillons ont été prélevés dans le but de tester les influences de la profondeur, du bassin et de la latitude sur les communautés. Aucune d'elles ne s'est révélée significative. Dans une troisième étude, j'ai utilisé une approche métagénomique afin de révéler les acteurs écologiquement majeurs dans les sédiments, leurs rôles et de reconstruire leurs génomes. Cela nous a permis notamment de mettre en évidence le rôle primordial des Thaumarchaeota dans le cycle de l'azote et la production primaire de molécules de carbone. Les chloroflexi et les protéobacteries ont aussi un rôle important dans la surface des sédiments du lac Baikal. Ce travail de thèse participe à la connaissance globale de la diversité microbienne sur la planète en mettant en lumière des environnements peu étudiés. De plus, l'étude de la surface des sédiments du lac Baikal apporte de nouvelles données sur le sujet de la transition eau douces/eau marines des microbes. Enfin, la métagénomique a permis de révéler le cycle des nutriments et les microorganismes y participant dans ces échantillons de sédiment. En résumé, ce travail vient mettre en lumière l'écologie microbienne d'écosystèmes suboxiques, notamment la surface des sédiments du lac Baikal.

Title: Metabarcoding and metagenomic approaches to decipher microbial communities in suboxic environments

Keywords: metabarcoding; metagenomic; suboxic environments; lake Baikal; cave Movile; microbial ecology

Abstract: Microbial ecology is the science of micro-organisms and their biotic and abiotic interactions in a given ecosystem. As technology has advanced, molecular techniques have been widely used to overcome the limitations of classical approaches such as culturing and microscopy. Indeed, the development of Next Generation Sequencing (NGS) technologies in the past twenty years has largely helped to unravel the phylogenetic diversity and functional potential of microbial communities across ecosystems. Nonetheless, most of the environments studied through these techniques concentrated on relatively easily accessible, tractable and host-related ecosystems such as plankton (especially in marine ecosystems), soils and gut microbiomes. This has contributed to the rapid accumulation of a wealth of environmental diversity and metagenomic data along with advances in bioinformatics leading to the development of myriads of tools. Oxygen-depleted environments and especially their microbial eukaryote components are less studied and may lead to future phylogenetic and metabolic discoveries. In order to address this, we conducted analyses on two poorly studied suboxic ecosystems: Movile Cave (Romania) and lake Baikal sediments (Siberia, Russia). In this task, we aimed at unveiling the taxonomic and functional diversity of microorganims in these environments. To do so, I first evaluated the available bioinformatics tools and implemented a bioinformatics pipeline for 16S/18S rRNA gene-based metabarcoding analysis, making reasoned methodological choices. Then, as a case study, I carried out metabarcoding analyses of the water and floating microbial mats found in Movile Cave in order to investigate its protist diversity. Our study showed that Movile Cave, a sealed off chemosynthetic ecosystem, harbored a substantial protist diversity with species spanning most of the major eukaryotic super groups. The majority if these protists were related to species of freshwater and marine origins. Most of them were putatively anaerobic, in line with the cave environment, and suggesting that in addition to their predatory role, they might participate in prokaryote-protist symbioses. In a second study, I applied my metabarcoding pipeline to explore unique and relatively unexplored environment of Lake Baikal sediments. I first applied a metabarcoding approach using 16S and 18S rRNA genes to describe prokaryotic as well as protist diversity. Overall, the communities within these ecosystems were very diverse and enriched in ammonia-oxidizing Thaumarchaeota. We also identified several typical marine taxa which are likely planktonic but accumulate in sediments. Finally, our sampling plan allowed us to test whether differences across depth, basin or latitude affected microbial community structure. Our results showed that the composition of sediment microbial communities remained relatively stable across the samples regardless of depth or latitude. In a third study, we applied metagenomics to study the metabolic potential of communities associated to Baikal sediments and to reconstruct metagenome-assembled genomes (MAGs) of dominant organisms. This revealed the considerable ecological importance of Thaumarchaeota lineages in lake Baikal sediments, which were found to be the major autotrophic phyla and also very implicated in the nitrogen cycle. Chloroflexi and Proteobacteria-related species also appeared ecologically important. This PhD thesis reveals the taxonomic diversity of poorly studied suboxic ecosystems and therefore contributes to our knowledge of microbial diversity on Earth. Additionally, the analyses of surface sediment samples in lake Baikal adds new light on freshwater-marine transitions. The metagenomic analyses reported here allowed us to postulate a model of nutrient cycle carried out by microorganismsin these sediments. Overall, this work sheds light on the microbial ecology of oxygen-depleted environments, and most notably lake Baikal surface sediments.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France