UNIVERSITÉ DU QUÉBEC À MONTRÉAL

PROCESSUS D'ASSEMBLAGE DES COMMUNAUTÉS MICROBIENNES ENDOLITHIQUES ET SESSILES DE LA SUBSURFACE TERRESTRE SUPERFICIELLE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN BIOLOGIE

PAR
JULIA ARIELLE MEYER

AVRIL 2022

REMERCIEMENTS

Tout d'abord, je tiens à remercier ma directrice de recherche Dre. Cassandre Lazar de m'avoir permis d'intégrer son laboratoire de recherche. Je te remercie pour ta bienveillance et ton soutien. Ton enthousiasme et ton intérêt envers les communautés microbiennes de la subsurface terrestre m'ont inspirés ! À tes côtés, j'ai appris énormément et j'en suis très reconnaissante.

Je remercie également les membres de mon laboratoire : Jean-Christophe Gagnon, Karine Villeneuve et Benjamin Groult. J'ai apprécié votre compagnie, nos échanges et nos moments de réflexion. Mention spéciale à Benjamin pour les excellents schémas de la subsurface ! J'aimerais également remercier toutes les personnes qui ont participé de près ou de loin à l'acquisition de mes données : un grand merci à Marie Larocque et son équipe de nous avoir permis de participer aux opérations de forage sans lesquelles nous n'aurions jamais pu récupérer les échantillons provenant de la subsurface terrestre superficielle. Je remercie également Sheri Zahkari de les avoir récoltés avec grand soin ! Merci à Agnieska Adamowicz et à Jean-François Hélie pour l'encadrement des analyses de carbone total, d'azote total, de carbone organique total et de carbone inorganique total réalisées au laboratoire de géochimie des isotopes stables légers. Merci à Jean-Carlos Montero Serrano pour son interprétation semi-quantitative des principaux composants minéralogiques des échantillons. Je tiens également à remercier Geneviève Bourret pour son accompagnement et son aide dans mon projet de séquençage. Tu as toujours été sympathique et tu as répondu avec une grande patience à mes nombreuses questions concernant les PCR et le séquençage.

Je remercie finalement ma famille et mes amis. Vous avez rendu ces deux années beaucoup plus agréables ! Une mention spéciale à mes sœurs. Vous avez su me motiver même pendant la pandémie ! J'ai adoré nos séances de travail et nos échanges de savoirs (et de script R !).

TABLE DES MATIÈRES

LISTE DES FIGURES v

LISTE DES TABLEAUX vii

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES viii

LISTE DES SYMBOLES ET DES UNITÉS x

RÉSUMÉ 1

INTRODUCTION 1

CHAPITRE I ARTICLE DE RECHERCHE : CHANGE IN DIVERSITY, COMPOSITION, AND ABSOLUTE ABUNDANCE OF SESSILE AND ENDOLITHIC BACTERIAL, ARCHAEAL, AND EUKARYOTIC COMMUNITIES WITH SOIL DEPTH IN TWO SOIL PROFILES FROM THE SHALOW TERRESTRIAL SUBSURFACE IN THE LAURENTIANS (QC, CANADA) 8

1.1 Introduction 9

1.2 Materials and Methods 11

1.2.1 Study Sites 11
1.2.2 Sampling 12
1.2.3 Chemical Characteristics of Geological Material and Water Samples 14
1.2.4 DNA Extraction 14
1.2.5 PCR Amplification and Illumina Sequencing 15
1.2.6 Digital PCR 15
1.2.7 Sequence Processing 16
1.2.8 Statistical Analysis 16

1.3 Results 18

1.3.1 Variation of Chemical Characteristics with Depth 18
1.3.2 Taxonomic α-Diversity : Variation with Depth and Correlation with Chemical Characteristics of Geological Material 20
1.3.3 Phylogenetic α-Diversity and Correlation with Geological Material Characteristics 21
1.3.4 β-Diversity and Correlation with Geological Material Characteristics 21
1.3.5 Absolute Bacterial Abundance 23
1.3.6 Variation in the Relative Abundance of Dominant Bacteria, Arhaea, and Eukaryote Microorganisms at the Phylum and Genus Levels, with Depth 25
1.3.7 Interactions among Bacterial, Eukaryotic and Archaeal major ASVs 28
1.3.8 Microbial Source Tracking 29

1.4 Discussion 31

1.4.1 The Effect of Depth on Endolithic and Sessile Communities 31
1.4.2 Effect of Abiotic Characteristics on Endolithic and Sessile Communities 33
1.4.3 Potential Biotic Interactions between Bacterial, Eukaryotic and Archaeal Microorganisms 35
1.4.4 Vertical Fluid Fluxes as Sources of Microbial Communities in the Shallow Terrestrial Subsurface 36

1.5 Conclusions.....38

DISCUSSION GÉNÉRALE.....41

Annexe A : SUPPLEMENTARY FIGURES.....44

Annexe B : SUPPLEMENTARY TABLES.....52

Annexe C : SCRIPTS.....64

RÉFÉRENCES.....125

LISTE DES FIGURES

Figure 0.1. Représentation schématique de la subsurface terrestre superficielle et des différents types de communautés de microorganismes qui vivent dans cet habitat (B.Groult). 2

Figure 1.1 Location of the two study sites. Both sites are in the northwest of Montreal in the Laurentians region (Quebec, Canada). 12

Figure 1.2. Diagrams of the 15.2 cm diameter boreholes, showing the depth at which samples of rock-associated (red crosses, sediments samples during drilling) and planktonic (blue crosses, samples in the well water after drilling) subsurface communities were taken (a) at site 1 (Ascension) and (b) site 2 (Notre-Dame-du-Laus). 13

Figure 1.3. Chemical characteristics of geological material along depth at site 1 (black circles) and site 2 (grey squares) with pH (a), total carbon (TC) (b), total organic carbon (TOC) (c), and total inorganic carbon (TIC) (d) along depth. Surface sample at site 1 was removed from the TC, TOC and TIC figures for a better visual representation. Bedrock samples are represented by hollow symbols. pH of the groundwater samples (located at 18 m and 1.75 m of depth, at site 1 and 2 respectively) are represented by stars in (a). For full details on the changes of soil characteristics, see Table S1 and S2. 19

Figure 1.4. Mineral composition along depth at site 1 (a) and at site 2 (b). OM, organic matter. Bedrock samples are indicated by hollow symbols. For full details on the mineralogical composition, see Table S4 and S5. 20

Figure 1.5. Microbial community compositional structure in the subsurface samples as indicated by non-metric multidimensional scaling plots (nMDS) on Bray-Curtis dissimilarity matrices. Geological material textures are color coded and depth is indicated for each sample (subscript). With: (a) bacterial composition at site 1, stress =0.113; (b), bacterial composition at site 2, stress= 0.159; (c) archaeal composition at site 1, stress=0.0094; (d) archaeal composition at site 2, stress=0.137; (e) eukaryotic composition at site 1, stress=0.0107; (f) eukaryotic composition at site 2, stress =0.133. mS, medium sand; mcSG, medium to coarse sand and gravel; cSG, coarse sand and gravel; R, Bedrock; fmS, fine to medium sand; fSSIC, fine sand with silt and clay; fvfS, fine to very fine sand; SG, sand and gravel; SI, silt; SIC, Silt and clay; GW, groundwater. 22

Figure 1.6. Relative abundance (%) of each microbial phylum across different depths (m) at site 1 and at site 2 . With: (a) bacterial relative abundance at site 1, (b) bacterial relative abundance at site 2, (c) archaeal relative abundance at site 1, (d) archaeal relative abundance at site 2, (e) eukaryotic relative abundance at site 1, (f) eukaryal relative abundance at site 2. Phyla with an average relative abundance of less than 1 % were categorized as “Other”. Bedrock samples are represented by hollow symbols. Groundwater samples are located at 18 m and 1.75 m of depth, at site 1 and 2 respectively and are represented by blue stars. 27

Figure 1.7. Spearman correlations between dominant bacterial, archaeal and eukaryotic ASVs. Both positive (blue) and negative (red) correlations between ASVs were calculated. Data were filtered to remove ASVs from the surface soil and the groundwater samples. Correlations are considered signifcant if $p < 0.05$ and signifcant correlations are noted with a *. 29

Figure 1.8. FEAST estimations of upper layers (source) contributing to the deeper layer (sink) across different soil depths (m) at site 1 and at site 2. With: (a, d) the sources contribution for bacterial communities; (b, e) the sources contribution for archaeal communities; (c, f) the sources contribution for eukaryotic communities The unknown sources are represented in grey. Groundwater samples are located at 18 m and 1.75 m of depth, at site 1 and 2 respectively and are represented by blue stars. 30

LISTE DES TABLEAUX

Tableau 1.I. Absolute abundances expressed in number of bacterial 16S rRNA gene copies per gram of geological material (copies g^{-1}) detected in each sample. Only values with good quality thresholds are presented. 24

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

ADN or DNA	Acide désoxyribonucléique (« Deoxyribonucleic acid»)
ANOSIM	Analyse de similarité (« Analysis of similarities »)
ASV	Variant de séquence d'amplicon (« Amplicon sequence variant »)
c.-à-d. ou i.e.	C'est-à-dire
CERMO-CF	Centre d'Excellence en Recherche sur les Maladies Orphelines – Fondation Courtois
cSG	Gravier et sable grossier (« Coarse sand and gravel »)
dPCR	PCR digitale
Etc.	Et cetera
FEAST	Fast Expectation–mAximization microbial Source Tracking
MNTD	« Mean Nearest Taxon Distance »
mS	Sable moyen (« Medium sand »)
NaH_2PO_4	Dihydrogénophosphate de sodium
Na_2HPO_4	Hydrogénophosphate de sodium
NCBI	Centre National d'Information sur la Biotechnologique
nMDS	«Non-metric multidimensional scaling »
PCR	Réaction en chaîne par polymérase
PD	Diversité Phylogénétique (« Phylogenetic Diversity »)
PES	Polyéthersulfone
P/EtOH	Tampon phosphate-éthanol

p.ex. ou *e.g*.	Par exemple
pH	Potentiel hydrogène
R	Roche mère (« Bedrock »)
rRNA	Acide ribonucléique ribosomique («ribosomal ribonucleic acid »)
ses	Tailles d'effet standardisées (« standardized effect sizes »)
SG	Gravier et sable (« Sand and gravel »)
SI	Limon (« Silt »)
SIC	Argile et limon («Silt and clay»)
TC	Carbone total («Total carbon »)
TIC	Carbone inorganique total («Total inorganic carbon »)
TN	Azote total («Total nitrogen »)
TOC	Carbone organique total («Total organic carbon »)
UCP	Production ultra propre («Ultra clean production »)
UQÀM	Université du Québec à Montréal
w/v	poids/volume (« weight/volume »)
XRD	Diffraction aux rayons X (« X-ray diffraction »)

LISTE DES SYMBOLES ET DES UNITÉS

α	Alpha
Å	Ångström
B	Beta
$CaCO_3$	Calcite
C	Carbone
Cu	Cuivre
°C	Degrés Celsius
(°, ', '')	Degré, minute, seconde
$CaMg(CO_3)_2$	Dolomite
~	Environ
g	Gramme
h	Heure
$KAlSi_3O_8$	K-feldspar
m	Mètre
µm	Micromètre
µM	Micromolaire
mL	Millilitre
min	Minute
M	Molaire

$NaAlSi_3O_8$–$CaAl_2Si_2O_8$	Plagioclase
K	Potassium
%	Pourcentage
SiO_2	Quartz
sec	Seconde
θ	Thêta
V	Volt

RÉSUMÉ

Résumé : Les communautés microbiennes jouent un rôle important dans les écosystèmes de la subsurface terrestre. La plupart des études dans cet habitat se sont concentrées sur les communautés planctoniques et ne ciblent que des groupes microbiens spécifiques (bactéries, ou archées ou eucaryotes). Par conséquent, une compréhension des processus qui régissent l'ensemble des groupes microbiens appartenant aux communautés endolithiques et sessiles fait toujours défaut. Cette étude vise à comprendre (i) l'effet de la profondeur et (ii) des facteurs biotiques sur ces communautés ; (iii) de déterminer la proportion de ces communautés qui sont formées par les mouvements de flux verticaux et (iv) de comparer leur composition avec les communautés microbiennes détectées dans l'eau de l'aquifère. Pour ce faire, nous avons collecté des échantillons dans deux profils de sol (~ 0–50 m), effectué des extractions d'ADN et un séquençage Illumina. Les résultats suggèrent que les changements dans les caractéristiques du profil des sédiments de la subsurface terrestre avec la profondeur représentent un filtre écologique et phylogénétique efficace pour la plupart des communautés d'archées et de bactéries. De plus, les mouvements de flux verticaux depuis la surface jouent un rôle majeur dans les processus d'assemblage des communautés microbiennes sous la surface. En complément, nos résultats ont montré que la plupart des interactions biotiques entre les trois domaines étaient positives, ce qui peut indiquer des associations potentielles coopératives ou mutualistes telles que l'alimentation croisée ou les relations syntrophiques dans la subsurface terrestre superficielle. Nos résultats soulignent également l'importance d'échantillonner à la fois les phases liquide et solide d'un aquifère lorsqu'il s'agit d'étudier sa microbiologie globale.

Mots-clés : subsurface terrestre; microbes de la subsurface; génomique; bactéries; archées; eucaryotes; microorganismes sessiles; microorganismes endolithiques; interactions biotiques

INTRODUCTION

Mise en contexte

Sur Terre, la plus grande diversité d'organismes vivants est retrouvée au sein des microorganismes. Omniprésents, ils vivent dans presque tous les habitats de cette planète, et sont même retrouvés dans des habitats extrêmes comme les sources chaudes, les glaciers, ou encore la subsurface terrestre (Griebler et Lueders, 2009).

Cette dernière constitue le plus vaste des écosystèmes terrestres (Edwards *et al*., 2012). Localisée sous la surface des continents, elle se situe sous les horizons du sol de surface et s'étend jusqu'à des centaines voire des milliers de mètres de profondeur (subsurface profonde) (Smith *et al*., 2018). La subsurface terrestre superficielle constitue sa zone la moins profonde et est typiquement décrite comme étant située entre les horizons du sol de surface et la roche mère (< 50 m de profondeur) (Smith *et al*., 2018). Cet environnement est composé de sédiments, de roches, de gaz et de pores au travers desquels s'écoule de l'eau souterraine qui s'accumule au niveau des aquifères (Smith et *al*., 2018; Kohlhepp *et al*., 2016) (**Figure 0.1**). C'est également un habitat complexe et actif d'un point de vue microbien. Il abrite des communautés très diverses composées principalement de bactéries, d'archées, mais également d'eucaryotes (Bomberg et Ahonen, 2017; Griebler et Lueders, 2009) qui peuvent vivre sous une forme planctonique (*c.-à-d*. flottant librement dans l'eau), sessile (*c.-à-d*. attachées à la surface des solides) (Goldsheider *et al*., 2006) ou endolithique (*c.-à-d*. vivant à l'intérieur de la matrice des roches) (Anthony *et al*., 2012) (**Figure 0.1**). Les communautés microbiennes sessiles sont actives et d'importance globale, et seraient probablement plus abondantes que les communautés planctoniques (Griebler et Lueders, 2009). Elles jouent un rôle capital pour la durabilité à long terme de la subsurface terrestre, car elles participent aux cycles des nutriments du sol et des sédiments, à la formation du sol, à la dégradation des contaminants ainsi qu'au maintien des eaux de la subsurface (Chu *et al*., 2016).

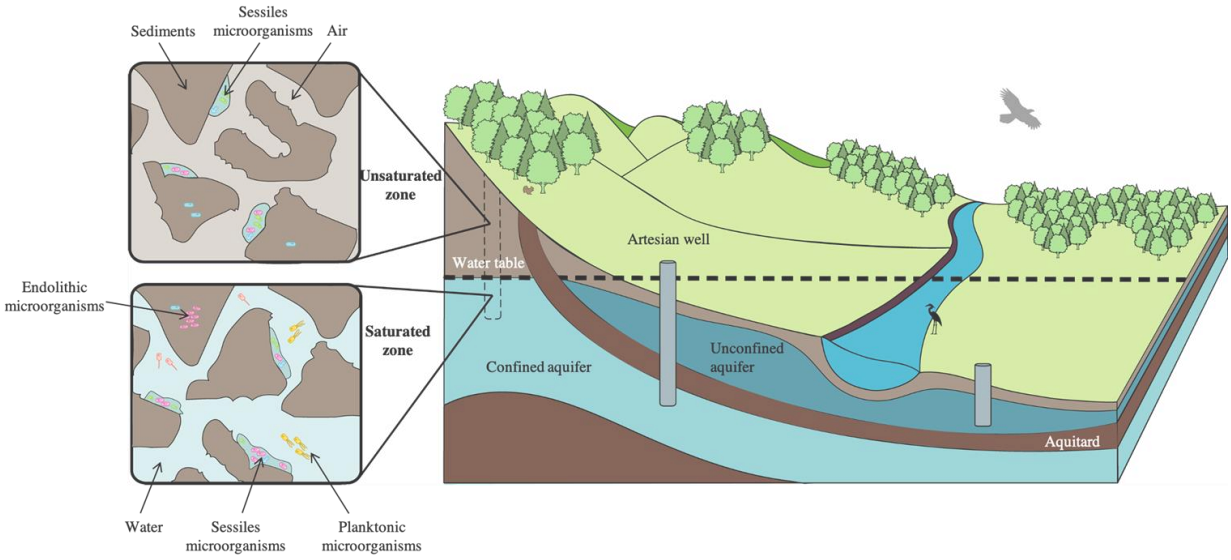


Figure 0.1. Représentation schématique de la subsurface terrestre superficielle et des différents types de communautés de microorganismes qui vivent dans cet habitat (B.Groult).

Les connaissances actuelles concernant les communautés microbiennes souterraines sont toujours à un stade précoce. Il y a seulement quelques dizaines d'années, l'idée que la distribution des microorganismes actifs était limitée au sol de surface, ou au mieux aux dix premiers mètres du sol était largement répandue au sein de la communauté scientifique (Reith, 2011). Depuis la découverte de leur existence dans les horizons profonds de la subsurface terrestre, ces dernières sont devenues de plus en plus une cible d'exploration (Colman *et al*., 2017). Cependant, la subsurface terrestre demeure l'un des habitats les moins explorés sur Terre (Alain *et al*., 2011). À ce jour, les détails sur la diversité de ces communautés, ainsi que les interactions existantes entre elles restent largement inexplorés (Govil *et al*., 2019). Encore très peu d'informations existent à propos de la formation des microbiomes de la subsurface (Herrmann *et al*., 2019), et des facteurs qui régissent l'assemblage, la structure et le maintien de ces communautés. De plus, à cause de la difficulté d'accéder aux roches et aux sédiments sous le sol, la plupart des recherches microbiennes se sont intéressées aux communautés planctoniques présentes dans l'eau de la subsurface, accessibles par des puits forés dans la subsurface (Lazar *et al*., 2019). Par conséquent, les communautés endolithiques et sessiles qui vivent sur/dans les sédiments et les roches de la subsurface ont reçu relativement peu d'intérêt (Lazar *et al*., 2019).

Questions de recherche

Ainsi, de nombreuses questions concernant l'écologie des communautés microbiennes de la subsurface terrestre restent sans réponses : (i) Quels sont les facteurs abiotiques et biotiques qui gouvernent l'assemblage des communautés microbiennes endolithiques et sessiles de la subsurface terrestre superficielle ? (ii) Quelles sont les sources/origines des microorganismes retrouvés dans la subsurface terrestre superficielle ? (iii) Quelles sont les différences et similarités entre les communautés microbiennes attachées aux sédiments et celles suspendues dans l'eau, en particulier au niveau de l'aquifère où les deux communautés sont présentes ?

État des connaissances et problématique

La diversité et la composition des communautés microbiennes présentes dans la subsurface terrestre superficielle varient considérablement en fonction des facteurs abiotiques et biotiques. À de grandes échelles spatiales, ces caractéristiques sont corrélées à des variables édaphiques, telles que le pH, la teneur en eau, les apports en nutriments et les facteurs de stress environnementaux (Smith *et al*., 2018; Andrew *et al*., 2012). Dans la subsurface terrestre superficielle, l'eau et les nutriments peuvent être fournis par les roches et sédiments qui contiennent des minéraux oxydés et réduits (Colman *et al*., 2017) ou provenir de la surface par infiltration (Satyanarayana *et al*., 2019). La matière organique enterrée en profondeur peut également servir de source de nutriments pour les microorganismes de la subsurface, mais sa présence dans les horizons profonds de la subsurface reste relativement rare (Satyanarayana *et al*., 2019). Parmi les facteurs abiotiques qui façonnent l'assemblage des communautés de la subsurface terrestre superficielle, la profondeur a reçu un intérêt grandissant depuis ces dernières années. En plus d'être un paramètre relativement simple à mesurer, il s'est révélé très efficace pour expliquer les variations de la diversité et de la composition dans la subsurface terrestre superficielle (p. ex., Kim *et al*., 2016; Eilers *et al*., 2012; Fierer *et al*., 2003; LaMontagne *et al*., 2003; Agnelli *et al*., 2004; Goberna *et al*., 2005; Will *et al*., 2010; Hansel *et al*., 2008). Ce facteur peut également être relié aux mécanismes de formation des communautés microbiennes, car les mouvements verticaux de l'eau (provenant de la surface) apportent des microorganismes dans la subsurface terrestre superficielle (Amy *et al*., 1992). Combiner les informations recueillies à partir des facteurs abiotiques avec les facteurs biotiques (p. ex., les interactions entre espèces) permettrait d'obtenir une vision approfondie sur les mécanismes régissant l'assemblage de ces communautés car notre compréhension globale sur le sujet reste limitée (Xu *et al*., 2021). Les études sur les communautés microbiennes dans la subsurface terrestre superficielle sont toujours très peu nombreuses et la plupart des recherches se sont intéressées à l'effet de la profondeur en regardant les

variations de la structure des communautés sur les premiers mètres du sol uniquement (Kim *et al*., 2016; Xu *et al*., 2021). De plus, les études dans cet habitat se sont également souvent confinées à un groupe spécifique de microorganismes (p. ex., archées, bactéries ou eucaryotes), ce qui rend difficile l'obtention d'une compréhension systémique des communautés microbiennes de la subsurface et des interactions existantes entre elles. Finalement, même si différentes caractéristiques structurelles ont été examinées, peu d'études ont étudié simultanément l'effet de la profondeur sur la diversité α, l'abondance absolue et la composition de ces communautés microbiennes (Xu *et al*., 2021).

Mis à part les mouvements d'eau provenant de la surface, les microorganismes de la subsurface terrestre superficielle peuvent s'être retrouvés dans cet habitat à travers les flux latéraux ou auraient pu coloniser les sédiments et roches entre les évènements de déposition (Amy *et al.*, 1992). Dans ce cas, ces microorganismes peuvent représenter des cellules vivantes qui ont survécu pendant de longues périodes (année jusqu'à des milliers d'années) (Amy *et al.*, 1992). Cependant, puisque les études sur l'origine des microorganismes de la subsurface sont rares, il existe très peu de données à ce sujet. L'apprentissage des origines des communautés microbiennes peut donc améliorer considérablement notre compréhension actuelle de la formation des communautés microbiennes de la subsurface superficielle, et pourrait également fournir des informations sur les conséquences de certaines activités humaines qui touchent les sols telles que les pratiques agricoles (p. ex., l' introduction de produits dans les sols) (Shenhav *et al*., 2019).

Finalement, une comparaison des communautés planctoniques par rapport aux communautés endolithiques et sessiles retrouvées au niveau de l'aquifère est nécessaire, car la plupart des études menées dans ce milieu n'ont échantillonné que les communautés planctoniques. Pourtant, les différences entre les communautés planctoniques et sessiles sont communément acceptées (Smith *et al*., 2018). Les communautés sessiles pourraient être avantagées par rapport aux communautés planctoniques parce que certains microorganismes qui la composent possèdent la capacité de dégrader les surfaces rocheuses afin d'en ressortir des éléments essentiels à leur survie (Park *et al*., 2009; Stevens et McKinley, 1995; Bennett *et al*., 2001). Ces dernières ont également la possibilité de former des biofilms, ce qui leur confère une protection contre les toxines ou les prédateurs (Uroz *et al.,* 2015; Griebler *et al*., 2002) et leur permet de former des communautés synergiques qui peuvent effectuer des processus que les espèces individuelles ne peuvent pas (Beveridge *et al*., 1997; Griebler *et al*., 2002). D'un autre côté, les microorganismes planctoniques ont l'avantage de pouvoir se déplacer en quête de nutriments et d'énergie. Cependant, il existerait un équilibre entre les processus d'attachement et de détachement entre les deux communautés,

surtout au niveau de l'aquifère (Goldsheider *et al*., 2006). Ainsi, les études menées sur le fonctionnement des écosystèmes de la subsurface terrestre basées uniquement sur les communautés planctoniques n'incluent pas le microbiome total de l'aquifère et pourraient être sujettes à de la mésinterprétation (Lazar *et al*., 2019).

Objectifs

Les objectifs principaux de ce projet de maîtrise sont les suivants:

(1) Comprendre l'effet de la profondeur et des paramètres abiotiques associés sur la diversité α, la composition et l'abondance absolue des communautés microbiennes (bactéries, archées et eucaryotes) endolithiques et sessiles dans la subsurface terrestre superficielle;

(2) Déterminer les interactions biotiques potentielles existantes entre les taxons des communautés endolithiques et sessiles de bactéries, d'archées et d'eucaryotes au sein de la subsurface terrestre superficielle;

(3) Déterminer la proportion des communautés microbiennes de la subsurface terrestre superficielle qui sont formées par les mouvements de flux verticaux;

(4) Déterminer les différences de composition entre les communautés planctoniques détectées dans l'eau de l'aquifère et les communautés microbiennes sessiles détectées dans les sédiments avoisinants.

Hypothèses et prédictions

De nombreux facteurs abiotiques qui façonnent la structure des communautés microbiennes (pH, texture du sol, quantité en carbone organique, disponibilité des nutriments, contenu en eau et niveau d'oxygène, etc.) varient avec la profondeur (Eilers *et al*., 2012). De plus, les horizons plus profonds de la subsurface terrestre superficielle sont généralement caractérisés par des conditions extrêmes de pH, de concentrations en nutriments, et en énergie (Bomberg et Ahonen, 2017; Govil *et al.*, 2019; Xu *et al*., 2021) et des environnements très hétérogènes sont formés pour les communautés microbiennes le long du profil de sol (Xu *et al*., 2021). La plupart des taxons archées possèdent une niche écologique large ou préfèrent vivre dans des environnements extrêmes (Xu *et al*., 2021). En revanche, la plupart des bactéries et des eucaryotes ne peuvent pas endurer de telles conditions de stress (Eilers *et al*., 2012; Xu *et al*., 2021). Par conséquent, nous formulons les hypothèses et prédictions suivantes :

(1) La profondeur agit comme un filtre de l'habitat en sélectionnant les taxons les mieux adaptés aux conditions abiotiques extrêmes retrouvées dans les horizons profonds de la subsurface terrestre superficielle. Nous prédisons : que le long du gradient de profondeur, la diversité taxonomique α diffère entre les groupes microbiens en raison de leurs différences physiologiques. Plus précisément, avec l'augmentation de la profondeur, la diversité taxonomique α et l'abondance absolue des communautés bactériennes et eucaryotes diminuent, alors que la diversité taxonomique et l'abondance absolue des archées augmentent; la diversité phylogénétique α diminue pour les trois domaines, car les microorganismes de la communauté sont limités par une exigence de traits de tolérance au stress connus sous le nom de filtration de l'habitat ou phylogénétique; la composition globale des communautés bactériennes, archées et eucaryotes changent avec la profondeur.

(2) Dans tous les milieux et notamment dans des milieux où les conditions peuvent être extrêmes comme dans la subsurface terrestre superficielle, il existe des interactions positives (p. ex., codépendance) et négatives (p. ex., compétition) entre les taxons des communautés de bactéries, d'archées et d'eucaryotes. Nous nous attendons à observer des corrélations interspécifiques négatives, principalement avec les taxons archées (meilleurs compétiteurs dans les milieux extrêmes) et des corrélations positives– qui suggèrent que les taxons utilisent des ressources différentes ou sont codépendants - entre les taxons des communautés eucaryotes et bactériennes.

(3) Par infiltration, les flux verticaux contribuent à l'assemblage des communautés microbiennes de la subsurface terrestre superficielle. Ces flux traversent les différents horizons de la subsurface et transportent avec eux les communautés microbiennes retrouvées à la surface et/ou au niveau des échantillons des étages supérieurs vers les étages plus profonds. Par conséquent, nous prédisons que les échantillons provenant des horizons du sol supérieurs (sources) contribuent à la formation des communautés microbiennes des étages sous-jacents (puits).

(4) Bien que l'eau de l'aquifère et les roches environnantes soient relativement similaires dues à leur proximité, les différences de substrat, de caractéristiques (p. ex., pH, niveau d'oxygène, connectivité) et de style de vie pourraient causer une différenciation entre les communautés planctoniques et sessiles. Nous nous attendons à ce que les communautés suspendues soient relativement différentes des communautés attachées environnantes malgré la présence de populations microbiennes communes dues aux échanges entre les deux communautés.

CHAPITRE I

ARTICLE DE RECHERCHE : CHANGE IN DIVERSITY, COMPOSITION, AND ABSOLUTE ABUNDANCE OF SESSILE AND ENDOLITHIC BACTERIAL, ARCHAEAL, AND EUKARYOTIC COMMUNITIES WITH SOIL DEPTH IN TWO SOIL PROFILES FROM THE SHALOW TERRESTRIAL SUBSURFACE IN THE LAURENTIANS (QC, CANADA)

Julia Meyer[1], Sheri Zakhary[1], Marie Larocque[2], and Cassandre S. Lazar[1]

Manuscrit publié dans le journal Microorganims à l'issue spéciale "Microbe-Driven Migration and Transformation of Elements through the Earth's Critical Zone" dans la section Environmental Microbiology sous le numéro 1469595

[1] Department of Biological Sciences, University of Quebec at Montreal, C.P. 8888, Succ. Centre-Ville, Montréal, Québec, H3C 3P8, Canada

[2] Department of Earth and Atmospheric Sciences and GEOTOP, University of Quebec at Montreal, C.P. 8888, Succ. Centre-Ville, Montréal, Québec, H3C 3P8, Canada

Abstract: Microbial communities play an important role in terrestrial subsurface ecosystems. Most studies of this habitat have focused on planktonic communities that are found in the groundwater of aquifer systems and only target specific microbial groups. Therefore, a systematic understanding of processes that govern the assembly of endolithic and sessile microbial communities is still missing. This study aims to understand the effect of depth and biotic factors on these communities, to better unravel their origins and to compare their composition with the microbial communities detected in groundwater. To do so, we collected samples from two profiles (~ 0–50 m) in aquifer sites in the Laurentians (Quebec, Canada), performed DNA extractions and Illumina sequencing. Results suggest that changes in geological material characteristics with depth represent a strong ecological and phylogenetical filter for most archaeal and bacterial communities. Additionally, vertical movement of water from the surface play a major role in shallow subsurface microbial assembly processes. Furthermore, biotic interactions between bacteria and eukaryotes were mostly positive which may indicate cooperative or mutualistic potential associations, such as cross-feeding and/or syntrophic relationships in the terrestrial subsurface. Our results also point toward the importance of sampling both the geological formation and groundwater when it comes to studying its overall microbiology.

Keywords: terrestrial subsurface; genomics; bacteria; archaea; eucaryote; planctonic; microbial interactions

1.1 Introduction

In the shallow terrestrial subsurface (< 50 m below surface), microorganisms play an important role in soil nutrition cycling, soil respiration, soil formation, ecosystem biochemical processes, contaminant degradation, as well as groundwater maintenance (Chu *et al*., 2016; Hartmann *et al*., 2009). In recent years, many subsurface inventories have surveyed the distribution patterns or processes that govern microbial community assembly (Hartmann *et al*., 2009; Chu *et al*., 2016). However, due to the difficulty in accessing rocks and sediments below ground, most of these inventories have focused solely on the planktonic communities that live suspended in the aquifer groundwater of this habitat (Lazar *et al*., 2019). Yet, endolithic and sessile communities that live in rocks and are attached to sediments constitute the majority of cells in the subsurface (Flynn *et al*., 2013). Consequently, most of these subsurface inventories bear the risk of misconceptions (Lazar *et al*., 2019).

Depth plays a major role in shaping microbial communities in subterranean ecosystems (Kim *et al*., 2016; Eilers *et al*., 2012; Fierer *et al*., 2003; LaMontagne *et al*., 2003; Agnelli *et al*., 2004; Goberna *et al*., 2005; Will *et al*., 2010; Hansel *et al*., 2008). Depth is also related to the mechanisms of microbial community formation because vertical movements of water from the surface bring microorganisms into the terrestrial subsurface (Amy *et al*., 1992). However, despite the existing literature on the effect of depth on microbial communities, our overall understanding remains imcomplete. Primary, studies on this subject are still limited and most of them describe the effect of depth on large spatial scales (by comparing the surface with the subsurface), or by focusing exclusively on near-surface soil horizons (Kim *et al*., 2016; Xu *et al*., 2021; Eilers *et al*., 2012). Studies also often target specific microbial domains (e.g., archaea, bacteria, or eukaryote), making it difficult to gain a systematic understanding of the total microbial community diversity and the interactions between organisms composing them (Xu *et al*., 2021; Hartmann *et al*., 2009). Moreover, although different microbial community structural characteristics have been examined in previous study, few studies have simultaneously studied the effect of depth on absolute abundance, phylogenetic diversity, taxonomic diversity, and composition (Xu *et al*., 2021). Yet, these characteristics provide complementary information. Analysing all these characteristics allos us o get more insight intoo the processes at the origin of the assembly of communities by making it possible to quantify the importance of different ecological and evolutionary processes such as dispersion, competition, and filtration of the environment. In addition, learning the origins of microbial communities may not only significantly improve our current understanding of how shallow subsurface microbial communities are

formed, but could also inform on the consequences of some human activities that affect the soil (e.g., agricultural inputs) (Shenhav *et al*., 2019).

Therefore, the objectives of this study are to (1) determine which factors govern microbial community diversity in samples from surface to the bedrock (i.e., shallow terrestrial subsurface), and (2) estimate the proportion of subsurface communities that are formed through vertical fluid fluxes originating from the surface, in two Laurentian sites harbouring aquifers. We analyzed abiotic factors such as the physico-chemical, geochemical, and mineralogical characteristics of the geological material with depth, but also biotic factors such as potential interactions between microorganisms. Total microbial community was studied though sequencing of archaeal and bacterial 16S rRNA genes, as well as 18S rRNA eukaryotic genes.

1.2 Materials and Methods

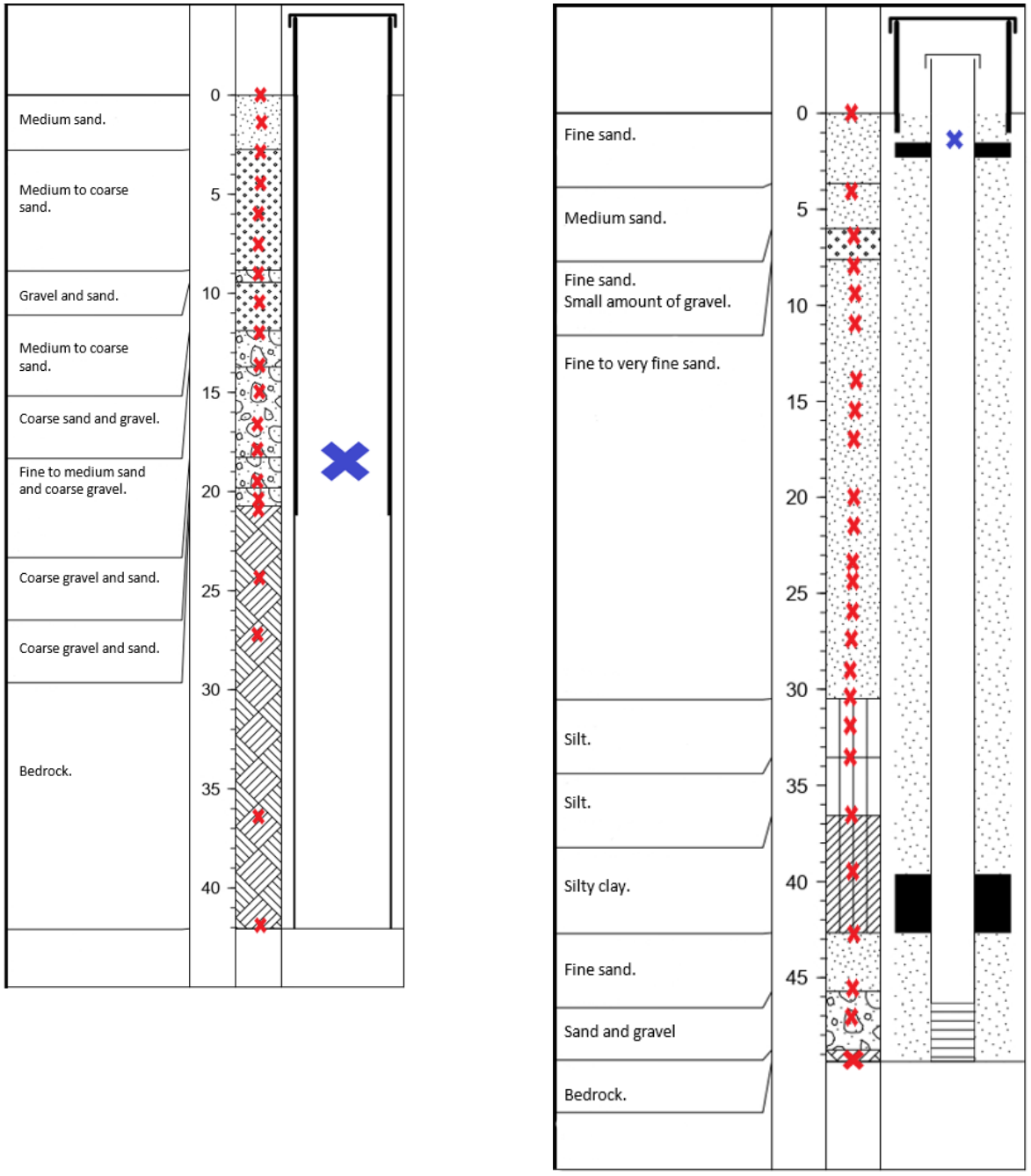
1.2.1 Study Sites

We selected two sites in the Laurentians region (Quebec, Canada) for this study. The first site is in the municipality of Ascension (46°36'42.66 "North, 74°47'3.911" East) and the second site is in the municipality of Notre-Dame-du-Laus (46°0 '59.148 "North, 75°34'37.667" East) (**Figure 1.1**). Both sites differ in terms of the size and relative abundance of mineral particles. At site 1 (Ascension), the geological material was mostly composed of sand and had a smaller amount of clay compared to site 2 (Notre-Dame-du-Laus) (**Table S1 and S2**). Both sites are located in a forested area. At site 1, the vegetation type is dominated by resinous trees mainly composed of black spruce. Site 2 is also located in a resinous tree dominated forest, but pine is the dominant species. Well-drained loam is the soil type in both sites. For site 1, climatic data from the period of 1981–2010 were taken from the station located 25 km southward (Government of Canada, 2021a). Mean annual precipitation is 1028.9 mm and annual average temperature is 4 °C. At site 2, climate data from the closest weather station for which a 1981–2010 climatic normal was available is located 33 km west (Government of Canada, 2021b). Mean annual precipitation is 939.9 mm and annual average temperature is 4.7 °C. We acquired samples during a regional scale aquifer characterization project (Department of Earth and Atmospheric Sciences, UQAM). During this project, two new wells were drilled using a double rotation drill in a destructive mode during the summer of 2019. Bedrock was reached at 20.73 m at site 1 (drilling stopped at 42.06 m), and 48.77 m at site 2 (drilling stopped at 49.38 m).

Figure 1.1. Location of the two study sites. Both sites are in the northwest of Montreal in the Laurentians region (Quebec, Canada).

1.2.2 Sampling

We collected a total of 46 samples, i.e., 20 samples at site 1 (**Figure 1.2.a**) and 26 samples at site 2 (**Figure 1.2.b**). Sampling was carried out vertically from the surface down to the bedrock, from the material extracted during the drilling. Estimation of depth for each sample was based on the operator's information during drilling. It is estimated that these depths are precise at ± 0.5 to 1 m. We collected the soil, sediments, and crushed bedrock material rising from the subsurface in sterile 50 mL Falcon tubes on the same day as the drilling. To minimize potential contamination, we subsampled them by taking the materials in the center of the container used for sampling. We collected water used for drilling (from the surface water of a local stream) in a 50 mL Falcon tube, to identify potential contamination from microbial communities potentially introduced into the samples during drilling. Groundwater samples were collected in the newly drilled wells at site 1 (18 m depth) and site 2 (1.75 m depth), using a submersible pump (12V/24V Mini-Monsoon, Waterra, Canada). It is important to underline that the sampled groundwater from site 1 represents a mix of water coming from the open borehole, i.e., between 20.73 and 42.06 m.



(a) (b)

Figure 1.2. Diagrams of the 15.2 cm diameter boreholes, showing the depth at which samples of rock-associated (red crosses, sediments samples during drilling) and planktonic (blue crosses, samples in the

well water after drilling) subsurface communities were taken (a) at site 1 (Ascension) and (b) site 2 (Notre-Dame-du-Laus).

1.2.3 Chemical Characteristics of Geological Material and Water Samples

We determined the geological material texture in the field by touch (Ritchey *et al*., 2015). Prior to further analysis, subsamples were dried at 60 °C for 48 h and ground into a homogeneous powder using a mortar and pestle. We measured soil pH in triplicate in a 1:2 (w/v) soil/water ratio using an Accumet XL600 pH meter (Fisher Scientific, Waltham, MA, USA) after 1 h of incubation (Eckert and Sims, 1995). Then, we conducted analyses for total nitrogen content, total carbon content, total organic carbon content and total inorganic carbon of geological material at the light stable isotope geochemistry laboratory of the GEOTOP center at UQAM using the NC 2500 elemental analyzer (Carlo Erba, Milan, Italy). We determined the total organic carbon content after pre-treatment of the samples with 5 % HCl to remove inorganic carbon. The mineral composition of the samples was determined at the research center NanoQAM (University of Quebec in Montreal, Quebec, Canada) with the X-ray diffraction (XRD) method using the X-ray diffractometer D8 Advance (Bruker, Billerica, MA, USA). The instrument was adapted with a copper tube (Cu K-α = 154178 Å) and samples were scanned from 5° to 70° 2-θ with a step of 0.02° 2-θ. The resulting diffractograms were converted to percentage by weight of minerals using the powdR program Retrieved 10 December 2020, from https://CRAN.R-project.org/package=powdR, accessed on 23 December 2021) (Butler et Hillier, 2021) and quartz as internal standard. For the water samples, we only determined the pH. To do so, we used an Oakton pH meter (Waterproof pH/DO 450 Meter, Oakton, CA, USA).

1.2.4 DNA Extraction

DNA was extracted from 10 g of samples following the instructions of the commercial DNeasy PowerMax Soil Kit (Qiagen, Hilden, Germany) with some modifications (Lazar *et al*., 2017). Obtaining representative DNA extracts proved very challenging due to the low cellular biomass, the adsorption of cells to geological material, and the frequent co-extraction of enzymatic inhibitors (Alain *et al*., 2011; Lazar *et al*., 2019). Therefore, following Direito et al. (2012), we optimized the kit by using a 1 M of NaH2PO4/Na2HPO4 buffer in the initial steps of the kit manufacturer's protocol (Lazar *et al*., 2019). We tested one control sample in order to detect extraction-process contamination by applying sterilized water to the DNA extraction kit. We then filtered the water samples (drilling fluid and groundwater from sites 1 and 2) through a sterile 0.2 µm polyether sulfone filter (Sartorius, Midisart, Germany). DNA was extracted using the DNeasyPowerWater Kit (Qiagen, Germany). We stored the obtained DNA extracts at -20 °C.

1.2.5 PCR Amplification and Illumina Sequencing

Sequencing was performed at the Center of Excellence in Research on Orphan Disease – Fondation Courtois (CERMO-CF) at UQAM. We amplified the DNA extracted from each sample with universal primers that target the hypervariable V3-V4 region of the bacterial 16S rRNA genes, the V3-V4-V5 region of the archaeal 16S rRNA genes, and the V7 region of the eukaryotic 18S rRNA genes. For this, we used the primer pair B341F (5′ -CCT ACG GGA GGC AGC AG -3′) (Muyzer *et al*., 1993) - B785R (5′ - GAC TAC CGG GGT ATC TAA TCC -3′) (Klindworth *et al*., 2013), A340F (5'-CCC TAC GGG CYC CAS CAG-3 ') (Baker *et al*., 2003) - A915R (5'-GTG CTC CCC CGC CAA TTC CT-3 ') (DeLong, 1992) and E960F (5′ -GGCTTAATTTGACTCAACRCG-3′) (Gast *et al*., 2004) - NSR1438R (5′ -GGGCATCACAGACCTGTTAT-3′) (Van de Peer *et al*., 2000). PCR amplification was carried out using the UCP HiFidelity PCR kit (Qiagen, Hilden, Germany). We performed the amplification under the following PCR conditions: initial denaturation at 98 °C for 30 seconds, denaturation at 98 °C for 30 seconds, primer annealing for 30 seconds (at 57 °C for bacteria, 58 °C for archaea and 56.6 °C for eukaryote), extension at 72 °C for 1 minute, and a final extension at 72 °C for 10 minutes. The denaturation, annealing and extension steps were repeated 33 times for bacteria and 40 times for both archaea and eukaryote. Sequencing was performed using an Illumina MiSeq (2 × 300) and the MiSeq Reagent Kit v3 (600 cycles, Illumina), according to manufacturer's instructions. The obtained sequences were deposited in the National Center for Biotechnology In-formation (NCBI) under the BioProject ID: PRJNA758373.

1.2.6 Digital PCR

Digital PCR amplification used for the estimation of absolute microbial abundance was challenging because of the overall low microbial biomass, especially for archaeal and eukaryotic genes. Therefore, we only show digital PCR results for the 16S rRNA genes of the bacteria (the number for archaea and eukaryotes were below the quality threshold). We performed these analyses using the QuantStudio 3D Digital PCR (ThermoFisher, Waltham, MA, USA) instrument and the QuantStudio 3D PCR Master Mix v2 (ThermoFisher, USA). We used the primers B341F (5′ -CCT ACG GGA GGC AGC AG -3′) (Muyzer *et al*., 1993) - B785R (5′ -GAC TAC CGG GGT ATC TAA TCC-3′) (Klindworth *et al*., 2013) and the reaction procedure was as follows: 96 °C for 10 minutes; 39 PCR cycles of 56 °C for 3 minutes, annealing at 56 °C for 3 minutes, then 98 °C for 30 seconds, and a final extension at 56 °C for 2 minutes. The dPCR results were expressed as gene copies per gram of geological material (copies g^{-1}).

1.2.7 Sequence Processing

Sequence data was processed using the DADA2 Pipeline (v.1.18.1) (Retrieved 29 June 2021, from https://benjjneb.github.io/dada2/tutorial_1_8.html, accessed on 23 December 2021) described by Callahan et al. (2016). Based on the quality profiles generated for both forward and reverse reads, we choose to truncate the bacterial forward reads at position 290, and the reverse reads at position 250. For eukaryotic sequences, we trimmed both forward and reverse reads at position 250. Because of the low quality of the reverse reads, we only kept the forward reads for archaeal sequences and trimmed them at position 210. Then, we determined the unique amplicon sequence variants (ASVs) with the inference algorithm as implemented in DADA2 (Callahan *et al*., 2016). Afterwards, we merged the forward and reverse reads using mergePairs and chimeras were removed with removeBimeraDenovo. We assigned taxonomy using IDTAXA taxonomic classification approach (Murali *et al*., 2018) along with SILVA SSU (v.138) database (Retrieved May 01, 2021, from Latin silva, forest, http://www.arb-silva.de, accessed on 23 December 2021) (Quast *et al*., 2012). The non-classified archaeal and eukaryotic sequences underwent a second classification using a datasetconstructed based on Liu et al. (2018) and Zhou et al. (2018) for the 16S rRNA archaeal genes, and on the PR^2 database (Retrieved 12 May 2021, from https://pr2-database.org/) for the 18S rRNA eukaryotic genes. For further analysis of the microbiome data, we removed contaminant sequences present in controls (drilling fluids, blank DNA kit extraction, and negative sample from the PCR amplifications prior to sequencing) from the output sequence table. Processed sequences were aligned with DECIPHER::alignSeqs and a phylogenetic tree was constructed with FastTree (v.2.1.10) (Price *et al*., 2009).

1.2.8 Statistical Analysis

All statistical analyses were performed with R (v.4.0.3) using phyloseq (v.1.34.0) (McMurdie and Holmes, 2013), and vegan (v.2.5-7) packages (Oksanen *et al*., 2013). In order to adjust for differences in library sizes across samples, we used the median sequencing depth normalization method (Pereira *et al*., 2018) and only used samples with more than 1000 sequences. Taxonomic α-diversity was estimated with Shannon and Simpson indices while phylogenetic α-diversity was estimated with Faith's PD index (Faith, 1992) and the mean nearest taxon distance (MNTD) index (Webb *et al*., 2008). The combined use of both metrics allows for a better understanding of assemblage structure, while the use of only one metric could lead to incomplete interpretations and biased conclusions (Mazel *et al*., 2016). We also calculated the standardized phylogenetic diversity measure of both indices (SES_{PD} and SES_{MNTD}) using the null model "taxa.labels" (999 randomization) in Picante R package (v.1.8.2) (Kembel *et al*., 2010). The standardized

phylogenetic diversity measure expresses how different the observed phylogenetic diversity value is (in units of standard deviations (sd) from the average (mean) phylogenetic diversity in the randomly generated communities. Positive values indicate phylogenetic evenness (co-occurring sequences more phylogenetically distantly related than expected by chance), while negative values indicate phylogenetic clustering (co-occurring sequences more closely related than expected by chance) (Kembel et al., 2011). We determined the relationships between abiotic factors and depth, and abiotic factors and microbial α-diversity through Spearman correlation analyses. For β-diversity analysis, sample-sample distances were determined with the Bray-Curtis distance and visualized with non-metric multidimensional scaling (nMDS). A Mantel test was used to determine the effects of abiotic factors on β-diversity of microbial communities. In addition, we used an analysis of similarities (ANOSIM) to determined if the communities were significantly different among samples with different sedimentary textures. To estimate the proportion of upper soil horizons (sources) contributing to the formation of the microbial communities in deeper soil layers (sinks), we used Fast Expectation-mAximization microbial Source Tracking (FEAST) (Shenhav *et al*., 2019). Finally, we investigated potential biotic interactions between bacterial, archaeal, and eukaryotic endolithic and sessile communities from the terrestrial subsurface at the ASV-level. Because the focus for the biotic interactions were the endolithic and sessile communities from the terrestrial subsurface, we filtered the data to remove the groundwater samples and the samples taken at the surface. We only kept the samples containing sequences from the three domains and the 15 most abundant ASVs from each domain. Then, Spearman correlations were calculated using the relative abundance of each ASV and visualised in a Heatmap. For all statistical analysis, we considered p-value <0.05 to be statistically significant.

All the R codes used for this study are available in the supplementary material (p.64-124).

1.3 Results

1.3.1 Variation of Chemical Characteristics with Depth

The chemical characteristics of the geological material (pH, total nitrogen, total carbon, total organic carbon, total inorganic carbon) varied with depth and between sites (**Figure 1.3**). Subsurface pH varied from neutral to basic and significantly increased along soil depth at site 1 (Spearman's r = 0.83, p-value = 1.12×10^{-5}, S = 194.17, **Table S3**) with the most pronounced changes occurring between depths 20-30 m (site 1) and 0-10 m (site 2) (**Figure 1.3.a**). The total nitrogen percentage values were below our detection limits for all samples except for the near-surface horizons samples (**Table S1** and **Table S2**). The total carbon, total organic carbon and total inorganic carbon content varied between sites. At site 1, total carbon, total organic carbon, and total inorganic carbon content were higher in the surface soil sample than in the subsurface samples (**Table S1**). In the subsurface samples, the total carbon content was variable throughout the profile (ranging from 0.01 % to 0.06 % across all subsurface samples, **Figure 1.3.b**). Total organic carbon content was higher in the top 13 m samples in comparison to deeper samples where the % of total organic carbonwere all below our detection limits. At site 2, the total carbon content in the subsurface significantly increased with depth (Spearman's r = 0.91, p-value = 1.92×10^{-10}, S = 225.51, **Table S3**). Total organic carbon content was variable throughout the soil profile (ranging from 0.01 % to 2.27 % across all samples), decreasing by over six orders of magnitude from the four shallowest samples to the following twenty samples and finally increased by nearly 45 orders of magnitude in the second last deepest samples (**Figure 1.3.c**). At both sites, the % of total inorganic carbon in the subsurface significantly increased with increasing depth (Site 1 : Spearman's r = 0.58, p-value = 9.53×10^{-3}, S = 480.97; Site 2 : Spearman's r = 0.92, p-value = 6.52×10^{-11}, S = 204.62, **Table S3**) (**Figure 1.3.d**). Sedimentary texture changed with depth at both study sites. At site 1, as depth increased, soil texture varied continuously from finer materials to coarser materials while at site 2, soil texture varied generally from coarser materials (e.g., medium sand) in the surface to finer materials (e.g., clay) (**Table S1 and S2**). At site 1, pH, % total carbon, and % total inorganic carbon increased in the bedrock samples (**Figure 1.3**, hollow circles). At both sites, groundwater pH was close to neutrality (**Table S1 and S2**).

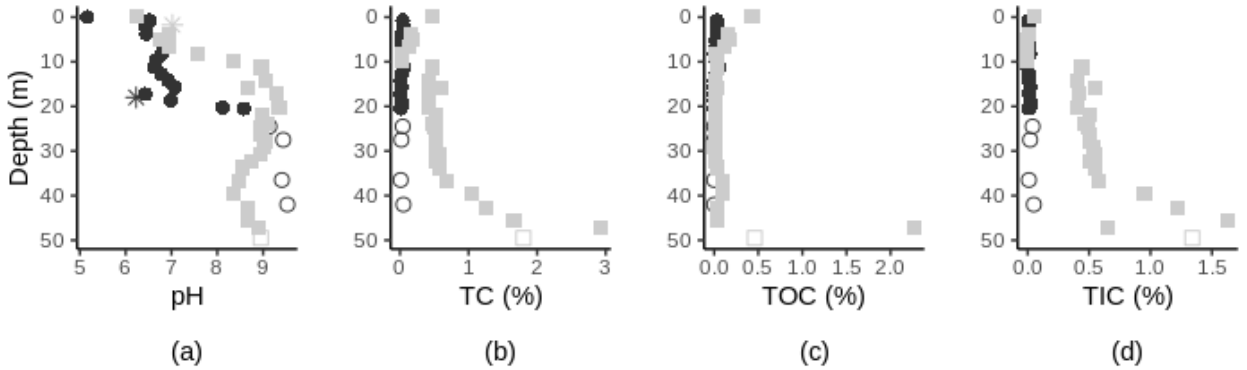


Figure 1.3. Chemical characteristics of geological material along depth at site 1 (black circles) and site 2 (grey squares) with pH (a), total carbon (TC) (b), total organic carbon (TOC) (c), and total inorganic carbon (TIC) (d) along depth. Surface sample at site 1 was removed from the TC, TOC and TIC figures for a better visual representation. Bedrock samples are represented by hollow symbols. pH of the groundwater samples (located at 18 m and 1.75 m of depth, at site 1 and 2 respectively) are represented by stars in (a). For full details on the changes of soil characteristics, see Table S1 and S2.

The subsurface samples taken in the scope of this study were characterized by a mosaic of different minerals (**Figure 1.4**). They were predominantly composed of quartz (SiO_2), plagioclase ($NaAlSi_3O_8$–$CaAl_2Si_2O_8$) and K-feldspar ($KAlSi_3O_8$). The surface sample at site 1 was mostly composed of organic matter (~ 85 %) (**Figure 1.4.a**). The bottom three samples at site 2 were characterized by the presence of detrital carbonates (calcite and dolomite) which are abundant in rocks from the St. Lawrence Platform (**Figure 1.4.b**). At both sites, the relative abundance of quartz significantly decreased with depth (Site 1 : Spearman's r = -0.48, p-value = 0.03, S = 1974.20) (Site 2 : Spearman's r = -0.47, p-value = 0.02, S = 4296.50). There were also strong significant relationships between depth and the content of plagioclase (Site 1 : Spearman's r = 0.72, p-value = 5.62×10^{-4}, S = 378), calcite ($CaCO_3$) (Site 2 : Spearman's r = 0.63, p-value =6.41×10^{-4}, S = 1096.80) and dolomite ($CaMg(CO_3)_2$) (Site 2 : Spearman's r = 0.60, p-value = 1.06×10^{-3}, S = 1155.90).

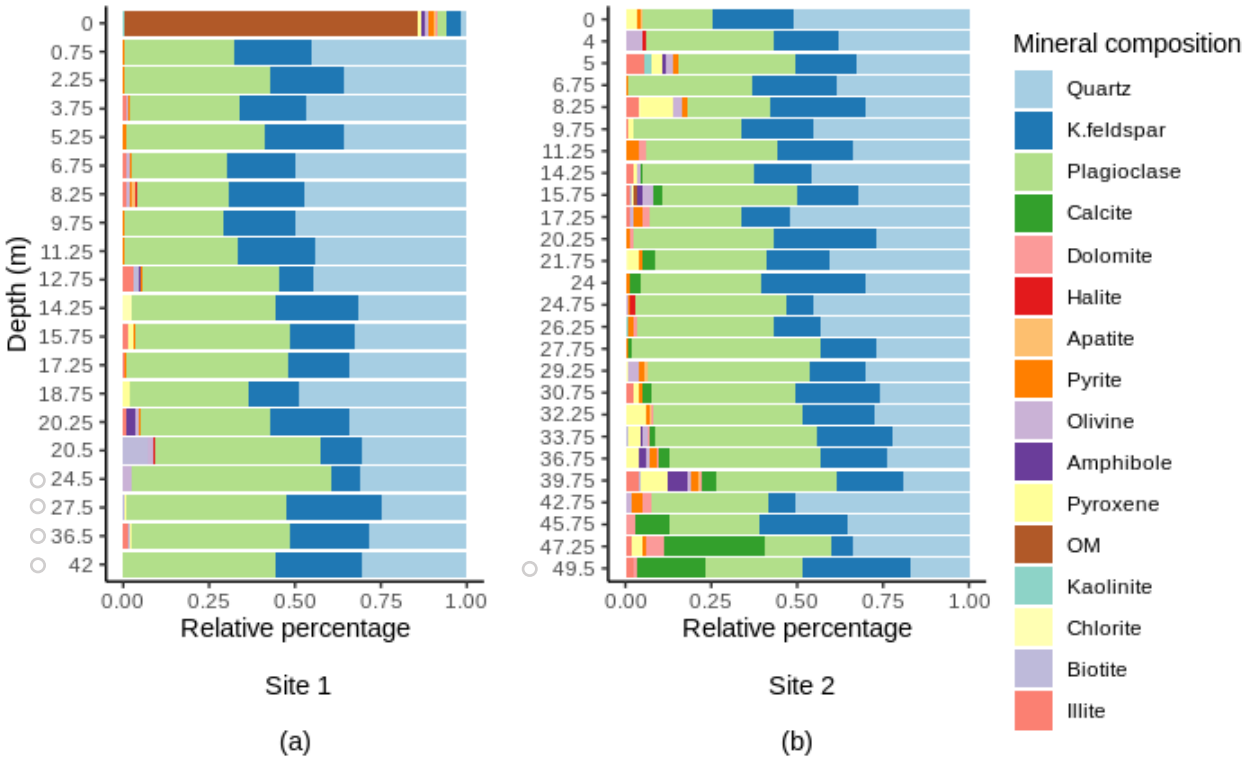


Figure 1.4. Mineral composition along depth at site 1 (a) and at site 2 (b). OM, organic matter. Bedrock samples are indicated by hollow symbols. For full details on the mineralogical composition, see Table S4 and S5.

1.3.2 Taxonomic α-Diversity : Variation with Depth and Correlation with Chemical Characteristics of Geological Material

Bacterial taxonomic α-diversity decreased significantly with depth whether diversity was expressed by Shannon or Simpson indexes (Spearman's r < -0.50, p-value <0.05, **Table S6** and **Figure S1**) with the most pronounced change occurring when the bedrock was reached, between 20 and 24 m of depth at site 1 (**Figure S1a** and **S1b**). We found significant relationships between bacterial taxonomic α-diversity metrics and pH, % total organic carbon, and % total inorganic carbon at site 1, and with % total carbon and % total inorganic carbon at site 2 (**Table S6**). Archaeal taxonomic α-diversity peaked at intermediate depth (~ 10 - 15 m) and was lower at the surface and in the deepest horizons (**Figure S1e** and **S1f**). The archaeal taxonomic α-diversity significantly decreased along the soil profile (0 – 46 m) at site 2 (Shannon : Spearman's r = -0.73, p-value = 8.88×10^{-5}, S = 3972; Simpson : Spearman's r = -0.62, p-value = 1.58×10^{-3}, S = 3724, Table S7). We found negative significant relationships between the archaeal taxonomic α-diversity metrics and the % of total carbon, % of total inorganic carbon at site 2 (Spearman's r < -0.55, p-value <0.01, **Table S7**). Eukaryotic α-diversity did not show clear trends with depth or any other geological

material characteristics (p-value > 0.05, in all cases, **Table S8**). However, some of our samples, especially the deepest ones contained a very low number of sequences for archaea and eukaryotes (< 1000 reads) and these samples were not used for comparisons of taxonomic α-diversity for any of the subsequent analyses.

1.3.3 Phylogenetic α-Diversity and Correlation with Geological Material Characteristics

At site 2, bacterial phylogenetic α-diversity significantly increased with depth when measured with MNTD index (Spearman's r = 0.80, p-value = 3.29×10^{-6}, S = 514, **Table S6**) or decreased with depth when measured with Faith's PD (Spearman's r = -0.75, p-value = 3.11×10^{-5}, S = 4538, **Table S6**, **Figure S1c** and **S1d**). As observed with the taxonomic α-diversity indices, the most pronounced changes in phylogenetic diversity occured when the beadrock was reached. Using Spearman correlation, we also found significant relationships between the bacterial phylogenetic α-diversity metrics and the % of total carbon and the % of total inorganic carbon at site 2 (**Table S6**). Archaeal phylogenetic α-diversity changed significantly with depth when measured with Faith's PD index (**Table S7**) and showed opposite patterns in the studied sites (**Figure S1g** and **S1h**). In the depth interval where archaeal sequences were detected at site 1 (0.75 -20 m), we found an increase of Faith's PD with depth (Spearman's r = 0.73, p-value = 6.32×10^{-3}, S = 98). By contrast, in the 0 – 46 m depth interval where archaeal sequences were detected at site 2, the archaeal phylogenetic α-diversity decreased with increasing depth (Faith's PD : Spearman's r < -0.73, p-value = 9.42×10^{-5}, S= 3968). Faith's PD was also correlated with the % of total carbon (both sites), the % of total organic carbon (both sites) and the % of total inorganic carbon (site 2) (p-value <0.05, in all cases, **Table S7**). Using only samples containing more than 1000 sequences (we did not use 22 samples, including six samples from the site 1 and 16 of the deepest samples at site 2), we did not find a clear relationship between the phylogenetic eukaryotic α-diversity metrics and soil depth at either sites (**Table S8**; **Figure S1k** and **S1l**). In addition, the majority of the standardized effect sizes of MNTD (SES_{MNTD}) and Faith's PD (SES_{PD}) values obtained using the null model were below zero ($SES_{MNTD} < 0$ and $SES_{PD} < 0$), meaning that the sequences in the terrestrial subsurface samples's were more closely related to each other than expected by chance (Kembel *et al*., 2011).

1.3.4 β-Diversity and Correlation with Geological Material Characteristics

The β-diversity levels (i.e., the Bray-Curtis indices) changed significantly with depth, for all microbial domains and at both study sites (p-value < 0.05, **Table S9**). At both sites, depth, texture and pH were the major factors shaping bacterial communities (**Figure 1.5**, **Table S9**). For the archaeal community, we found

moderate correlations between the ß-diversity levels and, pH and total carbon content (**Table S9**). Finally, the eukaryotic community composition was more strongly correlated with the % of total carbon (site 1) and the % of total organic carbon (site 2) rather than with depth (**Table S9**). We did not find any strong correlations between the microbial composition and the mineralogy composition. At both sites and for all microbial domains, the community composing the groundwater samples was distinct from the the community in the geological material (**Figure 1.5**).

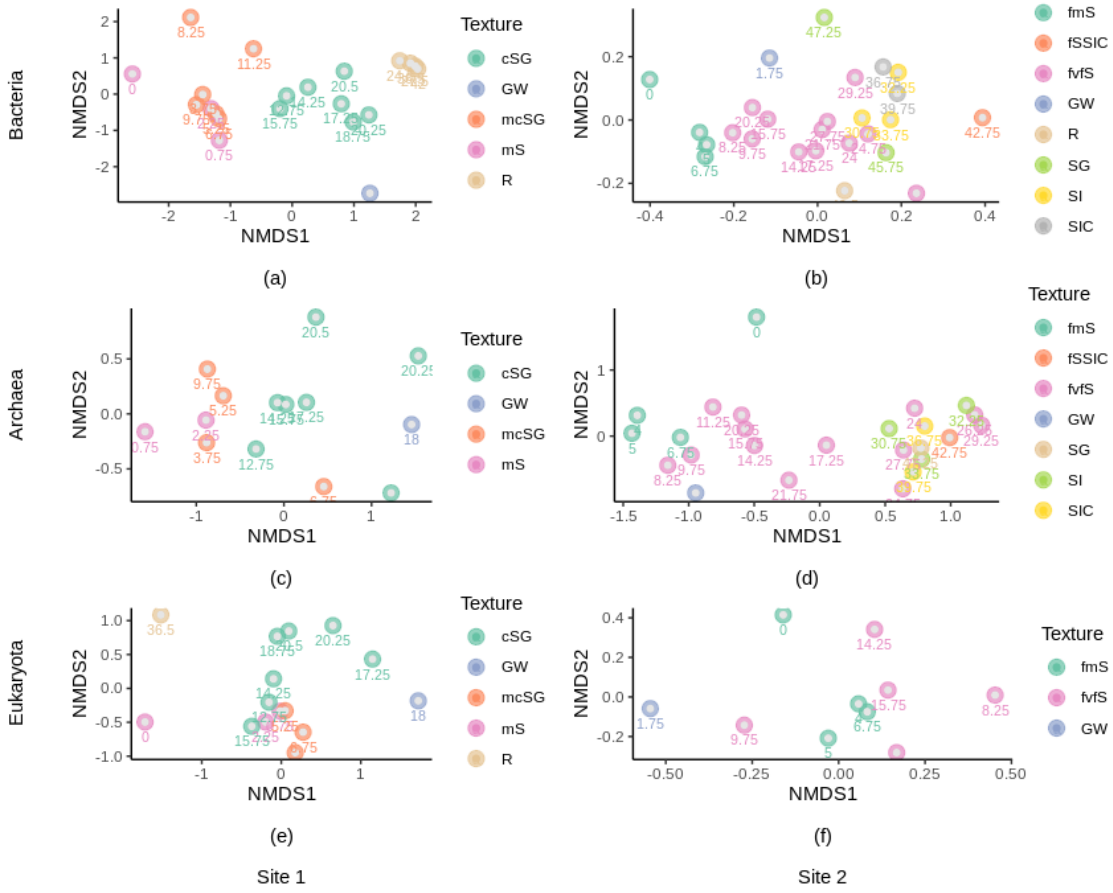


Figure 1.5. Microbial community compositional structure in the subsurface samples as indicated by non-metric multidimensional scaling plots (nMDS) on Bray-Curtis dissimilarity matrices. Geological material textures are color coded and depth is indicated for each sample (subscript). With: (a) bacterial composition at site 1, stress =0.113; (b), bacterial composition at site 2, stress= 0.159; (c) archaeal composition at site 1, stress=0.0094; (d) archaeal composition at site 2, stress=0.137; (e) eukaryotic composition at site 1, stress=0.0107; (f) eukaryotic composition at site 2, stress =0.133. mS, medium sand; mcSG, medium to coarse sand and gravel; cSG, coarse sand and gravel; R, Bedrock; fmS, fine to medium

sand; fSSIC, fine sand with silt and clay; fvfS, fine to very fine sand; SG, sand and gravel; SI, silt; SIC, Silt and clay; GW, groundwater.

1.3.5 Absolute Bacterial Abundance

Absolute bacterial abundance determined with dPCR was clearly higher in the soil surface in comparison to the subsurface (**Table 1.I**). Using only samples with good quality treshold, we only found significant correlation at site 2. The absolute bacterial abundance declined significantly with depth at site 2 (Spearman's r = -0.68, p-value = 1.20×10^{-3}, S = 1910, **Table S10**). We also found negative correlations between bacterial absolute abundance, the % of total carbon and the % of total inorganic carbon (site 2, **Table S10**).

Tableau 1.I. Absolute abundances expressed in number of bacterial 16S rRNA gene copies per gram of geological material (copies g^{-1}) detected in each sample. Only values with good quality thresholds are presented.

Site	Samples	Depth (m)	Abundance (Copies g^{-1})
Site 1	RR-01	0	682273937.5
	RR-02	0.75	109185
	RR-03	2.25	3618757.75
	RR-04	3.75	252068
	RR-05	5.25	344150.25
	RR-06	6.75	440495.5
	RR-11	14.25	203594.5
	RR-12	15.75	569690.5
	RR-14	18.75	475723.25
	RR-16	20.5	18367280.5
Site 2	NDDL-01	0	539684562.5
	NDDL-02	4	222415.5
	NDDL-03	5	190000.75
	NDDL-04	6.75	273854.25
	NDDL-05	8.25	83918.75
	NDDL-06	9.75	520608
	NDDL-07	11.25	365820.5
	NDDL-08	14.25	367227
	NDDL-09	15.75	432781.5
	NDDL-10	17.25	145616.25
	NDDL-11	20.25	89283.75
	NDDL-12	21.75	34321.5
	NDDL-13	24	498270.75
	NDDL-14	24.75	35184.25
	NDDL-17	29.25	60160.5
	NDDL-20	33.75	33647.25
	NDDL-21	36.75	18212
	NDDL-22	39.75	97846
	NDDL-25	47.25	20474

Samples have been named according to the site's location and the order of sampling. RR: Riviere- Rouge (Site 1), NDDL : Notre-Dame-du-Laus (Site 2).

1.3.6 Variation in the Relative Abundance of Dominant Bacteria, Arhaea, and Eukaryote Microorganisms at the Phylum and Genus Levels, with Depth

At site 1, the most dominant bacterial phyla were Proteobacteria (51.62 % average relative abundance across all samples), Acidobacteriota (23.94 %), and Actinobacteriota (9.37 %). At site 2, the dominant bacterial phyla were represented by Proteobacteria (32.93 %), Chloroflexi (15.60 %), and Acidobacteriota (9.64 %) (**Figure 1.6.a** and **1.6.b**). The dominant bacterial genera belonged to Ralstonia (15.90 %), and Acinetobacter (2.33 %) at site 1, and Acinetobacter (8.30 %) at site 2 (**Figure S4a** and **S4b**). Their relative abundances were highly variable across the collected samples with the most changes occurring when the bedrock was reached, around 20- 25 m of depth at site 1 (**Figure S4a** and **S4b**). Relative abundance of the phylum Proteobacteria increased with depth at site 1 (**Table S11**) while relative abundances of Acidobacteriota and Myxococcota declined exponentially with depth at both sites (**Table S11**) (**Figure S2 and S3**). Some phyla like Actinobacteria were relatively most abundant at the surface and in the deepest horizons while the opposite pattern was observed for Nitrospirota (site 1, **Figure S2**). Likewise, most of the genera at site 1: *Bryobacter*, Candidatus Koribacter, *Acidothermus* decreased exponentially in relative abundance and were not found in the deepest samples (**Table S11**, **Figure S4**). The opposite pattern was observed for *Ralstonia*, the relative abundance of which increased with depth (**Table S11**) and peaked in relative abundance between 20 – 25 m of depth where this genus comprised up to 50 % of the community (**Figure S4**). The relative abundances of the other most abundant phyla and genera did not exhibit any clear shifts with depth (**Table S11**, **Figure S5**). Sessile/endolithic and planktonic communities shared similar taxa but differed in their relative abundances (**Figure 1.6**). Sessile/endolithic bacterial communities surrounding the groundwater contained more Acidobacteria than the planktonic communities (**Figure 1.6.a and 1.6.b**). At site 2, there were more Chloroflexi (23.97 % vs 0.06 % and 1.56 % in the surrounding solid materials), sva0485 (11.15 % vs 0.00 % and 0.56 %), Firmicutes (11.81% vs 0.12 % and 4.87%), Desulfobacterota (17.98 % vs 1.15 % and 1.30 %), and less Myxoccocota (0.08 % vs 4.281 % and 2.71 %) in the groundwater (**Figure 1.6.a and 1.6.b**). In addition, some abundant bacterial phyla were solely detected in the geological material and not in the groundwater (e.g., *Ralstonia*, Acinetobacter, Bryobacter, Rhodoplanes and Candidatus Koribacter, site 1, **Figure 1.6.a**).

At both sites, the archaeal communities were dominated by the Crenarchaeota phylum which comprised up to 80 % of the total community (**Figure 1.6.c** and **1.6.d**). At site 1, the archaeal community was also represented by the Woesearchaeota (13.10 %) and Thermoplasmatota (1.62 %). Nearly all the Crenarchaeota sequences (%) were classified as members of group 1.1c (54.43 %, **Figure S6c**), which was

dominant in the surface soil sample, and decreased with depth (**Table S11**, **Figure S7**). At site 1, the Woesearchaeota subgroup 24 increased in relative abundance with depth (**Table S11**, **Figure S6c**, **Figure S7**). At site 2, unclassified Bathyarchaeia, Crenarchaeota and Group 1.1c were dominant (59.77 %, 23.39 % and 3.32 % respectively). The relative abundance of unclassified Crenarchaeota and unclassified Group 1.1c decreased with depth, whereas unclassified Bathyarchaeia increased (**Table S11**, **Figure S8**). Methanogens such as Methanosarcina, Methanoregula or the uncultured Rice Cluster II, and the methanotroph cand. Methanoperendens were detected until 20 m depth and were also in the groundwater (**Figure S8**). The Marine Group II was detected in a vast majority in the groundwater.

At site 1, the most dominant eukaryotic phyla belonged to Basidiomycota (26.42 %), Ascomycota (14.87 %), Phragmoplastophyta (10.64 %), Chytridiomycota (7.65 %), Cercozoa (4.24 %), Vertebrata (3.70 %), Dinoflagellata (2.06 %) and Arthropoda (1.92 %) (**Figure 1.6.e**). At site 2, the most dominant eucaryal phyla were Phragmoplastophyta (25.99 %), Ascomycota (21.69 %), Basidiomycota (19.20 %), Arthropoda (5.88 %), Cercozoa (4.47 %), Nematozoa (2.99 %), Mucoromycota (2.51 %) and Metazoa (2.37 %) (**Figure 1.6.f**). Some of these phylum (e.g., Basidiomycota, Nematozoa, Vertebrata, Dinoflagellata, Mucoromycota) changed significantly in relative abundance with soil depth at site 1 (**Table S11**). The most dominant eukaryotic genera at site 1 belonged to Geranomyces (~ 4.42 %), Archaeorhizomyces (~ 1.69 %) and Oikomonas (~ 1.05 %, only detected in the grounwater) while the most abundant eukaryote genera at site 2 were Venturia (~ 1.97 %) and Archaeorhizomyces (~ 1.66 %) (**Figure S6e** and **S6f**). The unidentified genera (referred to as the unknown genera) accounted for up to 80 % of the average relative abundance across all samples analysed. At site 2, we found less Phragmophyta (1.73 % vs 80.92% and 29.53%), more Ascomycota (51.12 % vs 5.83 % and 13.18 %), and more Arthropoda (31.11 % vs 0.82 % and 1.83 %) in the groundwater than in the geological material. In addition, some abundant eukaryotic taxa were solely detected in the rocks and geological material and not in the groundwater (e.g., Geranomyces, site 1 and Archaeorhizomyces, site 2).

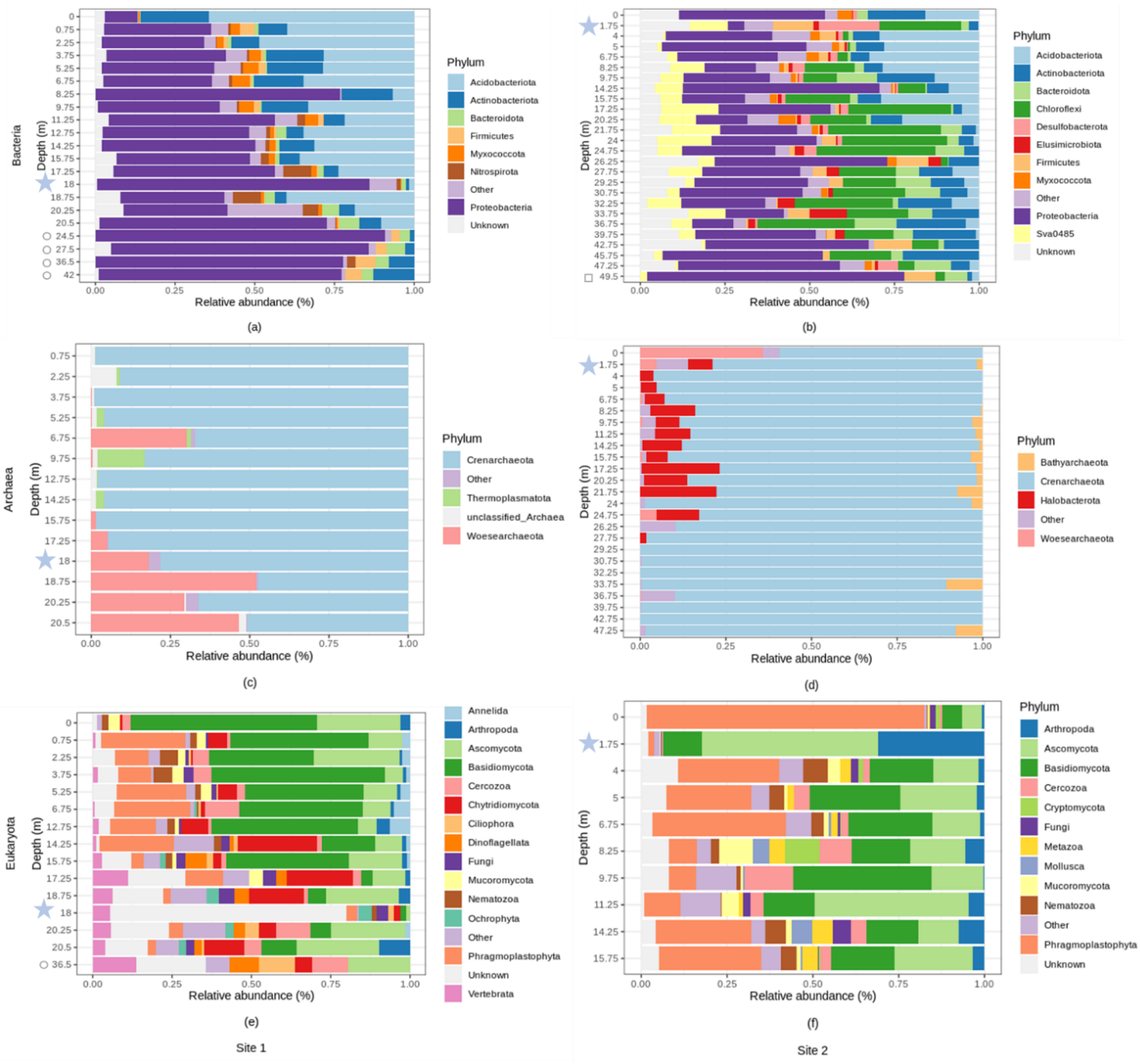


Figure 1.6. Relative abundance (%) of each microbial phylum across different depths (m) at site 1 and at site 2 . With: (a) bacterial relative abundance at site 1, (b) bacterial relative abundance at site 2, (c) archaeal relative abundance at site 1, (d) archaeal relative abundance at site 2, (e) eukaryotic relative abundance at site 1, (f) eukaryal relative abundance at site 2. Phyla with an average relative abundance of less than 1 % were categorized as "Other". Bedrock samples are represented by hollow symbols. Groundwater samples are located at 18 m and 1.75 m of depth, at site 1 and 2 respectively and are represented by blue stars.

1.3.7 Interactions among Bacterial, Eukaryotic and Archaeal major ASVs

There were significant positive, negative and neutral correlations between the most abundant ASVs of the three domains (**Figure 1.7**). Most of these interactions were positive, between bacterial and eukaryotic ASVs (especially at site 1) while they were mostly negative between archaeal and eukaryotic dominant ASVs at site 2 (**Figure 1.7.f**). Bacterial and eukaryotic interactions prediminantly occurred between ASVs belonging to the phylum Phragmoplastophyta (e.g., Euk312 at site 1 or Euk-274 or/and Euk-263 at site 2) or the phylum Basidiomycota (e.g., ASV-Euk 716 at site 1) (**Figure 1.7.a** and **1.7.d**). Bacterial ASV Bac-1974 (site 1) and 3764 (site 2) belonging to the Xanthobacteraceaea family were strongly positively correlated with most of the eukaryotic ASVs (**Figure 1.7.a** and **1.7.d**). In contrast, Acinetobacter spp. (Bac-2469 at site 1, Bac-1775, Bac-1785, Bac-1773 at site 2) were negatively correlated with most eukaryotic ASVs (**Figure 1.7.a** and **1.7.d**). Most archaeal interactions occurred with the Crenarchaeota Group 1.1c at site 1 (e.g., Arc-4, Arc-291, Arc-6, Arc-276) or were unclassified Crenarchaeota at site 2 (**Figure 1.7.d** and **1.7.e**). These taxa correlated both positively and negatively with bacterial. At site 1, eukaryotic ASV 105 (phylum Chytridiomycota), was strongly negatively correlated with most of the archaeal ASVs (**Figure 1.7.c**).

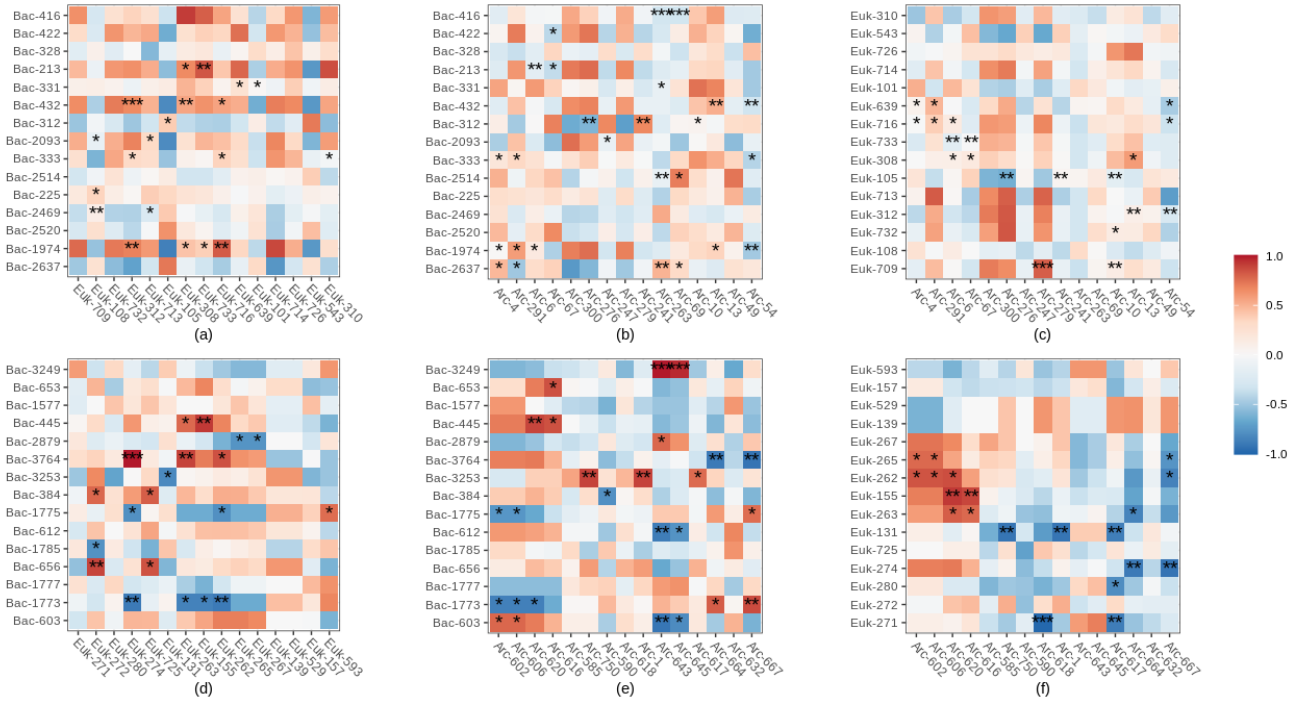


Figure 1.7. Spearman correlations between dominant bacterial, archaeal and eukaryotic ASVs. With correlations between bacterial and eukaryotic ASVs at site 1 (**a**) and site 2 (**d**), bacterial and archaeal ASVs at site 1 (**b**) and site 2 (**e**), eukaryotic and archaeal ASVs at site 1 (**c**) and site 2 (**f**). Both positive (red) and negative (blue) correlations between ASVs were calculated using the relative abundance of dominant ASV. Data were filtered to remove the groundwater samples and the samples taken at the surface. Significant correlations are noted with a * (when $p < 0.05$), ** ($p < 0.01$) or *** ($p < 0.001$).

1.3.8 Microbial Source Tracking

According to the FEAST analyses, upper samples contribute to the formation of shallow subsurface bacterial, archaeal and eukaryotic communities (**Figure 1.8**). They contribute on average 48.18 % (site 1) and 61.56 % (site 2) (percentage of the contribution of upper soil horizons) for the formation of archeal communities in the profile (**Figure 1.8.b, e**); 35.91 % (site 1) and 26.26 % (site 2) for the bacterial communities **Figure 1.8.a, d**); and 19.14 % (site 1) and 13.66 % (site 2) for the eukaryotic communities **Figure 1.8.c, f**). The remaining proportions were due to "unknown sources". The contribution of the groundwater community to the geological material/rock communities was extremely low, especially for bacterial and eukaryotic communities (either non existent or below 0.001 %) (**Figure 1.8**).

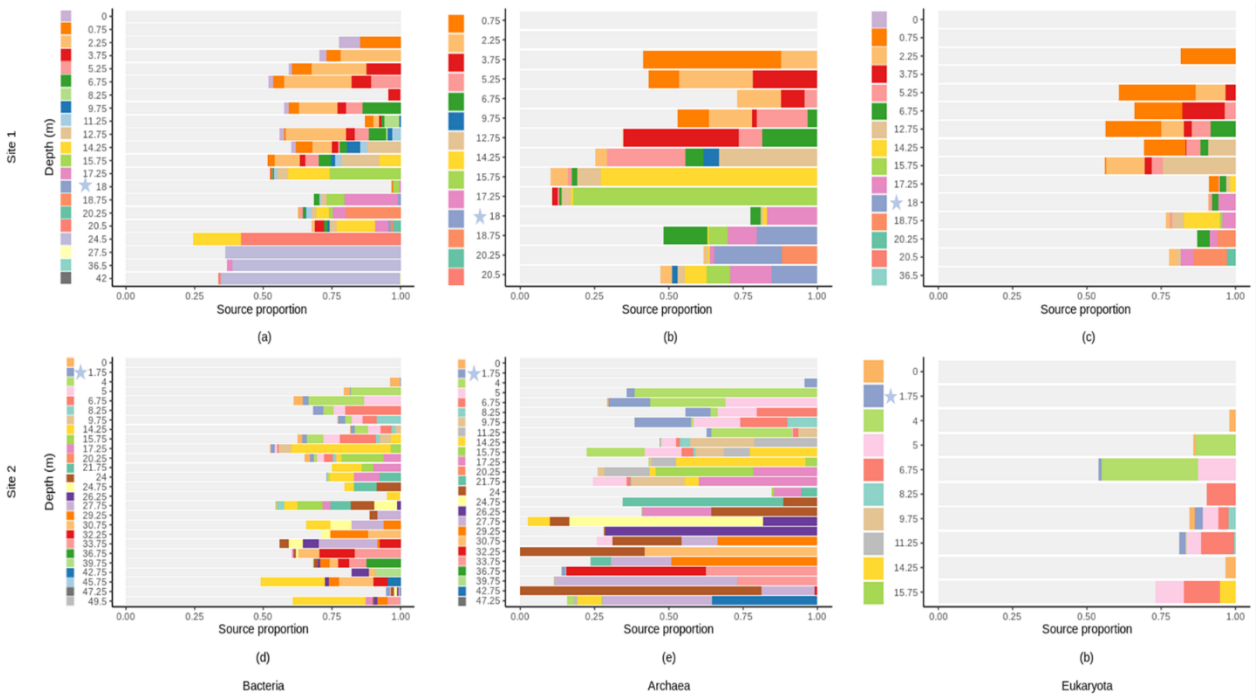


Figure 1.8. FEAST estimations of the proportion of communities from upper layers (source) contributing to the formation of communities in the deeper layer (sink) across different soil depths (m) at site 1 and at site 2. (a, d) for bacterial communities; (b, e) for archaeal communities; (c, f) for eukaryotic communities. The unknown sources are represented in grey. Each color indicates the sources which are the microbial communities from each sample (e.g., in (**a**), the sample collected at depth 2.25 m has sources which are the microbial communities from the upper layers (the surface soil at depth 0 m (purple) and the soil layer at depth 0.75 m (orange)). Groundwater samples are located at 18 m and 1.75 m of depth, at site 1 and 2 respectively and are represented by blue stars.

1.4 Discussion

1.4.1 The Effect of Depth on Endolithic and Sessile Communities

Our study showed that both bacterial taxonomic α-diversity and absolute abundancedecreased with soil depth. Archaeal taxonomic α-diversity followed the same pattern, while eukaryotic diversity did not differ significantly between soil depths at both sites. The decrease in bacterial taxonomic α-diversity with depth is commonly described (Tripathi *et al*., 2018; Fierer *et al*., 2003; LaMontagne *et al*., 2003; Agnelli *et al*., 2004; Goberna *et al*., 2005; Will *et al*., 2010; Baldrian *et al*., 2012; Ko *et al*., 2017; Hansel *et al*., 2008; Feng *et al*., 2019; Xu *et al*., 2021; Kim *et al*., 2016), whereas studies investigating the vertical pattern of archaea tend to show no consistent conclusion (Feng *et al*., 2019, Tripathi *et al*., 2018, Eilers *et al*., 2012, Cao *et al*., 2012). However, these studies only focused on the first centimeters or meters of subsurface soils. Our result suggests that changes in abiotic factors with depth represent a strong ecological filter (Ko *et al*., 2017). Therefore, most of the bacteria and archaea that live in the shallow terrestrial subsurface are less likely to thrive in deeper environments (Tripathi *et al*., 2018). In these deeper environment, bacteria, and archaea (even if they are less numerous and less diversified) probably have physiologies and properties that allow them to be more resistant to external stresses (e.g., better ability to retain water, sorb nutrients and protect against changes in local geochemistry).We associate our nonsignificant results for eukaryotic taxonomic α-diversity with the fact that we were not able to analyze the deeper samples due to very low sequence numbers. As supported by our inability to measure absolute abundances via digital PCR, this may be an indication that the eukaryotic microbial populations arevery small or absent in the deepest samples. In these deeper subsurface horizons, eukaryotes are probably limited in space, nutrients, are unable to cope with oxygen limitations, or some combination thereof (Akob and Küsel, 2011).

The variation in phylogenetic α-diversity patterns of microbial communities along soil depths showed contrasting results depending on the targeted microbial domains, the study site, and the index used for α-diversity measuring. Bacterial phylogenetic α-diversity along the profile tended to decrease when measured with Faith's PD index but increased with MNTD. For Archaea, only Faith's PD was correlated with depth, but patterns were opposite in the studied sites (an increase in the first study site (~ 0.75-20 m) and a decrease in the second study site (~ 0-47 m)). In addition, eukaryotic phylogenetic α-diversity did not show any trend with depth. Opposing pattern of phylogenetic diversity metrics might simply reflect differences in metric properties (Mazel *et al*., 2016). The index PD estimates the phylogenetic diversity of a community as the sum of the tree branch lengths connecting all species (Caruso *et al*., 2019). As a result,

Faith's PD and species richness are also often highly correlated (Tucker *et al*., 2017). This fact probably explains the behaviour of this index. The MNTD index is the average phylogenetic distance between any taxon and its closest relative (Caruso *et al*., 2019). This measure of phylogenetic diversity increased for bacteria in the deeper horizons. Similar results were observed by Chu et al. (2016), and this difference was linked to an increased importance of environmental filtering in surface soils in comparison to deep horizons (Chu *et al*., 2016). However, based on our results regarding geological material characteristics (which showed higher content of total carbon, total organic carbon, and total inorganic carbon in the surface soil sample than in the subsurface sample) and taxonomic α-diversity (which suggested that depth act as a strong ecological filter), we consider this unlikely to be the case here. Coexistence of phylogenetically distant bacteria in stressful and resource-poor sites— like in the deeper horizons of the shallow terrestrial subsurface—may suggest that competition between close relatives is exerting great selection pressure. Together with taxonomic diversity result, these results suggest that a small number (that is, low taxonomic α-diversity and low absolute abundance) of phylogenetically distant bacterial taxa can or must co-exist to survive the stressful conditions of the deeper horizons of the shallow terrestrial subsurface. In oligotrophic habitats, lack of nutrients can lead to a reduction in genome size by loss of expendable genes (Herrmann *et al*., 2019). Taken to extremes, this can lead to the loss of essential metabolic functions, which inevitably leads to dependencies on other possibly phylogenetically distant organisms (Herrmann *et al*., 2019). The fact that MNTD index showed no correlation with depth for both archaea and eukaryotes suggest no contributions of history and trait evolution to this community structure (Anacker and Harrisson, 2012). This result may suggest that (within the depth interval considered), archaeal and eukaryotic community membership is not limited by a requirement for shared stress tolerance traits, or that there are no dependencies on other phylogenetically distant organisms, or no strong competition between phylogenetically close organisms (Anacker and Harrisson, 2012). In addition, the majority of the standardized effect sizes of MNTD and Faith's values obtained for bacteria, archaea and eukaryotes using the null model were below zero, which shows that the microbial communities tended to be more phylogenetically clustered than would be expected by chance (Kembel *et al*., 2011). Therefore, in agreement with our taxonomic α-diversity results, microbial communities from the shallow terrestrial subsurface are structured by environmental filtering (Chu *et al*., 2016).

1.4.2 Effect of Abiotic Characteristics on Endolithic and Sessile Communities

The studied profiles from both sites (0 – 50 m) represent strong environmental gradients, with multiple abiotic factors changing with depth. One of the most pronounced changes through the profiles was the increase in pH with depth (toward more basic pH) and the relatively low nitrogen and carbon organic quantity. In the studied sites, sedimentary texture does not seem to be a major factor influencing the α-diversity patterns as opposite sedimentary texture patterns (coarser and looser at site 1 vs. finer and denser at site 2) are not reflected by opposite or different α-diversity patterns. Yet soil texture can affect microbial movement through soils and finer soils, especially clay minerals, increase water and nutrient retention which should increase survival time (Abu-Ashour *et al*., 1994; Hamarashid *et al*., 2010). Hence, if sedimentary texture was the predominant factor influencing microbial communities, the deeper and finer soil horizons at site 2 should have contained a higher diversity than the shallower and coarser soil horizons. Our results suggest that other factors are predominant in controlling the α-diversity of the subsurface microorganisms in the studied sites. We found significant relationships of bacterial absolute abundance and some of the bacterial and archaeal α-diversity metrics with abiotic factors such as pH, the % of total organic carbon, the % of total inorganic carbon, or/and the % of total carbon. Concentrations of total inorganic carbon tended to show a negative correlation with bacterial and archaeal taxonomic α-diversity. Bacterial and archaeal taxonomic α-diversity tended to show a negative correlation with the concentration of total inorganic carbon. This suggests that bacteria and archaea from the shallow terrestrial subsurface mostly acquire energy from organic carbon. However, because all these variables are moderately to strongly correlated with depth, we do not know if the changes induced by these variables are the sole factor driving the observed α-diversity patterns. Soil microbial communities may also be directly or indirectly affected by other factors such as water content, oxygen quantity, trophic status, spatial and climate factors (Cao *et al*., 2012; Hansel *et al*., 2008).

The bacterial, archaeal, and eukaryotic community compositions showed differences with depth through the whole profiles and mineral composition does not appear to play a role in the structure of microbial communities. In accordance with our results, it has been reported that soil microbial communities are strongly shaped by soil properties such as nutrient availability, pH, and soil texture, which vary considerably with soil depth (Griffiths *et al*., 2003; Hansel *et al*., 2008; Li *et al*., 2019). Changes in the composition of bacterial communities with pH, especially at site 1, is not surprising given the

predominance of members of the Acidobacteria phylum (e.g., *Candidatus Koribacter*, *Acidothermus*, *Bryobacter*). Although Acidobacteria are generally oligotrophic and might be more adapted to the environment in deep layers (Li *et al*., 2019), there is some evidence to suggest that their abundance within a community is primarily regulated by pH (Jones *et al*., 2009).

The overall structure of the bacterial communities also changed markedly with geological material texture, with the most pronounced changes occurring within the bedrock (site 1). The taxonomic α-diversity in site 1 also suddenly decreased at the level of the bedrock. Together, these results suggest that the conditions in the bedrock changed markedly and that many surface-dwelling bacteria are less likely to thrive in this environment. We associate this trend withthe little space for water and microbe colonization that the bedrock usually provides (Akob and Küsel, 2011). The increase in the relative abundance of Ralstonia, a member of Proteobacteria in this environment is probably due to their capabilities to cope with hostile life conditions, such as extreme pH or oligotrophic environments (Govil *et al*., 2019). As reported by many previous studies, ammonia-oxidizing archaea dominated the profiles (e.g., Hansel *et al*., 2008; Cao *et al*., 2012), which may be driving autotrophic nitrification in the deeper depths (Hansel *et al*., 2008; Cao *et al*., 2012). The relative abundanceof the Crenarchaeote phylum members decreased at site 1 but increased at site 2. Their decrease might be explained by the presence of group 1.1c at the first study site, a group that prefers acidic soils and has even been detected at a pH below 3 (Cao *et al*., 2012). This different trend suggests that changes in microbial community composition with depth are, to some degree, site specific and dependent on specific characteristics of the profiles being studied. At site 2, H_2/CO_2-utilizing autotrophic methanogens were identified (*Methanoregula*, *Methanosarcina*, Rice Cluster II), possibly linked to the higher total inorganic carbon content at this site. The presence of the methanotroph cand. Methanoperendes suggests a coupling between methane production, and nitrogen cycle (Bräuer *et al*., 2011; Mondav *et al*., 2014; Haroon *et al*., 2013).

The eukaryotic community composition was was not primarily controlled by depth. Instead, carbon content best explained the observed patterns. Predominant eukaryotic taxa at both study sites were mainly assigned to Basidiomycetes and Ascomycetes phyla, a result also observed by Hartmann et al. (2009) and Xu et al. (2021). The eukaryotic community might be controlled by other abiotic factors such as soil moisture (Li *et al*., 2017), or by vegetation types as suggested by Xu et al. (2021). Plant communities may play a major role in structuring eukaryotic microbial community composition, particularly for eukaryotic communities dominated by fungi (Shen *et al*., 2014). The decrease of the relative abundance of

Basidiomycota can probably be explained by the fact that the representatives of these fungi can form symbiotic relationship with plants (Xu *et al*., 2021). However, the decrease of Basidiomycota was only detected at site 1 and other studies have showed inconsistent pattern of fungal composition along soil depths (Kang *et al*., 2021). The inconsistent pattern may suggest that site-specific characteristics are also important determinants for shaping the fungal communities. Further studies on the ecology of these diverse eukaryotic taxa in the shallow terrestrial subsurface are clearly needed (Chen *et al*., 2019).

The detection of Vertebrata, Metazoa and Nematozoa in the geological material in deep soil horizons is surprising even if multicellular life in the deep subsurface has been previously detected in this habitat (Borgonie *et al.,* 2011). The presence of these organisms in the terrestrial subsurface can mean that the material examined contained the remains of the above animals passively infiltrating through the depth. Overall, this result highlights the importance of the surface community's role in structuring the diversity and composition of subsurface communities as many other organisms can passively be brought from the surface to the terrestrial subsurface (e.g., plants, fungi).

1.4.3 Potential Biotic Interactions between Bacterial, Eukaryotic and Archaeal Microorganisms

Soil microorganisms coexist in complex associations, including mutualism, competition, predation or neutral (Wan *et al*., 2020). In the shallow terrestrial subsurface, we found some positives correlations between bacterial-eukaryotic microorganisms while some correlations between archaeal-eukaryotic microorganisms were negatives (especially at site 2). Positive correlations may indicate the occupation of different niches (Mackenzie *et al*., 2019), cooperative or mutualistic potential associations such as cross-feeding and/or syntrophic relationships (Wan *et al*., 2020). Eukaryotic organisms are unable to obtain fixed nitrogen without nitrogen-fixing prokaryotes (Kneip *et al*., 2007). Hence, positive interactions detected between eukaryotic organisms and nitrogen-fixing bacteria, such as the Xanthobacteraceae family (e.g., ASV Bac-1974 at site 1 and Bac-3764 at site 2) suggest symbiotic interactions between these two domains (Kneip *et al*., 2007). Negative correlations could reveal antagonistic associations among species, such as competition for limited resources between archaeal and eukaryotic microorganisms, but it could also point out to predation and/or parasitisms interactions (Kneip *et al*., 2007). Some chytrids, for example are host-specific parasitic fungi, and may have a considerable negative impact on archaeal cells (Ibelings *et al*., 2004). Allelopathy might also be an important phenomenon in the subsurface. Soil microorganisms can produce and release allelochemicals and affect many organisms (either negatively or positively) like plants, algae, fungi (especially mycorrhizae, or pathogens), or nitrogen cycle bacteria (Reigosa *et al*., 1999). Our

data contained a large number of unknown taxa at the genus-level. Precise identification is critical to gain more precisions on the nature of the biotic interactions between the three domains. We also did not take into account the biotic interactions within each domain (e.g., bacterial-bacterial), but negative, positive and neutral correlations can also occur within domains and have a big influence on the whole community's structure.

1.4.4 Vertical Fluid Fluxes as Sources of Microbial Communities in the Shallow Terrestrial Subsurface

A major challenge of analyzing the compositional structure of microbial communities is to identify their potential origins and sources. The composition of each microbial community is typically composed of the members of several environmental sources, including different contaminants as well as other microbial communities that interacted with the sampled habitat (Shenhav *et al*., 2019). In the shallow subsurface, members of the microbial communities can be long-term descendants of microorganisms that colonized the geological material during deposition (Kieft *et al*., 1998), or represent younger/recent surface-sourced colonists or temporal survivors that disperse by diffusion, or by vertical and lateral fluid fluxes (Amy *et al*., 1992; Lazar *et al*., 2019; Kieft *et al*., 1998). Our objective was to estimate the proportion of shallow subsurface microbial communities that originated through vertical fluid fluxes by estimating the proportion of upper horizons (sources) contributing to the formation of the deeper soil layers (sinks). Our FEAST analysis confirmed a vertical colonization of the geological material and of the bedrock in the shallow terrestrial subsurface as upper horizons communities were sources for deeper layers (sinks).

The higher contribution of upper samples to the formation of archaeal communities might be explained by their better tolerance to extreme environments (Xu *et al*., 2021), which allow better survival in deeper layers from surface colonization. Downward fluid fluxes might be an additional explanation for the decrease in bacterial and archaeal taxonomic α-diversity and/or absolute abundance with depth. Indeed, microorganisms arriving in deeper horizons from the surface are more likely adapted to live under the surface or near surface conditions. However, the other microbial sources, collectively referred to as the "unknown source" still represent an important proportion of potential sources, especially for subsurface eukaryotic communities. Other potential sources might come from plants, or rainwater and more work is clearly needed (especially for eukaryotes) to unravel the origins of these complex subsurface microbial communities.

Another potential source of microbial cells might be the community found in the groundwater. However, the FEAST analyses suggest little to no contribution from planktonic populations to the overall sessile/endolithic community. Both communities have different compositions despite the presence of similar taxa which might suggest an exchange between the two communities. On the other hand, some microbial taxa were exclusively found in the groundwater or in the surrounding geological material and bedrock. This is the case with the heterotrophic nanoflagellate *Oikomonas*, or the potential nitrate-reducing Marine Group II archaea, who show a preference for the planktonic form as it they were found solely in the groundwater. In addition, differences in composition can partly be explained by the differences in abiotic factors in water compared to surrounding geological material and bedrock (Rinke *et al*., 2019; Cavalier-Smith *et al*., 1996). For example, the pH is closer to neutrality in the groundwater in comparison to surrounding geological material and bedrock (Table S1 and S2). However, other characteristics could also affect these communities (e.g., connectivity, nutrient inputs, etc.). Our results do not allow us to conclude whether rock-associated are significantly different from their planktonic counterparts. However, bar plot (**Figure 1.6**) and the nMDS analysis allowed us to observe marked differences in the composition of these two communities and other authors such as Lazar et al. (2019) suggested the existence of a unique rock matrix microbiome compared to the surrounding groundwater. In accordance with Griebler et al. (2002), our results underline the importance of sampling both the attached and the suspended communities when studies on all the communities of the aquifer are performed. Studies and predictions on the functioning of subsurface ecosystems based solely on groundwater samples might not include some taxa (although they might be rare) and therefore might be subject to misinterpretation (Lazar *et al*., 2019). A comparative study with a larger number of subsurface water samples and surrounding rocks is needed to draw robust conclusions about the existence of different communities in the aquifer (planktonic vs. rock-associated microbiome).

1.5 Conclusions

In the studied sites, many abiotic factors changed with depth (including soil texture, pH, carbon content, nitrogen content and mineralogy). Depth can therefore be seen as an ecological and phylogenetic filter for the subsurface microbial communities. Thus, most bacteria, archaea (and probably eukaryotes, although our data cannot confirm) are less likely to thrive in the deeper horizons. In addition, the significant effect of depth could be explained by vertical flow movements in the subsurface layers. Our results suggest that these vertical movements supply the earth's subsurface with microorganisms. It is therefore very likely that these communities—which are transported from the surface to the deeper soil horizons—are less adapted to the conditions of the deeper horizons and may use various survival strategies. We found that, in the deeper horizons, the competition is strong for the phylogenetically close taxa and that therefore, the cohabitation of a reduced number of phylogenetically distant bacterial taxa is an advantage (different ecological niche/form of co-dependency). Our results also suggest that cooperative or mutualistic potential associations between bacteria and microbial eukaryotes occur, such as cross-feeding and/or syntrophic relationships in the terrestrial subsurface.

The high heterogeneity of soil makes it difficult to achieve a systematic understanding of the microbial community distribution in subsurface soils across large-scale regions (Cao *et al*., 2012). To get a more comprehensive insight into the distribution and biogeography, particular attention should be paid to eukaryotic communities, and further studies using samples from different habitats and at a larger spatial scale are needed in the future (Cao *et al*., 2012). We believe that this study represents a step toward an insightful understanding of the overall assemblage processes affecting microbial communities in the shallow terrestrial subsurface

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, **Figure S1**: Bacterial, archaeal and eukaryotic α-diversity indexes along depth (m) at site 1 (black, circle) and site 2 (grey, square). With: (a, e, i) Shannon; (b, f, j) Simpson index; (c, g, k) Faith's PD and (d, h, l) MNTD index. Bedrock samples are represented by hollow symbols. Groundwater samples are represented by stars., **Figure S2**: Change in the relative abundance of most abundant bacterial phyla with depth at site 1., **Figure S3**: Change in the relative abundance of most abundant bacterial phyla with depth at site 2., **Figure S4**: Change in the relative abundance of most abundant bacterial genera with depth at site 1., **Figure S5**: Change in the relative abundance of most abundant bacterial genera with depth at site 2., **Figure S6**: Relative abundance (%) of each microbial genera across different depths (m) at site 1 and at site 2. With: (a) bacterial relative abundance at site 1, (b) bacterial relative abundance at site 2, (c) archaeal relative abundance at site 1, (d) archaeal relative abundance at site 2, (e) eukaryotic relative abundance at site 1, (f) eukaryotic relative abundance at site 2. Genera with an average relative abundance of less than 1 % were categorized as "Other". Bedrock samples are represented by hollow circles and hollow squares. Groundwater samples are represented by blue stars., **Figure S7**: Change in the relative abundance of most abundant archaeal genera with depth at site 1. Unc: uncultured., **Figure S8**: Change in the relative abundance of most abundant archaeal genera with depth at site 2. Unc. B : Uncultured Bathyarchaeia., **Table S1**: Abiotic characteristics measured along depth gradient at site 1., **Table S2**: Abiotic characteristics measured along depth gradient at site 2., **Table S3:** The correlations (r) determined by Spearman Correlation between abiotic characteristics and depth at site 1 and site 2., **Table S4:** XRD results showing the mineralogical composition (in %) of the samples collected at site 1., **Table S5**: XRD results showing the mineralogical composition (in %) of the samples collected at site 2., **Table S6:** Spearman correlation between bacterial α-diversity metrics and characteristics of geological material at site 1 and site 2., **Table S7:** Spearman Correlation between archaeal α-diversity metrics and characteristics of geological material at site 1 and site 2., **Table S8:** Spearman Correlation between eukaryotic α-diversity metrics and characteristics of geological material at site 1 and site 2., **Table S9:** Mantel and ANOSIM test between bacteria ß-diversity metrics and characteristics of geological material., **Table S10:** Spearman Correlation between absolute abundance and characteristics of geological material at site 1 and site 2., **Table S11**: Spearman correlation between relative abundances of most abundant bacterial, archaeal, and eukaryotic phyla and genera with characteristics of geological material at site 1 and site 2.

Author Contributions: Conceptualization, C.S.L.; methodology, J.M.; validation, J.M., C.S.L. and M.L.; formal analysis, J.M.; investigation, J.M. and S.Z.; resources, C.S.L.; data curation, J.M.; writing—original draft preparation, J.M.; writing—review and editing, C.S.L. and M.L.; visualization, J.M.; supervision, C.S.L.; project administration, C.S.L.; funding acquisition, C.S.L. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the 'Environmental Aquatic Genomics' Canada Research Chair and the NSERC-Discovery Grant RGPIN-2019-06670 awarded to C.S.L. This research was funded by the Quebec Ministry of Environment and Fight against climate change awarded to M.L.

Data Availability Statement: The obtained sequences were deposited in the National Center for Biotechnology Information (NCBI) under the BioProject ID: PRJNA758373.

Acknowledgments: The authors would like to thank Sylvain Gagné for his invaluable help in collecting the samples. We would also like to acknowledge the involvement of Lytana Lécuyer in the analysis of the Eukaryotic datasets.

Conflicts of Interest: The authors declare no conflict of interest.

DISCUSSION GÉNÉRALE

De nombreux facteurs abiotiques changent avec la profondeur (p. ex., texture, pH, contenu en carbone, en azote, minéralogie, etc.). Cette dernière peut donc être vue comme un filtre écologique et phylogénétique sur les communautés microbiennes retrouvées dans la subsurface terrestre superficielle. Ainsi, la plupart des bactéries, des archées (et probablement des eucaryotes, même si nos données ne permettent pas de le confirmer) sont moins susceptibles de prospérer dans les horizons plus profonds de la subsurface terrestre superficielle. En complément, l'effet marqué de la profondeur sur la diversité et la composition des communautés présentes dans la subsurface terrestre superficielle pourrait s'expliquer par les mouvements des flux verticaux. Ces mouvements verticaux approvisionnent la subsurface terrestre superficielle en microorganismes. Il est donc fort probable que ces communautés - qui sont transportées de la surface vers les horizons pédologiques plus profonds - soient moins adaptées aux conditions des horizons plus profonds et fassent usage de diverses stratégies de survie. Nos résultats suggèrent que, dans les horizons plus profonds, la compétition est forte pour les taxons phylogénétiquement proches et que par conséquent, la cohabitation d'un nombre réduit de taxons bactériens phylogénétiquement éloignés est un avantage (niches écologiques différentes/ forme de codépendance). Nos résultats montrent également qu'il existe de nombreuses interations positives entre les bactéries et les eucaryotes ce qui suggère la présence d'associations de coopération ou de mutualisme (p. ex., cross-feeding, relations synthrophiques) entre les organismes de ces deux domaines.

Nous avons observé que la diversité α et la composition des communautés microbiennes étaient influencées par des facteurs abiotiques comme le pH. Bien que le pH varie avec la profondeur comme de nombreux autres paramètres abiotiques, nous nous attendons à ce qu'il soit un facteur majeur influençant la diversité et la composition des communautés microbiennes (en particulier bactérienne) dans les profils de sol étudiés. En effet, le pH a tendance à augmenter de la neutralité vers des pH extrêmes dans les sites à l'étude. Or, le pH intracellulaire de la plupart des microorganismes se situe proche d'un pH neutre (pH = 7 ± 1) (Fierer et Jackson, 2006). Toute déviation du pH environnemental (extracellulaire) devrait imposer un stress à ces derniers (Fierer et Jackson, 2006). Un pH extrême affecte la structure de toutes les macromolécules (Parket *et al.*, 2017). Il rompt les liaisons hydrogène qui maintiennent ensemble les brins d'ADN, hydrolyse les lipides, dénature les protéines et altère la production d'ATP (Parket *et al*., 2017). De plus, le pH affecte les conditions environnementales extracellulaires en modulant les réactions géochimiques et l'activité chimique des protons (des acteurs clés dans les réactions redox, la dissolution

et la précipitation des minéraux). En retour, ces réactions déterminent la salinité, la composition des solutions aqueuses et contrôlent la biodisponibilité des nutriments ainsi que celle des oligoéléments. Le pH affecte également les activités des enzymes extracellulaires et la réactivité de la matière organique naturelle (Jin et Kirk, 2018).

Les analyses de cette étude ont été réalisées sur deux profils de sol provenant de sites relativement similaires (p. ex., climat, localisation, végétation et type de sol). Par conséquent, nos conclusions ne peuvent pas être généralisées à l'ensemble de la subsurface terrestre superficielle. D'autres études utilisant plusieurs réplicas d'échantillons provenant de différents habitats et à une plus grande échelle spatiale sont nécessaires. De plus, la grande hétérogénéité du sol de la subsurface rend difficile la compréhension systématique de la distribution des communautés microbiennes dans cet habitat. Dans notre étude, de nombreux facteurs abiotiques ont été pris en compte (tels que le pH, la minéralogie, le contenu en carbone total, en carbone inorganique total, en carbone organique total et en azote total). Toutefois, comme de nombreux facteurs abiotiques varient simultanément avec la profondeur, il serait pertinent de développer cette expérience avec d'autres paramètres tels que la température, le contenu en eau ou en oxygène. Ceci permettrait d'identifier plus clairement les causes des variations de la diversité, de l'abondance et de la composition des communautés microbiennes tout le long du profil de la subsurface terrestre superficielle.

Une attention particulière devrait être accordée aux communautés eucaryotes, car ce domaine a encore été très peu étudié dans la subsurface. Nous n'avons pas été en mesure d'estimer l'abondance absolue de ces communautés et nos données contenaient un grand nombre de taxons eucaryotes inconnus au niveau taxonomique du genre. Une identification plus précise est essentielle pour obtenir davantage d'informations sur ces communautés. Nous n'avons pas non plus pris en compte les interactions biotiques intradomaines (p. ex., les interactions bactérie-bactérie). Or, des interactions négatives, positives et neutres peuvent également se produire au sein des domaines et avoir une grande influence sur l'ensemble des communautés microbiennes.

Notre étude permet d'acquérir de nouvelles informations concernant l'implication des fluides verticaux dans l'assemblage des communautés de la subsurface terrestre superficielle. Cependant, d'autres sources potentielles non négligeables pourraient provenir d'ailleurs (p. ex., des plantes ou de l'eau de pluie). Par conséquent, des travaux supplémentaires incluant ces sources potentielles sont nécessaires pour

identifier plus amplement les origines des communautés microbiennes de la subsurface terrestre superficielle.

Nos résultats soulignent l'importance d'échantillonner à la fois les communautés attachées et suspendues lorsque des études sur l'ensemble des communautés de l'aquifère sont réalisées. En effet, les études et les prévisions sur le fonctionnement des écosystèmes souterrains basées uniquement sur des échantillons d'eau souterraine pourraient ne pas inclure certains taxons et pourraient donc être sujettes à une mauvaise interprétation (Lazar *et al.*, 2019). Une étude comparative avec un plus grand nombre d'échantillons d'eau souterraine et de roches environnantes est nécessaire pour tirer des conclusions solides sur l'existence de différentes communautés dans l'aquifère (communautés planctoniques vs communautés endolithiques et sessiles). Finalement, il serait intéressant d'élargir cette étude à la subsurface terrestre profonde en récupérant des échantillons à de grandes profondeurs (> 50 m).

Nous pensons que notre étude représente une étape vers une compréhension approfondie des processus d'assemblage globaux affectant les communautés de la subsurface terrestre superficielle. Nous pensons également que cette dernière facilitera une meilleure compréhension des principales communautés du sol (bactéries), mais aussi de la diversité et de la structure des autres communautés importantes (archées et eucaryotes).

ANNEXE A : SUPPLEMENTARY FIGURES

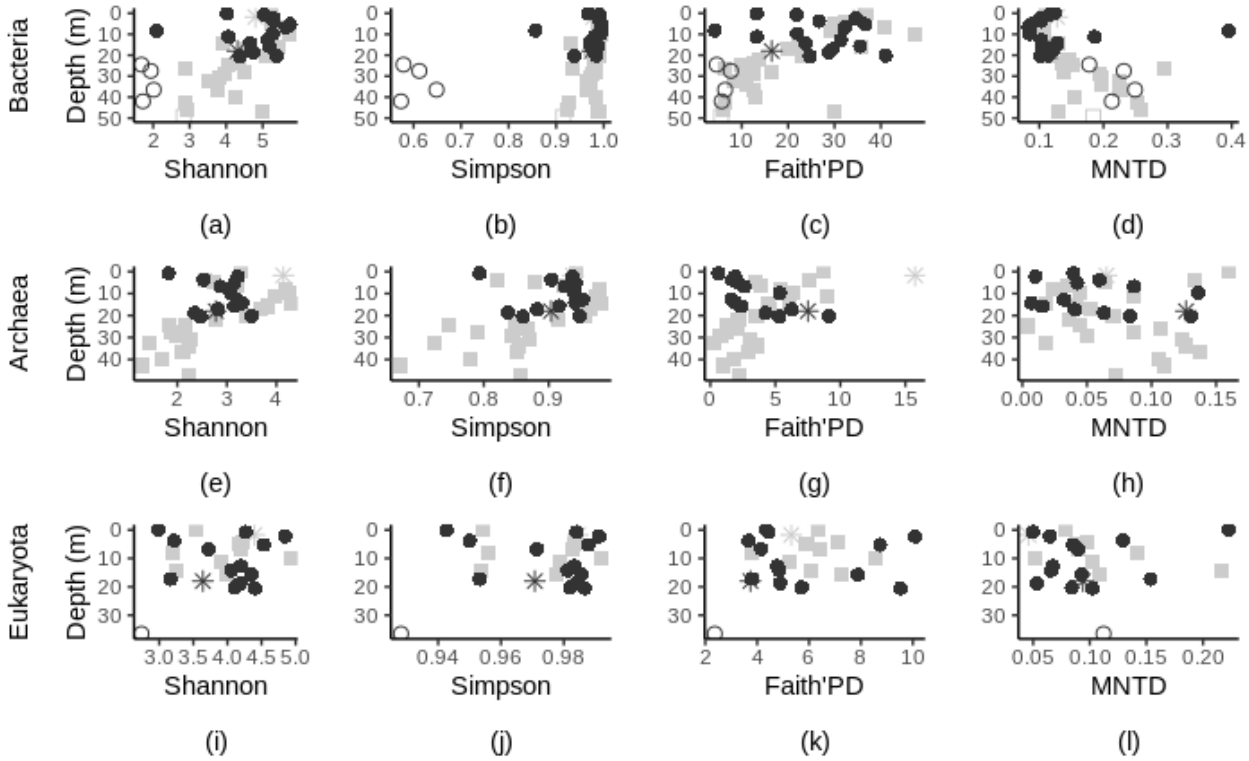


Figure S1. Bacterial, archaeal and eukaryotic α-diversity indexes along depth (m) at site 1 (black, circle) and site 2 (grey, square). With: (a, e, i) Shannon; (b, f, j) Simpson index; (c, g, k) Faith's PD and (d, h, l) MNTD index. Bedrock samples are represented by hollow symbols. Groundwater samples are represented by stars.

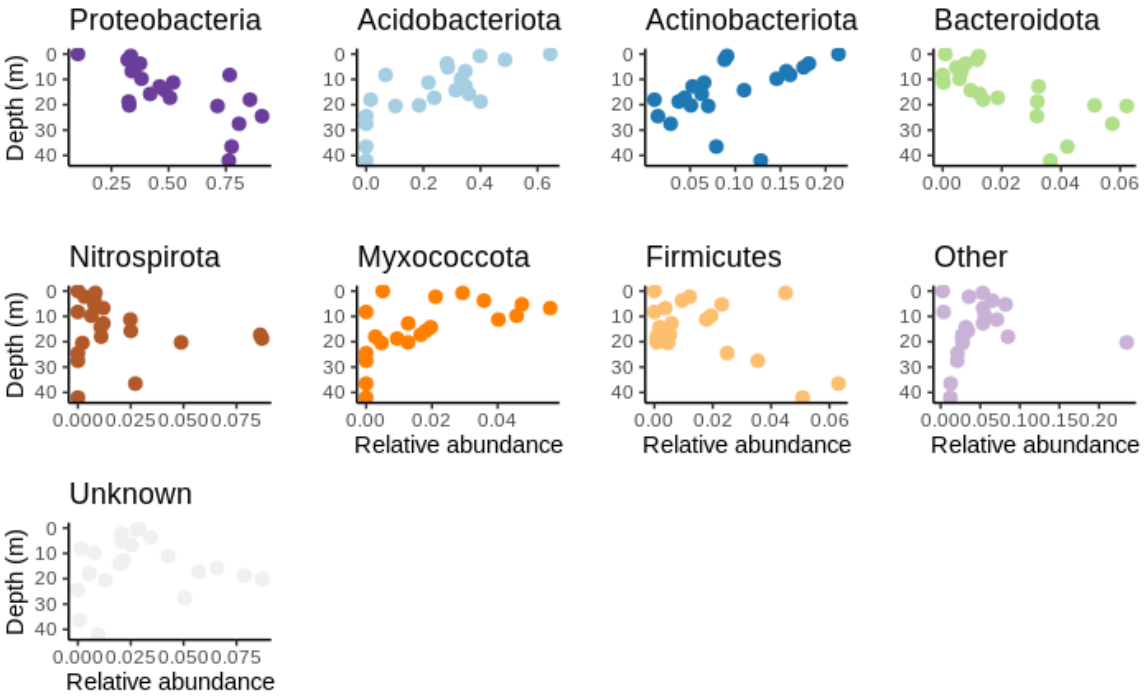


Figure S2. Change in the relative abundance of most abundant bacterial phyla with depth at site 1.

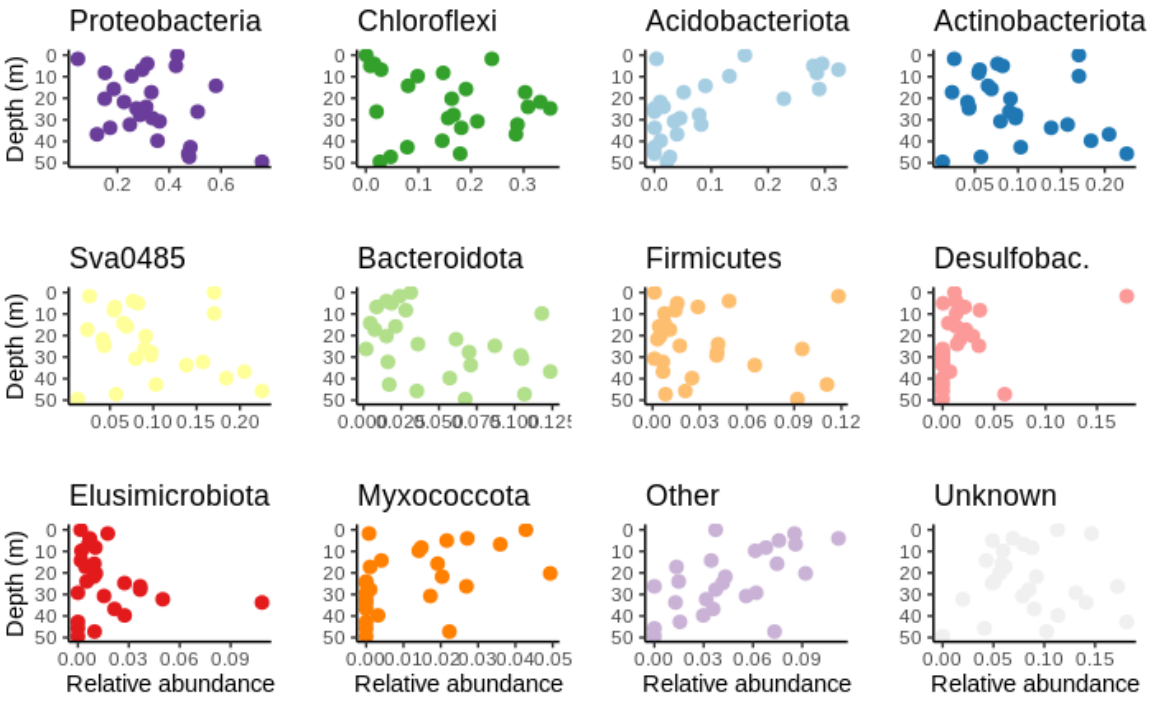


Figure S3. Change in the relative abundance of most abundant bacterial phyla with depth at site 2.

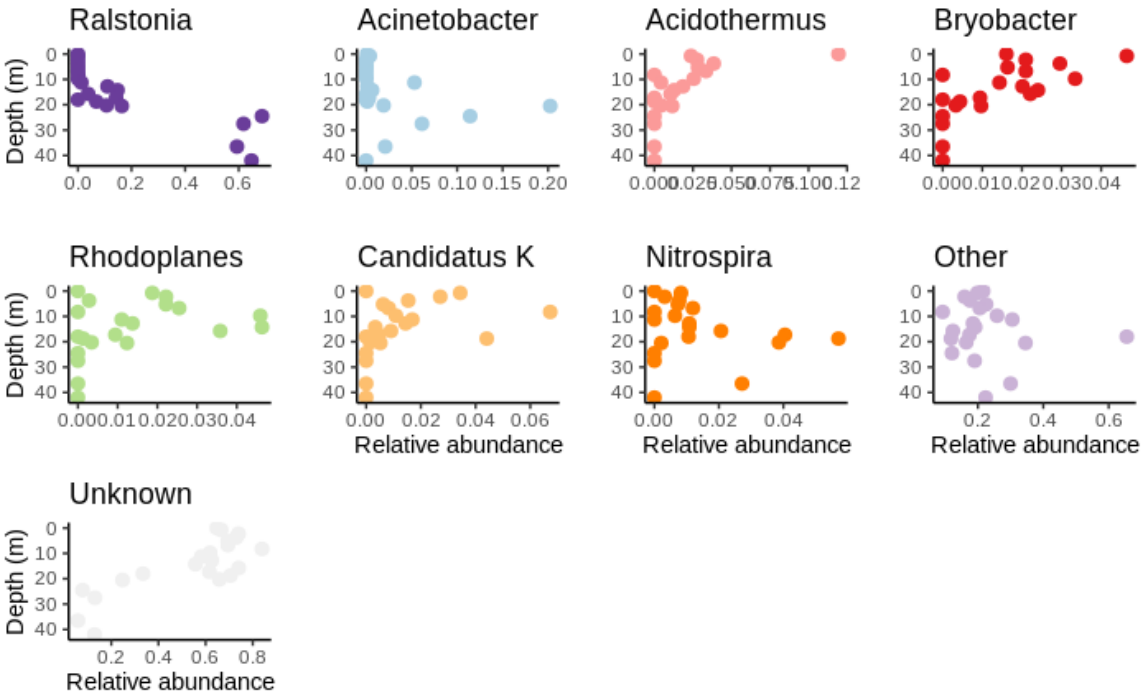


Figure S4. Change in the relative abundance of most abundant bacterial genera with depth at site 1.

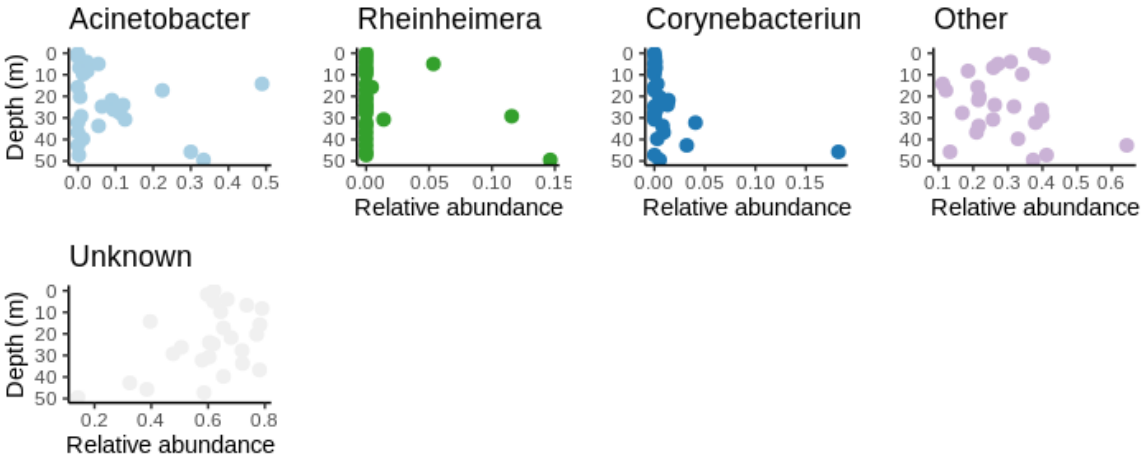


Figure S5. Change in the relative abundance of most abundant bacterial genera with depth at site 2.

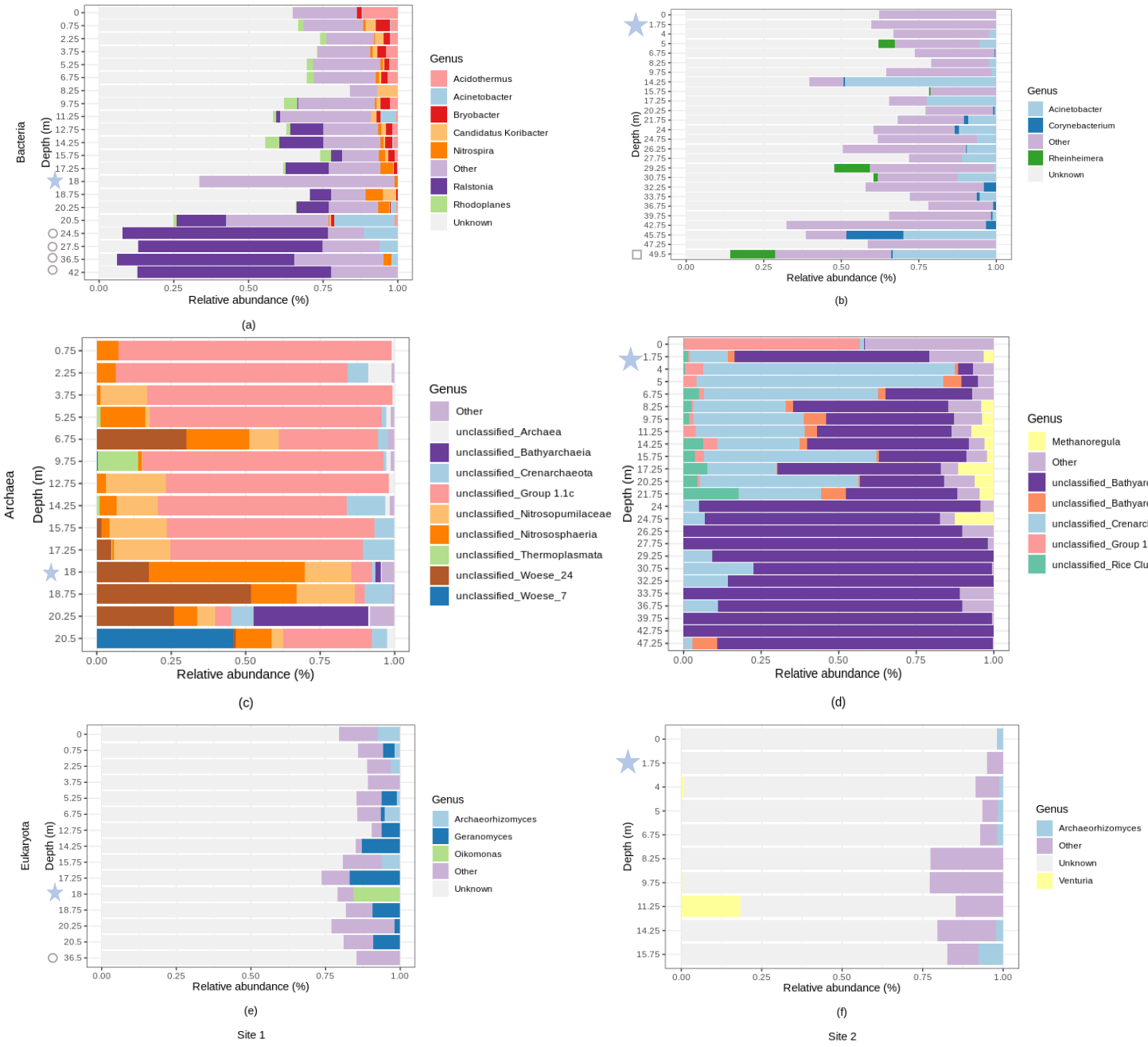


Figure S6. Relative abundance (%) of each microbial genera across different depths (m) at site 1 and at site 2. With: (a) bacterial relative abundance at site 1, (b) bacterial relative abundance at site 2, (c) archaeal relative abundance at site 1, (d) archaeal relative abundance at site 2, (e) eukaryotic relative abundance at site 1, (f) eukaryotic relative abundance at site 2. Genera with an average relative abundance of less than 1 % were categorized as "Other". Bedrock samples are represented by hollow circles and hollow squares. Groundwater samples are represented by blue stars.

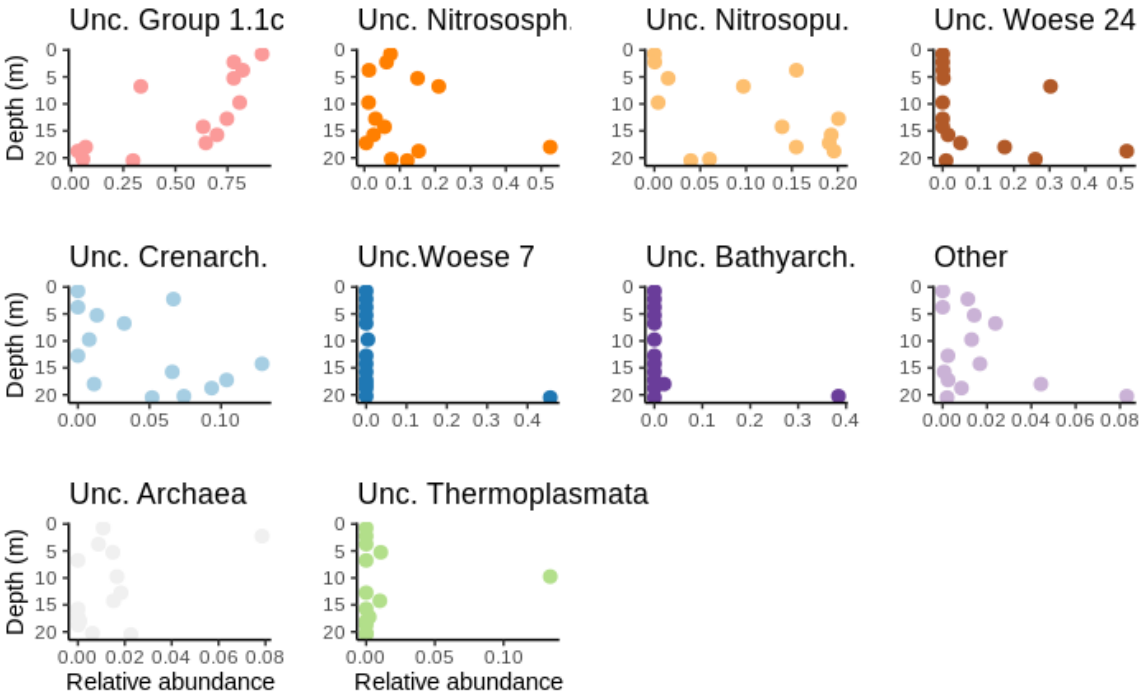


Figure S7. Change in the relative abundance of most abundant archaeal genera with depth at site 1. Unc : uncultured.

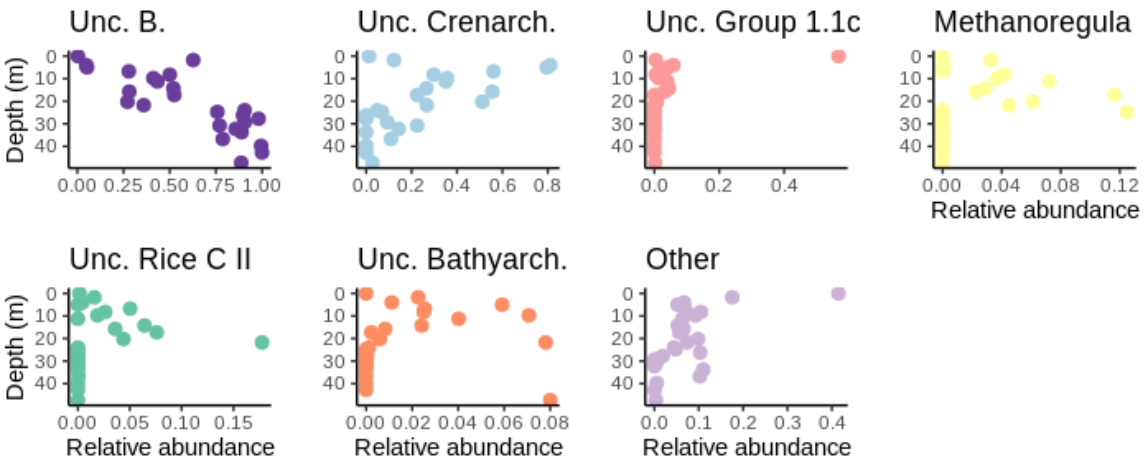


Figure S8. Change in the relative abundance of most abundant archaeal genera with depth at site 2. Unc. B : Uncultured Bathyarchaeia.

ANNEXE B : SUPPLEMENTARY TABLES

Table S1: Abiotic characteristics measured along depth gradient at site 1.

Sample	Depth (m)	Texture	pH	TN (%)	TC (%)	TOC (%)	TIC (%)
RR-01	0	mS	5.16 ± 0.20	1.52	48.24	30.73	17.51
RR-02	0.75	mS	6.53 ± 0.20	0.00	0.04	0.03	0.01
RR-03	2.25	mS	6.43 ± 0.09	0.00	0.05	0.04	0.01
RR-04	3.75	mcSG	6.45 ± 0.13	0.00	0.03	0.03	0.00
RR-05	5.25	mcSG	6.88 ± 0.09	0.00	0.03	0.02	0.01
RR-06	6.75	mcSG	6.84 ± 0.10	0.00	0.03	0.02	0.01
RR-07	8.25	mcSG	6.78 ± 0.05	0.00	0.02	0.00	0.02
RR-08	9.75	mcSG	6.66 ± 0.06	0.00	0.02	0.02	0.00
RR-09	11.25	mcSG	6.62 ± 0.20	0.00	0.06	0.06	0.00
RR-10	12.75	cSG	6.78 ± 0.22	0.00	0.03	0.02	0.01
RR-11	14.25	cSG	6.93 ± 0.08	0.00	0.01	0.00	0.01
RR-12	15.75	cSG	7.07 ± 0.07	0.00	0.02	0.00	0.02
RR-13	17.25	cSG	6.43 ± 0.15	0.00	0.01	0.00	0.01
RR-14	18.75	cSG	6.98 ± 0.09	0.00	0.02	0.00	0.02
RR-15	20.25	cSG	8.11 ± 0.09	0.00	0.02	0.00	0.02
RR-16	20.5	cSG	8.58 ± 0.31	0.00	0.01	0.00	0.01
RR-17	24.5	R	9.14 ± 0.05	0.00	0.04	0.00	0.04
RR-18	27.5	R	9.44 ± 0.07	0.00	0.02	0.00	0.02
RR-19	36.5	R	9.41 ± 0.04	0.00	0.01	0.00	0.01
RR-20	42	R	9.53 ± 0.06	0.00	0.05	0.00	0.05
RR-eau-puit	18.00	GW	7.23	-	-	-	-

mS, medium sand; mcSG, medium to coarse sand and gravel; cSG, coarse sand and gravel; R, Bedrock; GW, groundwater. ± report the standard deviation. 0.00 is for values that were below the limit of detection (LOD).

Table S2: Abiotic characteristics measured along depth gradient at site 2.

Sample	Depth (m)	Texture	pH	TN (%)	TC (%)	TOC (%)	TIC (%)
NDDL-01	0	fmS	6.23 ± 0.03	0.03	0.48	0.43	0.05
NDDL-02	4	fmS	6.93 ± 0.01	0.01	0.17	0.16	0.01
NDDL-03	5	fmS	6.76 ± 0.02	0.01	0.18	0.18	0.00
NDDL-04	6.75	fmS	6.95 ± 0.02	0.00	0.12	0.12	0.00
NDDL-05	8.25	fvfS	7.56 ± 0.02	0.00	0.02	0.02	0.00
NDDL-06	9.75	fvfS	8.36 ± 0.03	0.00	0.03	0.03	0.00
NDDL-07	11.25	fvfS	8.96 ± 0.02	0.00	0.47	0.03	0.44
NDDL-08	14.25	fvfS	9.07 ± 0.03	0.00	0.42	0.02	0.40
NDDL-09	15.75	fvfS	8.66 ± 0.03	0.00	0.59	0.04	0.55
NDDL-10	17.25	fvfS	9.30 ± 0.01	0.00	0.43	0.01	0.42
NDDL-11	20.25	fvfS	9.35 ± 0.02	0.00	0.42	0.01	0.41
NDDL-12	21.75	fvfS	8.97 ± 0.03	0.00	0.52	0.02	0.50
NDDL-13	24	fvfS	9.10 ± 0.04	0.00	0.47	0.01	0.46
NDDL-14	24.75	fvfS	8.96 ± 0.01	0.00	0.51	0.02	0.49
NDDL-15	26.25	fvfS	8.96 ± 0.03	0.00	0.52	0.01	0.51
NDDL-16	27.75	fvfS	9.05 ± 0.01	0.00	0.52	0.01	0.51
NDDL-17	29.25	fvfS	9.00 ± 0.01	0.00	0.53	0.00	0.53
NDDL-18	30.75	SI	8.94 ± 0.02	0.00	0.57	0.02	0.55
NDDL-19	32.25	SI	8.75 ± 0.02	0.00	0.52	0.02	0.50
NDDL-20	33.75	SI	8.54 ± 0.04	0.00	0.58	0.03	0.55
NDDL-21	36.75	SIC	8.46 ± 0.01	0.00	0.67	0.09	0.58
NDDL-22	39.75	SIC	8.37 ± 0.03	0.00	1.04	0.09	0.95
NDDL-23	42.75	fSSIC	8.66 ± 0.03	0.00	1.25	0.03	1.22
NDDL-24	45.75	SG	8.65 ± 0.01	0.00	1.66	0.03	1.63
NDDL-25	47.25	SG	8.89 ± 0.05	0.00	2.93	2.27	0.65
NDDL-26	49.5	R	8.95 ± 0.01	0.00	1.80	0.46	1.34
NDDL-eau-puit	1.75	GW	7.02	-	-	-	-

fmS, fine to medium sand; fvfS, fine to very fine sand; SI, silt; SIC, Silt and clay; fSSIC, fine sand with silt and clay; SG, sand and gravel; R, Bedrock; GW, groundwater ± report the standard deviation. 0.00 is for values that were below the limit of detection (LOD).

Table S3: The correlations (r) determined by Spearman Correlation between abiotic characteristics and depth at site 1 and site 2.

	pH	TN	TC	TOC	TIC
Depth (Site 1)					
r	**0.83*** **	-	**-0.37**	**-0.79*** **	**0.58***
S	194.17	-	1564.80	2041.50	480.97
Depth (Site 2)					
r	0.16	-	**0.91*** **	0.07	**0.92*** **
S	2177.60	-	225.51	2406.10	204.62

Values in bold are significant at p-value <0.05. TN, total nitrogen; TC, total carbon; TOC, total organic carbon; TIC, total inorganic carbon. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'. Only rocks and geological material from the subsurface were used for correlation.

Sample	Quartz	K-feldspar	Plagioclase	Calcite	Dolomite	Halite	Apatite	Pyrite	Olivine	Amphibole	Pyroxene	Organic matter	Kaolinite	Chlorite	Biotite	Illite
RR-01	1.5	4.6	2.6	0.0	0.3	0.1	0.5	1.5	1.2	0.8	1.1	84.9	0.6	0.0	0.0	0.0
RR-02	45.2	22.3	32.3	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RR-03	36.0	21.3	42.6	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RR-04	46.8	19.4	31.6	0.0	0.0	0.1	0.0	0.4	0.6	0.0	0.0	0.0	0.0	0.0	0.0	1.0
RR-05	35.7	23.1	40.5	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RR-06	49.8	19.7	27.7	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
RR-07	47.2	22.1	26.8	0.0	0.0	0.3	0.9	0.8	1.0	0.0	0.0	0.0	0.0	0.1	0.0	0.8
RR-08	49.9	21.1	28.7	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RR-09	44.0	22.7	33.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RR-10	44.6	10.0	39.9	0.0	0.0	0.0	0.0	0.2	0.0	0.6	0.1	0.0	0.0	0.0	1.6	3.0
RR-11	31.4	24.2	41.9	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	2.3	0.0	0.0
RR-12	32.5	19.0	45.2	0.0	0.0	0.0	0.0	0.6	0.0	0.0	1.2	0.0	0.0	0.0	0.0	1.6
RR-13	34.4	17.6	47.4	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.3
RR-14	48.6	14.7	34.6	0.0	0.0	0.0	0.0	0.1	0.0	0.0	1.9	0.0	0.0	0.0	0.0	0.0
RR-15	34.0	23.1	37.6	0.0	0.0	0.3	0.0	0.5	1.1	2.6	0.0	0.0	0.0	0.0	0.0	0.8
RR-16	30.4	12.0	48.5	0.0	0.0	0.2	0.0	0.4	1.4	0.0	0.0	0.0	0.0	0.0	7.1	0.0
RR-17	31.3	8.1	57.9	0.0	0.0	0.0	0.0	0.3	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RR-18	24.7	27.9	46.3	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.5	0.0
RR-19	28.2	23.3	45.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.4	1.4
RR-20	30.4	25.3	44.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table S4: XRD results showing the mineralogical composition (in %) of the samples collected at site 1.

Sample	Quartz	K-feldspar	Plagioclase	Calcite	Dolomite	Halite	Apatite	Pyrite	Olivine	Amphibole	Pyroxene	Organic matter	Kaolinite	Chlorite	Biotite	Illite
NDDL-01	50.8	23.9	20.3	0.0	0.0	0.1	0.6	0.8	0.0	0.0	3.5	0.0	0.0	0.0	0.0	0.0
NDDL-02	37.9	19.0	37.2	0.0	0.0	1.0	0.0	0.0	4.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-03	32.8	17.8	33.8	0.0	0.0	0.2	0.0	1.4	2.1	1.4	3.1	0.0	1.9	0.0	0.0	5.6
NDDL-04	38.4	24.5	36.3	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-05	30.0	27.9	23.9	0.0	0.0	0.0	0.1	1.6	2.6	0.0	9.7	0.0	0.0	0.1	0.0	4.1
NDDL-06	45.1	21.0	31.4	0.0	0.0	0.0	0.0	0.4	0.0	0.0	1.1	0.0	0.0	0.0	0.0	1.0
NDDL-07	34.0	21.9	38.2	0.0	1.7	0.0	0.0	4.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-08	45.7	16.8	32.4	0.9	0.0	0.0	0.0	0.0	0.6	0.0	1.2	0.0	0.0	0.0	0.0	2.4
NDDL-09	32.1	18.1	39.1	2.5	0.0	0.0	0.0	0.0	3.3	1.5	0.0	0.8	0.4	0.3	0.6	1.3
NDDL-10	52.0	14.4	26.6	0.0	2.1	0.0	0.0	2.5	1.2	0.0	0.0	0.0	0.0	0.0	0.0	1.2
NDDL-11	27.0	30.0	40.7	0.0	0.9	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-12	40.4	18.5	32.8	3.5	0.0	0.0	0.0	0.9	0.0	0.0	3.0	0.0	0.0	1.0	0.0	0.0
NDDL-13	29.9	30.8	34.8	3.2	0.0	0.0	0.0	1.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-14	45.3	8.0	43.8	0.0	0.0	1.5	0.0	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-15	43.2	13.6	39.7	0.0	1.1	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0
NDDL-16	26.8	16.5	55.0	0.8	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-17	29.9	16.3	47.3	0.0	0.0	0.0	0.8	2.0	2.7	0.0	0.0	0.0	0.0	1.0	0.0	0.0
NDDL-18	26.0	24.4	41.8	2.7	0.0	0.0	0.0	1.0	0.0	0.0	1.8	0.0	0.0	0.0	0.0	2.2
NDDL-19	27.4	21.3	43.2	0.0	0.7	0.0	0.2	1.0	0.0	0.0	6.2	0.0	0.0	0.0	0.0	0.0
NDDL-20	22.4	22.1	46.8	1.8	0.3	0.0	0.2	0.0	1.2	0.5	4.1	0.0	0.0	0.0	0.6	0.0
NDDL-21	23.7	19.3	44.2	3.1	0.8	0.0	0.0	1.9	0.9	2.0	4.1	0.0	0.0	0.0	0.0	0.0
NDDL-22	19.2	19.5	34.6	4.2	0.8	0.0	0.4	1.9	1.5	5.6	7.6	0.0	0.4	0.0	0.5	3.7
NDDL-23	50.7	7.8	34.0	0.0	2.5	0.0	0.0	3.0	1.6	0.0	0.0	0.0	0.0	0.0	0.4	0.0
NDDL-24	35.1	25.8	26.1	10.3	2.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NDDL-25	34.0	6.0	19.4	29.4	5.2	0.0	0.1	1.1	0.0	0.0	3.2	0.0	0.0	0.0	0.0	1.6
NDDL-26	17.2	31.4	28.4	19.9	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1

Table S5: XRD results showing the mineralogical composition (in %) of the samples collected at site 2.

Table S6: Spearman correlation between bacterial α-diversity metrics and characteristics of geological material at site 1 and site 2.

	Shannon	Simpson	PD	MNTD
Site 1				
Depth				
r	**-0.54**	**-0.67***	-0.28	0.42
S	2052	2226	1696	770
pH				
r	-0.40	**-0.51**	-0.23	0.30
S	1860.40	2011.50	1631.20	934.70
TN				
r	-	-	-	-
S	-	-	-	-
TC				
r	-0.00	0.13	-0.12	-0.02
S	1331	1163.40	1490.50	1361.70
TOC				
r	0.38	**0.54**	0.14	-0.36
S	826.16	610.54	1137.50	1805.20
TIC				
r	**-0.47**	**-0.57***	-0.24	0.43
S	1961.10	2088.20	1649.80	751.80
Site 2				
Depth				
r	**-0.77*****	**-0.62***	**-0.75*****	**0.80*****
S	4596	4208	4538	514
pH				
r	-0.33	-0.26	-0.38	0.26
S	3455.30	3288.30	3575.40	1924.70
TN				
r	-	-	-	-
S	-	-	-	-
TC				
r	**-0.65****	**-0.50**	**-0.59***	**0.73*****
S	4302.60	3912.80	4122.20	689.96
TOC				
r	0.18	0.15	0.30	-0.21
S	2129	2199.90	1828.10	3150
TIC				
r	**-0.71*****	**-0.58***	**-0.67****	**0.79*****
S	4447.70	4109.70	4343.40	540.65

Values in bold are significant at p-value < 0.05. TN: total nitrogen; TC: total carbon; TOC: total organic carbon; TIC: total inorganic carbon. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'.

Table S7: Spearman Correlation between archaeal α-diversity metrics and characteristics of geological material at site 1 and site 2.

	Shannon	Simpson	PD	MNTD
Site 1				
Depth				
r	0.06	-0.01	**0.73***	0.30
S	342	368	98	254
pH				
r	0.13	0.10	0.38	0.27
S	317.94	325.95	225.81	265.87
TN				
r	-	-	-	-
S	-	-	-	-
TC				
r	-0.01	0.08	**-0.67**	-0.20
S	367.10	333.99	609.22	437.46
TOC				
r	-0.08	0.03	**-0.67**	-0.12
S	393.96	352.23	606.86	407.87
TIC				
r	0.16	-0.03	0.28	-0.19
S	304.17	376.66	262.76	431.88
Site 2				
Depth				
r	**-0.73*****	**-0.62***	**-0.73*****	0.20
S	3972	3724	3968	1832
pH				
r	-0.07	0.01	-0.37	-0.27
S	2465.20	2285	3147.90	2918.70
TN				
r	-	-	-	-
S	-	-	-	-
TC				
r	**-0.62***	**-0.55***	**-0.51**	0.38
S	3736.80	3564.30	3481.10	1423.70
TOC				
r	0.12	0.03	**0.42**	0.31
S	2017.90	2220.90	1326.90	1576.50
TIC				
r	**-0.65****	**-0.57***	**-0.55***	0.36
S	3786.20	3611.60	3570.40	1464.10

Values in bold are significant at p-value < 0.05. TN: total nitrogen; TC: total carbon; TOC: total organic carbon; TIC: total inorganic carbon. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'.

Table S8: Spearman Correlation between eukaryotic α-diversity metrics and characteristics of geological material at site 1 and site 2.

	Shannon	Simpson	PD	MNTD
Site 1				
Depth				
	-0.13	-0.09	0.03	0.11
	516	496	440	406
pH				
	0.11	0.12	0.22	-0.12
	404.94	399,94	354.89	510.06
TN				
	-	-	-	-
	-	-	-	-
TC				
	0.22	0.17	0.07	-0.18
	356.36	377.94	421.09	536.17
TOC				
	0.09	0.03	-0.08	-0.01
	415.24	442.10	490.46	461.45
TIC				
	-0.03	0.03	0.27	-0.02
	467.9	442.10	331.89	463.21
Site 2				
Depth				
	-0.2	-0.23	0.07	0.58
	144	148	112	50
pH				
	-0.23	-0.25	-0.03	0.63
	148	150	124	44
TN				
	-	-	-	-
	-	-	-	-
TC				
	-0.23	-0.43	0.08	0.12
	148	172	110	106
TOC				
	0.39	0.27	0.25	-0.66
	73.61	87.73	89.75	198.66
TIC				
	-0.32	-0.48	0.06	0.45
	158.65	177.45	112.69	65.69

Values in bold are significant at p-value < 0.05. TN: total nitrogen; TC: total carbon; TOC: total organic carbon; TIC: total inorganic carbon. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'.

Table S9: Mantel and ANOSIM test between bacteria ß-diversity metrics and characteristics of geological material.

	Depth	pH	TN	TC	TOC	TIC	Mineralogy	Texture
Bacteria								
Site 1	**0.35***	**0.66****	-	0.07	-	-	**0.26***	**0.64****
Site 2	**0.53****	**0.37****	-	**0.28****	0.04	-	0.04	**0.47****
Archaea								
Site 1	**0.25**	0.12	-	0.14	-	-	0.07	0.20
Site 2	**0.59****	**0.22***	-	**0.23****	0.05	-	0.08	0.15
Eukaryote								
Site 1	**0.25***	**0.23***	-	**0.31***	-	-0.03	0.04	0.30
Site 2	**0.27**	0.19	-	0.13	**0.37**	-	0.15	-0.2

Values in bold are significant at p-value < 0.05. TN: total nitrogen; TC: total carbon; TOC: total organic carbon; TIC: total inorganic carbon. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'.

Table S10: Spearman Correlation between absolute abundance and characteristics of geological material at site 1 and site 2.

	p-value	r	S
Site 1			
Depth	0.76	0.12	146
pH	1	0.01	164
TN	-	-	-
TC	0.70	0.14	141.57
TOC	0.76	0.11	146.31
TIC	0.09	0.57	71.13
Depth	1.20×10^{-3}	**-0.68***	1910
pH	0.46	-0.18	1347.10
TN	-	-	-
TC	0.04	**-0.47**	1676.50
TOC	0.75	0.08	1052.30
TIC	0.02	**-0.53**	1748.90

Values in bold are significant at p-value < 0.05. TN: total nitrogen; TC: total carbon; TOC: total organic carbon; TIC: total inorganic carbon. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'.

Table S11: Spearman correlation between relative abundances of most abundant bacterial, archaeal, and eukaryotic phyla and genera with characteristics of geological material at site 1 and site 2.

	Depth (site 1)	**Depth (site 2)**
Bacterial phyla		
Proteobacteria	**0.68***	0.23
Acidobacteriota	**-0.72****	**-0.73*****
Actinobacteriota	**-0.65***	0.22
Bacteroidota	**0.77****	0.44
Nitrospirota	0.11	-
Myxococcota	**-0.66***	**-0.62****
Firmicutes	0.28	0.30
Other	-0.34	**-0.56***
Unknow	-0.11	0.11
Chloroflexi	-	0.25
Sva0485	-	-0.08
Desulfobacterota	-	-0.47
Elusimicrobiota	-	0.13
Bacterial genera		
Ralstonia	**0.94*****	-
Acinetobacter	0.49	0.08
Acidothermus	**-0.80*****	-
Bryobacter	**-0.70****	-
Rhodoplanes	-0.39	-
Candidatus Kirobacter	-0.55	-
Nitrospira	0.09	-
Rheinheimera	-	0.09
Corynebacterium	-	0.46
Other	0.06	0.23
Unknow	**-0.67***	-0.45
Archaeal phyla		
Crenarchaeota	**-0.62**	**0.58***
Woesearchaeota	**0.70***	-0.11
Thermoplasmatota	-0.19	-
Other	**0.61***	-0.19
Unclassified Archaea	-0.20	-
Halobacterota	-	**-0.61***
Bathyarchaeota	-	**-0.41**
Archaeal genera		
Unclassified Group 1.1c	**-0.84****	**-0.75*****
Unclassified Nitrososphaeria	0.08	-
Unclassified Nitrosopumilaceae	0.48	-
Unclassified Woese 24	**0.57**	-
Unclassified Crenarchaeota	**0.58**	**-0.70****

Unclassified Woese 7	0.49	-
Unclassified Bathyarchaeia	0.52	**0.84*** **
Other	0.18	**-0.5**
Unclassified Archaea	-0.20	-
Unclassified Thermoplasmata	0.09	-
Methanoregula	-	**-0.41**
Unclassified Rice Cluster II	-	**-0.54***
Unclassified Bathyarchaeota	-	**-0.49**
Eukaryotic phyla		
Basidiomycota	**-0.82****	0.07
Ascomycota	0.10	0.5
Phragmoplastophyta	-0.41	-0.33
Chytridiomycota	0.48	
Cercozoa	0.31	0.48
Vertebrata	**0.87*** **	-
Dinoflagellata	**0.86*** **	-
Arthropoda	0.14	0.55
Fungi	0.20	0.12
Nematozoa	**-0.76***	0.08
Annelida	-0.44	-
Ciliophora	0.48	-
Mucoromycota	**0.63**	0.27
Ochrophyta	**0.55**	-
Metazoa	-	0.37
Cryptomycota	-	-0.40
Mollusca	-	0.55
Other	**0.77***	0.23
Unknowm	**0.76***	-0.15
Eukaryotic genera		
Geranomyces	0.34	-
Archaeorhizomyces	**-0.64**	-
Oikomonas	NA	-
Venturia	-	0.22
Archaeorhizomyces	-	0.17
Unknown	-0.45	-0.68
Other	0.35	0.65

Values in bold are significant at p-value < 0.05. Signification codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'

ANNEXE C : SCRIPTS

1. Sequence processing (DADA2 Pipeline v.1.18.0)

Here we walk through version 1.18.0 of the DADA2 pipeline described by Callahan et al. (2016). Our starting point is a set of Illumina-sequenced paired-end fastq files that have been split by sample and from which the barcodes/adapters have already been removed. The product is an amplicon sequence variant (ASV) table, a higher-resolution analogue of the traditional OTU table, which records the number of times each exact amplicon sequence variant was observed in each sample.

The script showed are the one that were used for the sequences detected at site 1. The codes are the same for site 2. Just make sure that you set the right working directory and that you use the right control sample names (e.g., "Lazar-CTRL-neg-PCR-bac", "Kit-soil-bac", "NDDL-eau-forage-bac" for bacteria at site 2). Also make sure that you register both taxonomy table and ASV table with a correct name.

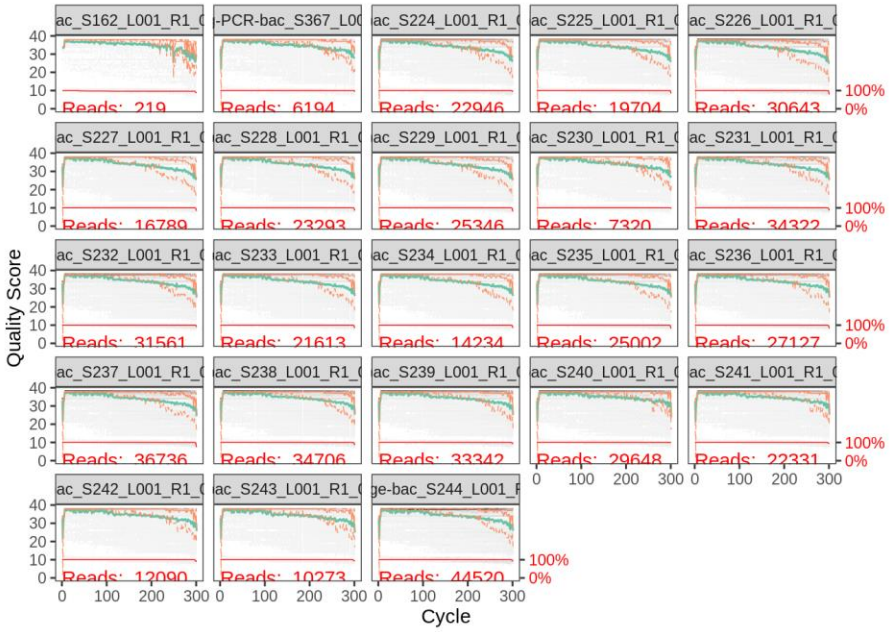
1.1 Bacterial sequences processing

1.1.1 Getting ready

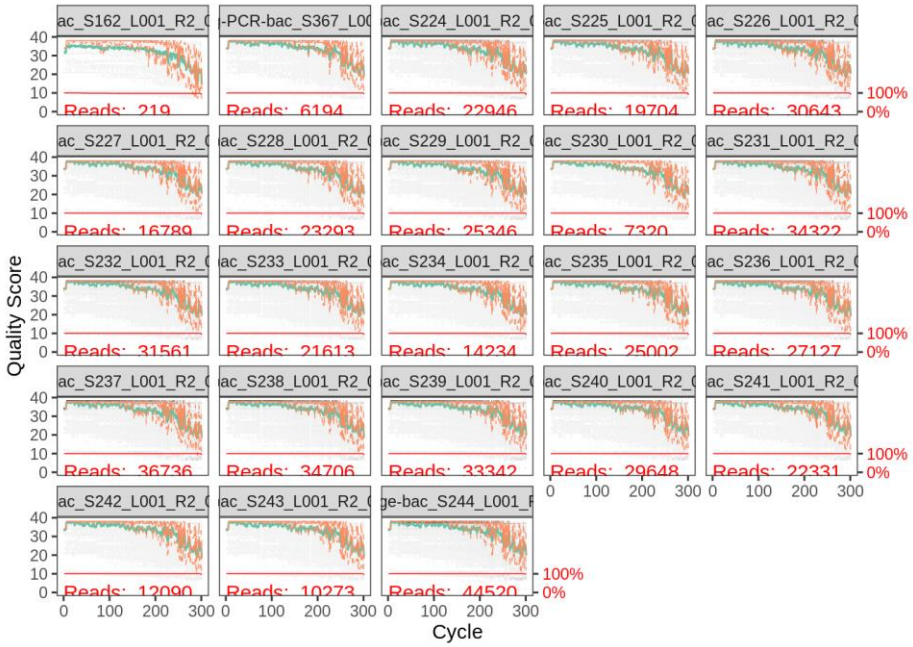
```
library(dada2); packageVersion("dada2")
setwd("~/ASV/CHAP1/BAC/site_1") # Set your working directory
path <- "~/ASV/CHAP1/BAC/site_1" # The directory containing the fastq files after unzipping.
list.files(path)
# Forward and reverse fastq filenames have format: SAMPLENAME_R1_001.fastq and SAMPLENAME_R2
_001.fastq
fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))
# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
```

1.1.2 Inspect read quality profiles

```
plotQualityProfile(fnFs[1:23]) # Visualize the quality profile of the forward reads
```




```
plotQualityProfile(fnRs[1:23]) # Visualize the quality profile of the reverse reads
```



The forward reads are relatively good quality. Based on these profiles, we decided to trimm the last few nucleotides at position 290 (trimming the last 10 nucleotides) to avoid less well-controlled errors that can arise there. We truncated the reverse reads at position 250 where the quality distribution crashes.

1.1.3 Filter and trim

```
# Place filtered files in filtered/ subdirectory
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(290,250),
```

```
        maxN=0, maxEE=c(2,2), truncQ=2, rm.phix=TRUE,
        compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```

1.1.4 Learn the Error Rates

```
errF <- learnErrors(filtFs, multithread=TRUE)
## 70527130 total bases in 243197 reads from 23 samples will be used for learning the error rates
errR <- learnErrors(filtRs, multithread=TRUE)
## 60799250 total bases in 243197 reads from 23 samples will be used for learning the error rates
plotErrors(errF, nominalQ=TRUE)
```

1.1.5 Dereplication

```
derepFs <- derepFastq(filtFs, verbose=TRUE)
derepRs <- derepFastq(filtRs, verbose=TRUE)
# Name the derep-class objects by the sample names
names(derepFs) <- sample.names
names(derepRs) <- sample.names
```

1.1.6 Sample Inference

```
dadaFs <- dada(derepFs, err=errF, multithread=TRUE)
dadaRs <- dada(derepRs, err=errR, multithread=TRUE)
dadaFs[[7]]
```

1.1.7 Merge paired reads

```
mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs, verbose=TRUE)
# Inspect the merger data.frame from the first sample
head(mergers[[1]])
```

1.1.8 Construct sequence table

```
seqtab <- makeSequenceTable(mergers)
dim(seqtab)
## [1]   23 4471
# Inspect distribution of sequence lengths
table(nchar(getSequences(seqtab)))
```

1.1.9 Remove chimeras

```
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRU
E)
## Identified 1101 bimeras out of 4471 input sequences.
dim(seqtab.nochim)
## [1]   23 3370
```

```
sum(seqtab.nochim)/sum(seqtab)
## [1] 0.9610272
```

1.1.10 Track reads through the Pipeline

```
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.nochim))
colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
rownames(track) <- sample.names
```

1.1.11 Assign taxonomy

```
install.packages("DECIPHER")
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("DECIPHER")
library(DECIPHER)
dna <- Biostrings::DNAStringSet(getSequences(seqtab.nochim)) # Create a DNAStringSet from the ASVs
load("~/SILVA_SSU_r138_2019.RData")
ids <- IdTaxa(dna, trainingSet, strand="both", processors=NULL, verbose=FALSE, threshold = 50)
ranks <- c("domain", "phylum", "class", "order", "family", "genus", "species")
taxid <- t(sapply(ids, function(x) {
      m <- match(ranks, x$rank)
      taxa <- x$taxon[m]
      taxa[startsWith(taxa, "unclassified_")] <- NA
      taxa
}))
colnames(taxid) <- ranks
rownames(taxid) <- getSequences(seqtab.nochim)
```

1.1.12 Get rid of potential contamination

```
ctrl.samples <- c("Lazar-CTRL-neg-PCR-bac", "Kit-soil-bac","RR-eau-forage-bac") # CHANGE to names of your negative controls
found.in.ctrls <- colSums(seqtab[ctrl.samples,])>0
seqtab.no.ctrl.seqs <- seqtab[,!found.in.ctrls]
```

1.1.13 Register

```
write.csv(as.data.frame(taxid), file = file.path(path,"BAC_S1_taxonomy_table.csv")) # Taxonomy table
write.csv(as.data.frame(seqtab.no.ctrl.seqs),file = file.path(path,"ASV_S1_table.csv")) # ASV table
write.csv(as.data.frame(t(seqtab.no.ctrl.seqs)),file = file.path(path,"ASV_S1_t_table.csv")) # Transpose ASV table
```

1.1.14 Decipher alignment

```
library(DECIPHER)
seqs <- getSequences(seqtab.no.ctrl.seqs)
names(seqs) <- seqs # This propagates to the tip labels of the tree
alignment <- AlignSeqs(DNAStringSet(seqs), anchor=NA, processors = NULL)
#Enregistrer les séquences allignées en format FASTA
writeXStringSet(alignment, file = file.path(path,"ASV_S1_align.fasta"), format="fasta")
# Now we will use this command on FastTree to generate a phylogenetic tree:
# -gtr -nt < ASV/CHAP1/BAC/site_1/ASV_S1_align.fasta > ASV/CHAP1/BAC/site_1/tree
```

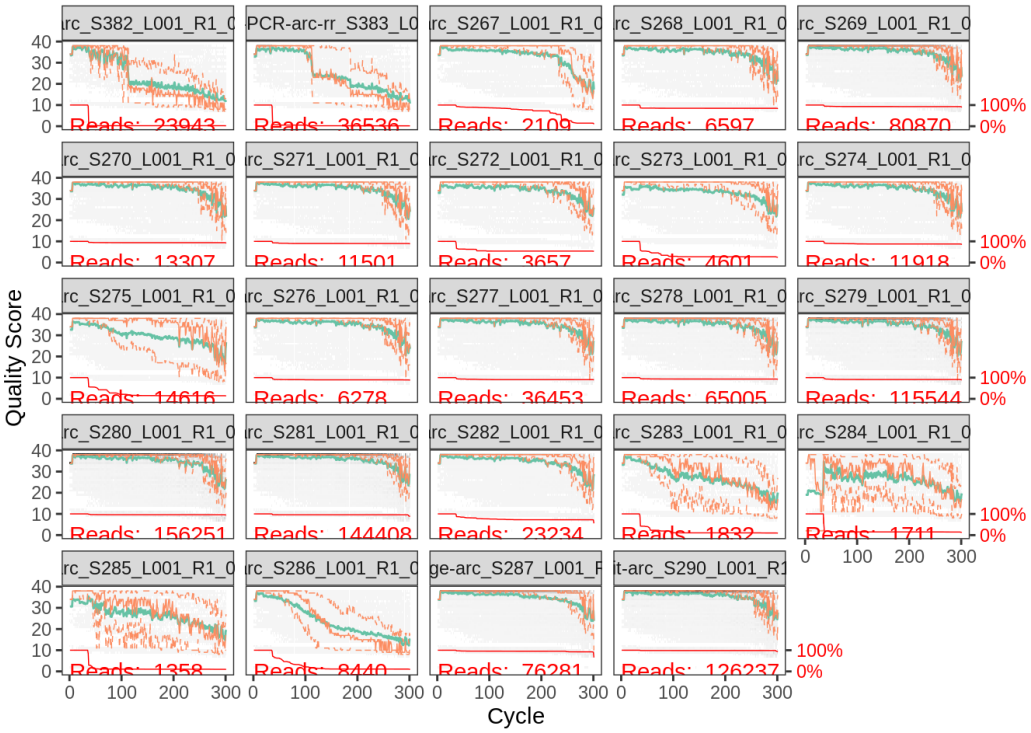
1.2 Archaeal sequences processing

1.2.1 Getting ready

```
library(dada2); packageVersion("dada2")
setwd("~/ASV/CHAP1/ARC/Site_1") # Set your working directory
path <- "~/ASV/CHAP1/ARC/Site_1" # The directory containing the fastq files after unzipping.
list.files(path)
# Forward fastq filename have format: SAMPLENAME_R1_001.fastq
fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
```

1.2.2 Inspect read quality profiles

```
plotQualityProfile(fnFs[1:24]) # Visualize the quality profile of the forward reads
```



Based on these profiles, we decided to truncate the forward reads at position 210 where the quality distribution generaly crashes.

1.2.3 Filter and trim

```
# Place filtered files in filtered/ subdirectory
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
out <- filterAndTrim(fnFs, filtFs, truncLen=c(210), maxN=0, maxEE=2 , truncQ=2, trimLeft=10, compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```

1.2.4 Learn the Error Rates

```
errF <- learnErrors(filtFs, multithread=TRUE)
plotErrors(errF, nominalQ=TRUE)
```

1.2.5 Dereplication

```
derepFs <- derepFastq(filtFs, verbose=TRUE)
# Name the derep-class objects by the sample names
names(derepFs) <- sample.names
```

1.2.6 Sample Inference

```
dadaFs <- dada(derepFs, err=errF, multithread=TRUE)
dadaFs[[7]]
```

1.2.7 Construct sequence table

```
seqtab <- makeSequenceTable(dadaFs)
dim(seqtab)
## [1]   24 1870
```

1.2.8 Remove chimeras

```
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRU
E)
## Identified 1028 bimeras out of 1870 input sequences.
dim(seqtab.nochim)
## [1]  24 842
sum(seqtab.nochim)/sum(seqtab)
## [1] 0.8596018
```

1.2.9 Track reads through the Pipeline

```
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), rowSums(seqtab.nochim))
colnames(track) <- c("input", "filtered", "denoisedF", "nonchim")
rownames(track) <- sample.names
```

1.2.10 Assign taxonomy

```
library(DECIPHER)
dna <- Biostrings::DNAStringSet(getSequences(seqtab.nochim)) # Create a DNAStringSet from the ASVs
load("~/SILVA_SSU_r138_2019.RData") # Change to the path of your training set
ids <- IdTaxa(dna, trainingSet, strand="both", processors=NULL, verbose=FALSE, threshold = 50)
ranks <- c("domain", "phylum", "class", "order", "family", "genus") # Ranks of interest
# Convert the output object of class "Taxa" to a matrix analogous to the output from assignTaxonomy
```

```
taxid <- t(sapply(ids, function(x) {
      m <- match(ranks, x$rank)
      taxa <- x$taxon[m]
      taxa[startsWith(taxa, "unclassified_")] <- NA
      taxa }))
colnames(taxid) <- ranks; rownames(taxid) <- getSequences(seqtab.nochim)
taxid <- as.data.frame(taxid)
taxint <- subset(taxid, is.na(phylum))
taxide <- subset(taxid, !(is.na(domain)))
dim(taxint)
## [1] 424   6
seqtabint <-as.data.frame(seqtab.nochim)
seqtabint <- seqtab.nochim[,colnames(seqtab.nochim) %in% rownames(taxint)]
load("~/ASV/CHAP1/ARC/Site_1/arc.cassandre.trainingset.RData") # Change to the path of your training set
dna <- DNAStringSet(getSequences(seqtabint)) # Create a DNAStringSet from the ASVs
ids <- IdTaxa(dna, trainingSet, strand="both", processors=NULL, verbose=FALSE, threshold = 50)
taxint <- t(sapply(ids, function(x) {
      m <- match(ranks, x$rank)
      taxa <- x$taxon[m]
      taxa[startsWith(taxa, "unclassified_")] <- NA
      taxa}))
colnames(taxint) <- ranks; rownames(taxint) <- getSequences(seqtabint)
taxint <-subset(as.data.frame(taxint), domain =="Archaea")
taxide <- taxide[!(rownames(taxide) %in% rownames(taxint)),]
taxid <- rbind(taxide, as.data.frame(taxint))

taxid <- as.data.frame(t(taxid))
taxid[] <- lapply(taxid, as.character)
taxid2<- tidyr::fill(taxid, names(taxid),.direction = "down")
taxid2<- sapply(taxid2, function(x){paste0("unclassified_", x)})
taxid[is.na(taxid)] <- taxid2[is.na(taxid)]
taxid <- t(taxid)
taxid[ taxid == "unclassified_NA" ] <- NA
```

1.2.11 Get rid of potential contamination

```
ctrl.samples <- c("Lazar-CTRL-neg-PCR-arc-rr", "Kit-soil-arc","RR-eau-forage-arc") # Change to names of y our negative controls
found.in.ctrls <- colSums(seqtab[ctrl.samples,])>0
seqtab.no.ctrl.seqs <- seqtab[,!found.in.ctrls]
```

1.2.12 Register

```
write.csv(as.data.frame(taxid), file = file.path(path,"ARC_S1_taxonomy_table.csv")) # Taxonomy table
write.csv(as.data.frame(seqtab.no.ctrl.seqs),file = file.path(path,"ASV_ARC_S1_table.csv")) # ASV table
```

```
write.csv(as.data.frame(t(seqtab.no.ctrl.seqs)),file = file.path(path,"ASV_ARC_S1_t_table.csv"))  # Transp
ose ASV table
```

1.2.13 Decipher alignment

```
library(DECIPHER)
seqs <- getSequences(seqtab.no.ctrl.seqs)
names(seqs) <- seqs # This propagates to the tip labels of the tree
alignment <- AlignSeqs(DNAStringSet(seqs), anchor=NA, processors = NULL)
#Enregistrer les séquences allignées en format FASTA
writeXStringSet(alignment, file = file.path(path,"ASV_ARC_S1_align.fasta"),format="fasta")
# Now we will use this command on FastTree to generate a phylogenetic tree :
# fasttree -gtr -nt < ASV/CHAP1/ARC/Site_1/ASV_ARC_S1_align.fasta > ASV/CHAP1/ARC/Site_1/tree.arc.
S1
```

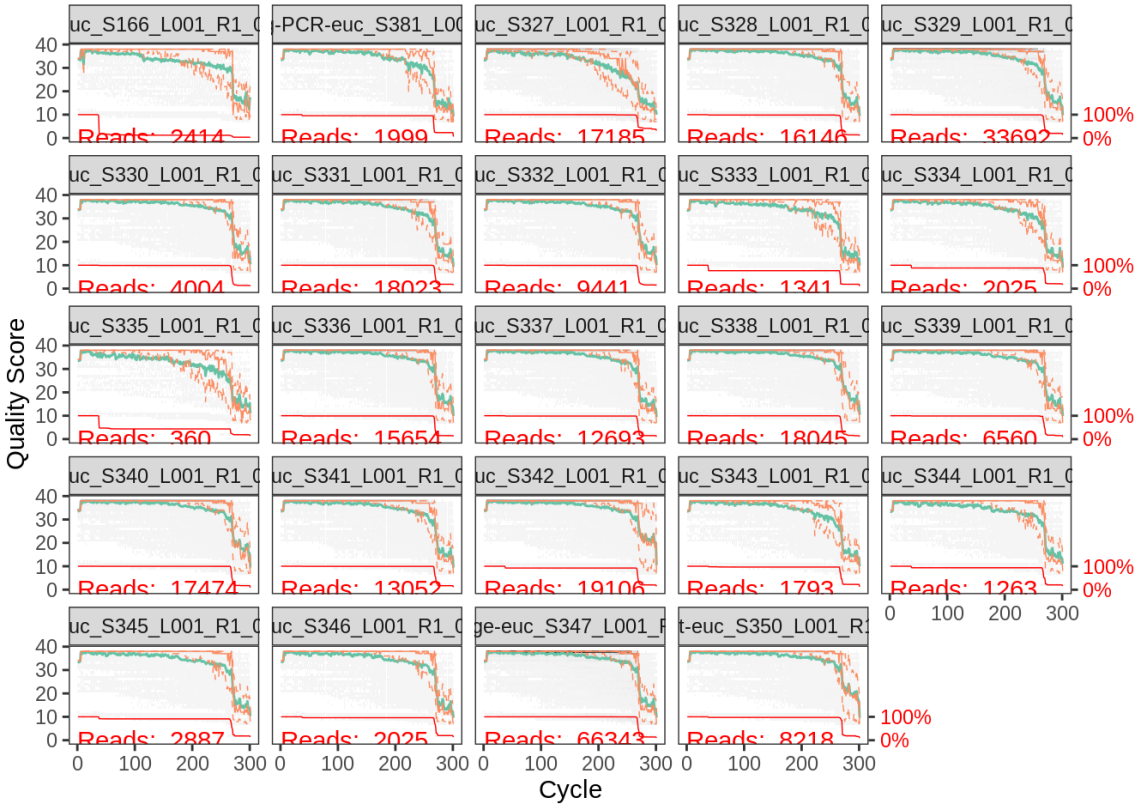
1.3 Eukaryal sequences processing

1.3.1 Getting ready

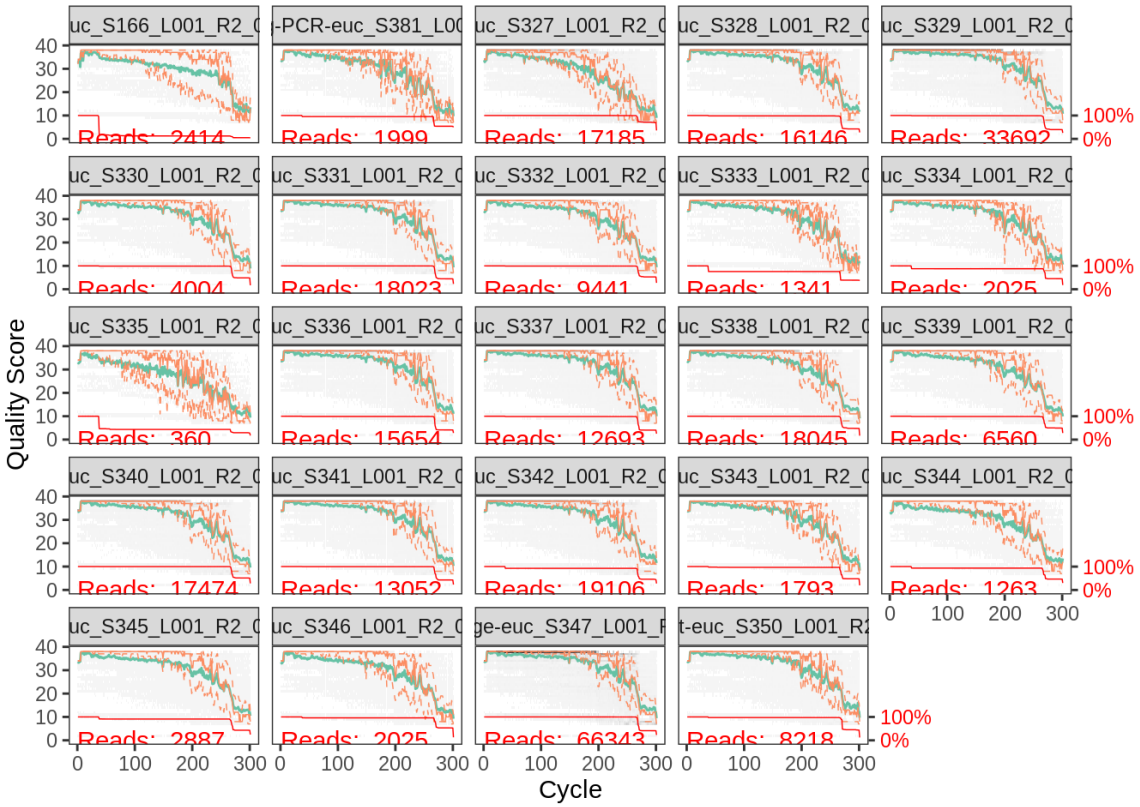
```
library(dada2); packageVersion("dada2")
setwd("~/ASV/CHAP1/EUC/Site_1") # Set your working directory
path <- "~/ASV/CHAP1/EUC/Site_1" # The directory containing the fastq files after unzipping.
list.files(path)
# Forward and reverse fastq filenames have format: SAMPLENAME_R1_001.fastq and SAMPLENAME_R2
_001.fastq
fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))
# Extract sample names, assuming filenames have format: SAMPLENAME_XXX.fastq
sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
```

1.3.2 Inspect read quality profiles

```
PlotQualityProfile(fnFs[1:24]) # Visualize the quality profile of the forward reads
```




```
plotQualityProfile(fnRs[1:24]) #visualize the quality profile of the reverse reads
```



Based on these profiles, we decided to truncate both forward and reverse reads at position 250.

1.3.3 Filter and trim

```
# Place filtered files in filtered/ subdirectory
filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(250,250), maxN=0, maxEE=c(2,2), truncQ=2, rm.p
hix=TRUE, compress=TRUE, multithread=TRUE) # On Windows set multithread=FALSE
```

1.3.4 Learn the Error Rates

```
errF <- learnErrors(filtFs, multithread=TRUE)
errR <- learnErrors(filtRs, multithread=TRUE)
plotErrors(errF, nominalQ=TRUE)
```

1.3.5 Dereplication

```
derepFs <- derepFastq(filtFs, verbose=TRUE)
derepRs <- derepFastq(filtRs, verbose=TRUE)
names(derepFs) <- sample.names # Name the derep-class objects by the sample names
names(derepRs) <- sample.names
```

1.3.6 Sample Inference

```
dadaFs <- dada(derepFs, err=errF, multithread=TRUE)
dadaRs <- dada(derepRs, err=errR, multithread=TRUE)
dadaFs[[7]]
```

1.3.7 Merge paired reads

```
mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs, verbose=TRUE)
head(mergers[[1]]) # Inspect the merger data.frame from the first sample
```

1.3.8 Construct sequence table

```
seqtab <- makeSequenceTable(mergers)
dim(seqtab)
## [1]   24 1339
table(nchar(getSequences(seqtab))) # Inspect distribution of sequence lengths
```

1.3.9 Remove chimeras

```
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRU
E)
## Identified 359 bimeras out of 1339 input sequences.
dim(seqtab.nochim)
## [1]  24 980
sum(seqtab.nochim)/sum(seqtab)
## [1] 0.7698131
```

1.3.10 Track reads through the pipeline

```
getN <- function(x) sum(getUniques(x))
track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.n
ochim))
# If processing a single sample, remove the sapply calls: e.g. replace sapply(dadaFs, getN) with getN(dada
Fs)
colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
rownames(track) <- sample.names
```

1.3.11 Assign taxonomy

```
library(DECIPHER)
dna <- Biostrings::DNAStringSet(getSequences(seqtab.nochim)) # Create a DNAStringSet from the ASVs
load("~/SILVA_SSU_r138_2019.RData")
ids <- IdTaxa(dna, trainingSet, strand="both", processors=NULL, verbose=FALSE, threshold = 50)
ranks <- c("domain", "phylum", "class", "order", "family", "genus", "species")
taxid <- t(sapply(ids, function(x) {
      m <- match(ranks, x$rank)
```

```
    taxa <- x$taxon[m]
    taxa[startsWith(taxa, "unclassified_")] <- NA
    taxa}))
colnames(taxid) <- ranks
rownames(taxid) <- getSequences(seqtab.nochim)
```

Keep the non classified sequences and assign taxonomy again with another dataset.

```
taxid <- as.data.frame(taxid)
taxint <- subset(taxid, is.na(phylum) & is.na(class))
taxide <- subset(taxid, !(is.na(phylum) & is.na(class)))
dim(taxint)
## [1] 264   7
seqtabint <-as.data.frame(seqtab.nochim)
seqtabint <- seqtab.nochim[,colnames(seqtab.nochim) %in% rownames(taxint)]
load("~/ASV/CHAP1/EUC/Site_1/pr2_trainingset.1000.RData.gz") # CHANGE TO THE PATH OF YOUR TRAI
NING SET
dna <- DNAStringSet(getSequences(seqtabint)) # Create a DNAStringSet from the ASVs
ids <- IdTaxa(dna, trainingSet, strand="both", processors=NULL, verbose=FALSE, threshold = 50)
taxint <- t(sapply(ids, function(x) {
    m <- match(ranks, x$rank)
    taxa <- x$taxon[m]
    taxa[startsWith(taxa, "unclassified_")] <- NA
    taxa }))
colnames(taxint) <- ranks; rownames(taxint) <- getSequences(seqtabint)
taxid <- rbind(taxide, as.data.frame(taxint))
```

1.3.12 Get rid of potential contamination

```
ctrl.samples <- c("Lazar-CTRL-neg-PCR-euc", "Kit-soil-euc","RR-eau-forage-euc") # CHANGE to names of y
our negative controls
found.in.ctrls <- colSums(seqtab[ctrl.samples,])>0
seqtab.no.ctrl.seqs <- seqtab[,!found.in.ctrls]
```

1.3.13 Register

```
write.csv(as.data.frame(taxid), file = file.path(path,"EUC_S1_taxonomy_table.csv")) # Taxonomy table
write.csv(as.data.frame(seqtab.no.ctrl.seqs),file = file.path(path,"ASV_EUC_S1_table.csv")) # ASV table
write.csv(as.data.frame(t(seqtab.no.ctrl.seqs)),file = file.path(path,"ASV_EUC_S1_t_table.csv")) # Transp
ose ASV table
```

1.3.14 Decipher alignment

```
library(DECIPHER)
seqs <- getSequences(seqtab.no.ctrl.seqs)
names(seqs) <- seqs # This propagates to the tip labels of the tree
```

```
alignment <- AlignSeqs(DNAStringSet(seqs), anchor=NA, processors = NULL)
# Register in the FASTQ format
writeXStringSet(alignment, file = file.path(path,"ASV_EUC_S1_align.fasta"),format="fasta")

# Now we will use this command on FastTree to generate a phylogenetic tree:
# fasttree -gtr -nt < ASV/CHAP1/EUC/Site_1/ASV_EUC_S1_align.fasta > ASV/CHAP1/EUC/Site_1/tree.euc.
S1
```

2. Sediment characteristics

2.1 Load library

```
library("tidyverse")
library("rstatix")
library("gridExtra")
library("cowplot")
```

2.2 Set directory

```
setwd("~/ASV/CHAP1/Soil_Characteristics")
```

2.3 Load data

```
Sediment_Characteristics <- read.csv(file="Sediment_Characteristics.csv", header=TRUE, sep=";", dec=".")
Sediment_Characteristics[1:5, 1:5]
```

```
##   Sample   Site       Category Depth Texture
## 1  RR.01 Site 1 Site 1 - Sessile  0.00      mS
## 2  RR.02 Site 1 Site 1 - Sessile  0.75      mS
## 3  RR.03 Site 1 Site 1 - Sessile  2.25      mS
## 4  RR.04 Site 1 Site 1 - Sessile  3.75    mcSG
## 5  RR.05 Site 1 Site 1 - Sessile  5.25    mcSG
```

2.4 Prepare the data

```
# Delete surface sample at site 1 for better visual representation
Sediment_Characteristics_wss <- Sediment_Characteristics[Sediment_Characteristics$Sample !="RR.01",]
Sediment_Characteristics_wss[1:5, 1:5]
```

```
##   Sample   Site       Category Depth Texture
## 2  RR.02 Site 1 Site 1 - Sessile  0.75      mS
## 3  RR.03 Site 1 Site 1 - Sessile  2.25      mS
## 4  RR.04 Site 1 Site 1 - Sessile  3.75    mcSG
## 5  RR.05 Site 1 Site 1 - Sessile  5.25    mcSG
## 6  RR.06 Site 1 Site 1 - Sessile  6.75    mcSG
```

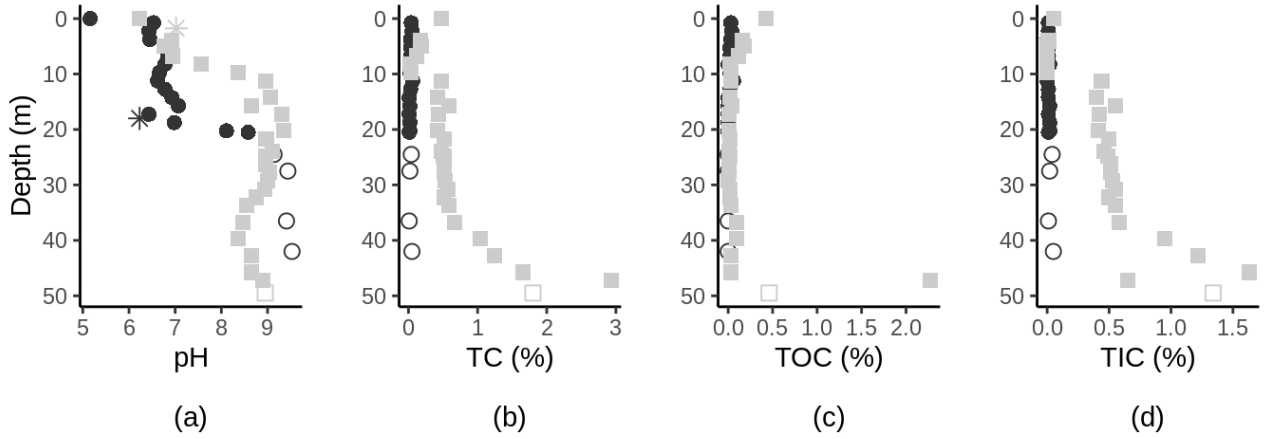
2.5 Plot Sediment Characteristics

```
pH <- ggplot(Sediment_Characteristics, aes(x = Depth, y = pH, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+ xlab("Depth (m)") + ylab("pH\n\n(a)") + theme_classic() + scale_shape_manual(values=c(1, 8,16,0,8,15))+ theme(legend.position="none")

TC <- ggplot(Sediment_Characteristics_wss, aes(x = Depth, y = TC, colour = Site,shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse()+ theme_bw() + xlab("") + ylab("TC
```

```
 (%)\n\n(b)") + theme_classic()+ scale_shape_manual(values=c(1, 8,16,0,8,15)) + theme(legend.position=
"none")

TOC <- ggplot(Sediment_Characteristics_wss, aes(x = Depth, y = TOC, colour = Site, shape= Category))+ ge
om_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+ xlab("") + ylab
("TOC (%)\n\n(c)") + theme_classic()+ scale_shape_manual(values=c(1, 8,16,0,8,15))+ theme(legend.posi
tion="none")

TIC <- ggplot(Sediment_Characteristics_wss, aes(x = Depth, y = TIC, colour = Site, shape= Category))+ geo
m_point(size=2.5) + coord_flip() + scale_x_reverse()+ theme_bw() + xlab("") + ylab("TIC (%)\n\n(d)") + sc
ale_color_grey() + theme_classic()+ scale_shape_manual(values=c(1, 8,16,0,8,15)) + theme(legend.positi
on="none")
```

2.6 Combine plots

```
plot_grid(pH,TC,TOC,TIC, ncol = 4, nrow = 2)
```



2.7 Spearman correlation

We used Spearman correlations to determine correlations between depth and sediment characteristics.

2.7.1 Define data from site 1 and 2

```
# SITE 1
Sediment_Characteristics_S1 <- Sediment_Characteristics[Sediment_Characteristics$Site=='Site 1',]
Sediment_Characteristics_S1 <- Sediment_Characteristics_S1[Sediment_Characteristics_S1$Texture!='G
W',] # Samples from site 1 and without the groundwater sample for statistical tests
Sediment_Characteristics_S1_wss <- Sediment_Characteristics_S1[Sediment_Characteristics_S1$Sample!
='RR.01',]  # Subsurface samples from first site (excluding the surface sample RR.01)

Sediment_Characteristics_S2 <- Sediment_Characteristics[Sediment_Characteristics$Site=='Site 2',]
Sediment_Characteristics_S2 <- Sediment_Characteristics_S2[Sediment_Characteristics_S2$Texture!='G
W',] # Samples from site 1 and without the groundwater sample for statistical tests
```

```
Sediment_Characteristics_S2_wss <- Sediment_Characteristics_S2[Sediment_Characteristics_S2$Sample!='NDDL.01',]  # Subsurface samples from first site (excluding the surface sample RR.01)
```

2.7.2 Test significance

```
# pH/TC/TOC/TIC as a function of depth (SITE 1)
cor.test(Sediment_Characteristics_S1_wss$Depth, Sediment_Characteristics_S1_wss$pH, method = "spearman") #pH
cor.test(Sediment_Characteristics_S1_wss$Depth, Sediment_Characteristics_S1_wss$TC, method = "spearman") #TC
cor.test(Sediment_Characteristics_S1_wss$Depth, Sediment_Characteristics_S1_wss$TOC, method = "spearman") #TOC
cor.test(Sediment_Characteristics_S1_wss$Depth, Sediment_Characteristics_S1_wss$TIC, method = "spearman") #TIC with only subsurface samples

# pH/TC/TOC/TIC as a function of depth (SITE 2)
cor.test(Sediment_Characteristics_S2_wss$Depth, Sediment_Characteristics_S2_wss$pH, method = "spearman") #pH
cor.test(Sediment_Characteristics_S2_wss$Depth, Sediment_Characteristics_S2_wss$TC, method = "spearman") #TC
cor.test(Sediment_Characteristics_S2_wss$Depth, Sediment_Characteristics_S2_wss$TOC, method = "spearman") #TOC
cor.test(Sediment_Characteristics_S2_wss$Depth, Sediment_Characteristics_S2_wss$TIC, method = "spearman") #TIC
```

3. Mineral composition

3.1 Set directory

```
setwd("~/ASV/CHAP1/Mineral_Composition")
```

3.2 Load library

```
library("RColorBrewer")
library("ggplot2")
library("cowplot")
```

3.3 Load data

```
# Load the data containing the mineral composition
Mineral_Composition <- read.csv(file="Mineral_Composition.csv", header=TRUE, sep=";", dec=".")
Mineral_Composition[1:5,1:5] # take a peek at the data (just the first five rows/columns)
##   Sample Depth   Site Quartz K.feldspar
## 1  RR.01  0.00 Site 1    1.5       4.6
## 2  RR.02  0.75 Site 1   45.2      22.3
## 3  RR.03  2.25 Site 1   36.0      21.3
## 4  RR.04  3.75 Site 1   46.8      19.4
## 5  RR.05  5.25 Site 1   35.7      23.1
```

3.4 Plot mineral composition (SITE 1)

```
Mineral_Composition_S1 <- Mineral_Composition[Mineral_Composition$Site=='Site 1',]# Samples from s
ite 1
Mineral_Composition_S1[1:5,1:5] # Take a peek at the data (just the first five rows/columns)
##   Sample Depth   Site Quartz K.feldspar
## 1  RR.01  0.00 Site 1    1.5       4.6
## 2  RR.02  0.75 Site 1   45.2      22.3
## 3  RR.03  2.25 Site 1   36.0      21.3
## 4  RR.04  3.75 Site 1   46.8      19.4
## 5  RR.05  5.25 Site 1   35.7      23.1
```

3.4.1 Prepare data

```
# First define the column "Depth" as character
Mineral_Composition_S1 ,"Depth"] <-as.character(Mineral_Composition_S1 ,"Depth"])
# Then get rid of the column "Sample"(column #1) and "Site" (column #3)
Mineral_Composition_S1 <- Mineral_Composition_S1[,-c(1,3)]
Mineral_Composition_S1[1:5,1:5]
##   Depth Quartz K.feldspar Plagioclase Calcite
## 1     0    1.5       4.6       2.6       0
## 2  0.75   45.2      22.3      32.3       0
## 3  2.25   36.0      21.3      42.6       0
## 4  3.75   46.8      19.4      31.6       0
```

```
## 5  5.25   35.7      23.1      40.5      0
# Then reshape the data frame
Mineral_Composition_S1_melt <- reshape2::melt(Mineral_Composition_S1)
## Using Depth as id variables
head(Mineral_Composition_S1_melt)
##   Depth variable value
## 1     0   Quartz   1.5
## 2  0.75   Quartz  45.2
## 3  2.25   Quartz  36.0
## 4  3.75   Quartz  46.8
## 5  5.25   Quartz  35.7
## 6  6.75   Quartz  49.8
```

3.4.2 Plot data

```
# Define the number of color that we want
mycolors = c(brewer.pal(name="Paired", n = 12), brewer.pal(name="Set3", n = 4))
# Now use the melted data frame to plot
Mineral_Composition_S1_Plot <-ggplot(Mineral_Composition_S1_melt, aes(x=value, y=factor(Depth, lev
el = c('42','36.5','27.5','24.5','20.5','20.25','18.75','17.25','15.75','14.25','12.75','11.25','9.75','8.25','6.75','
5.25','3.75','2.25','0.75','0')), fill= variable)) + geom_bar(position="fill", stat="identity") + xlab("Relative pe
rcentage\n\n Site 1\n\n(a)") + ylab("Depth (m)") + theme_classic()+ labs(fill = "Mineral composition") + s
cale_fill_manual(values = mycolors)
```

3.5 Plot mineral composition (SITE 2)

```
Mineral_Composition_S2 <- Mineral_Composition[Mineral_Composition$Site=='Site 2',] # Samples from
site 2
Mineral_Composition_S2[1:5,1:5]
##     Sample Depth   Site Quartz K.feldspar
## 21 NDDL.01  0.00 Site 2   50.8      23.9
## 22 NDDL.02  4.00 Site 2   37.9      19.0
## 23 NDDL.03  5.00 Site 2   32.8      17.8
## 24 NDDL.04  6.75 Site 2   38.4      24.5
## 25 NDDL.05  8.25 Site 2   30.0      27.9
```

3.5.1 Prepare data

```
# First define the column "Depth" as character
Mineral_Composition_S2 ["Depth"] <-as.character(Mineral_Composition_S2 ["Depth"])
# Then get rid of the column "Sample"(column #1) and "Site" (column #3)
Mineral_Composition_S2 <- Mineral_Composition_S2[,-c(1,3)]
Mineral_Composition_S2[1:5,1:5]
##    Depth Quartz K.feldspar Plagioclase Calcite
## 21     0   50.8      23.9       20.3       0
## 22     4   37.9      19.0       37.2       0
```

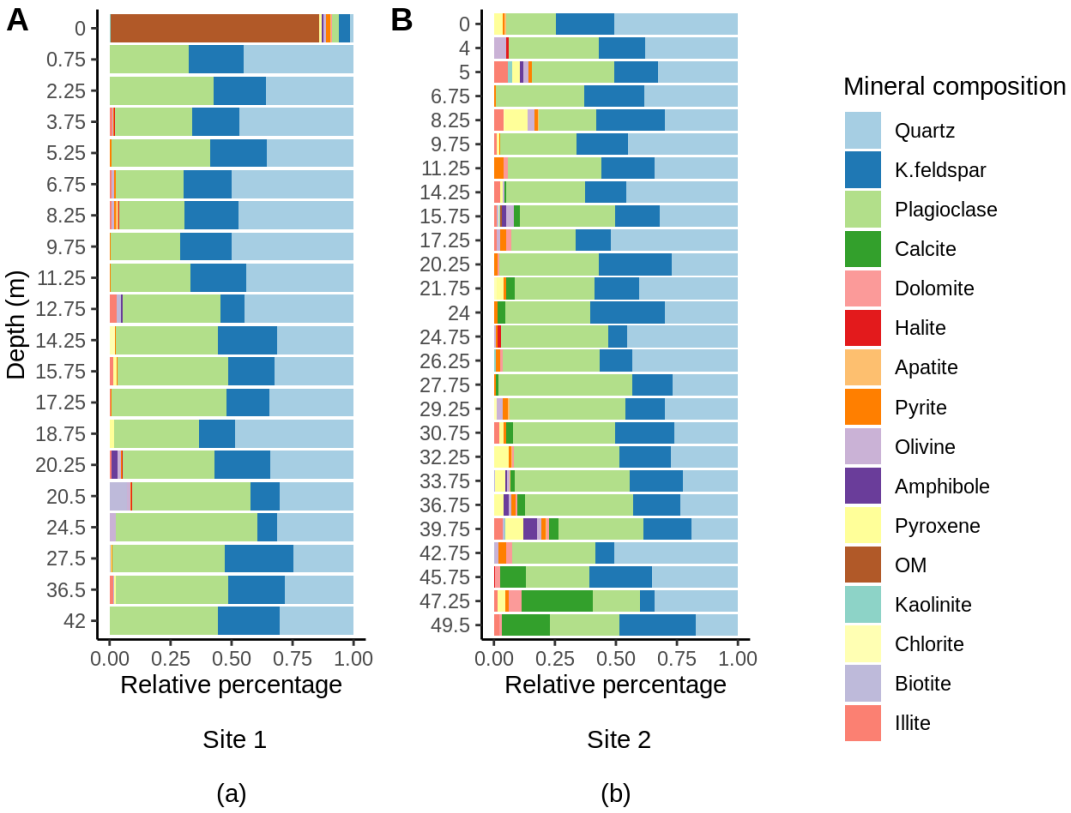
```
## 23    5   32.8      17.8      33.8      0
## 24  6.75   38.4      24.5      36.3      0
## 25  8.25   30.0      27.9      23.9      0
# Then reshape the data frame
Mineral_Composition_S2_melt <- reshape2::melt(Mineral_Composition_S2)
## Using Depth as id variables
head(Mineral_Composition_S2_melt)
##   Depth variable value
## 1     0   Quartz  50.8
## 2     4   Quartz  37.9
## 3     5   Quartz  32.8
## 4  6.75   Quartz  38.4
## 5  8.25   Quartz  30.0
## 6  9.75   Quartz  45.1
```

3.5.2 Plot data

```
# Define the number of color that we want
mycolors = c(brewer.pal(name="Paired", n = 12), brewer.pal(name="Set3", n = 4))
# Now use the melted data frame to plot
Mineral_Composition_S2_Plot <-ggplot(Mineral_Composition_S2_melt, aes(x=value, y=factor(Depth, level = c('49.5','47.25','45.75','42.75','39.75','36.75','33.75','32.25','30.75','29.25','27.75','26.25','24.75','24','21.75','20.25','17.25','15.75','14.25','11.25','9.75','8.25','6.75','5','4','0')), fill= variable)) + geom_bar(position="fill", stat="identity") + xlab("Relative percentage\n\n Site 2\n\n(b)") + ylab("") + theme_classic()+ labs(fill = "Mineral composition") + scale_fill_manual(values = mycolors)
```

3.6 Combine plots

```
# Keep only one legend, run the codes in the following order

Mineral_legend <- get_legend(Mineral_Composition_S1_Plot)
Mineral_Composition_S1_Plot <- Mineral_Composition_S1_Plot + theme(legend.position="none")
Mineral_Composition_S2_Plot <- Mineral_Composition_S2_Plot + theme(legend.position="none")
plot_grid(Mineral_Composition_S1_Plot, Mineral_Composition_S2_Plot, Mineral_legend, labels= c("A", "B"), widths=c(3.6,3.6,1), height =c(3.6, 3.6, 1), ncol = 3, nrow = 1)
```



3.11 Spearman correlation

Use the following codes to calculate correlation between each mineral composants and depth.

3.11.1 Prepare data

```
# Depth should now be a numerical value
Mineral_Composition_S1$Depth <- as.numeric(Mineral_Composition_S1$Depth) # Site 1
Mineral_Composition_S2$Depth <- as.numeric(Mineral_Composition_S2$Depth) # Site 2
```

3.11.2 Test significance

```
# Quartz, K-Feldpath, etc. as a function of depth (SITE 1)
cor.test(Mineral_Composition_S1$Depth, Mineral_Composition_S1$Quartz, method = "spearman")

# Quartz, K-Feldpath, etc. as a function of depth (SITE 2)
cor.test(Mineral_Composition_S2$Depth, Mineral_Composition_S2$Quartz, method = "spearman")

# Change Quartz with the other minerals (K.feldspar, Plagioclase, Calcite, Dolomite, etc.)
```

4. Microbiome data

This script was used to clean and normalise the datasets, calculate the diversity index, make the figures (staked bar Plot and nMDS), do a MANTEL and ANOSIM test. This script was the one used for the bacterial sequences detected at site 1 but the codes are similar for site 2 and for archaea and eukaryal data.

4.1 Set the working directory

```
setwd("~/ASV/CHAP1/BAC/site_1") # change to you working directory
```

4.2 Load libraries

```
library(microbiome) # Data analysis and visualisation
library(phyloseq) # Also the basis of data object. Data analysis and visualisation
library(microbiomeutilities) # Some utility tools
library(RColorBrewer) # Nice color options
library(ggpubr) # Publication quality figures, based on ggplot2
library(DT) # Interactive tables in html and markdown
library(data.table) # Alternative to data.frame
library(dplyr) # Data handling
```

4.3 Load data

```
asv_table <- read.csv("~/ASV/CHAP1/BAC/site_1/ASV_S1_t_table.csv", sep=",", row.names = 1)
taxonomy_table <- read.csv("~/ASV/CHAP1/BAC/site_1/BAC_S1_taxonomy_table.csv", sep=",", row.nam
es = 1)

Soil_Characteristics <- read.csv(file="Soil_Characteristics.csv", header=TRUE, sep=";", dec=".")
Mineral_Composition <- read.csv(file="Mineral_Composition.csv", header=TRUE, sep=";", dec=".")
Mineral_Composition <- Mineral_Composition[,-c(2,3)] # Get rid of the column #2 ('Depth') and column #
3 ('Site')

Abiotic_Factors <- merge(Soil_Characteristics,Mineral_Composition, by='Sample')
rownames(Abiotic_Factors) <- Abiotic_Factors$Sample # Add the rownames to samples file to simplify fur
ther analysis
```

4.4 Make a phyloseq object

```
# Transform into matrixes otu and tax tables (samples table can be left as data frame)
 asv_mat <- as.matrix(asv_table)
 tax_mat <- as.matrix(taxonomy_table)
ASV = otu_table(asv_mat, taxa_are_rows = TRUE)
TAX = tax_table(tax_mat)
SAMPLES = sample_data(Abiotic_Factors)
phyloseq_object <- phyloseq(ASV, TAX, SAMPLES)
```

4.5 Read the tree file

```
# Load tree file
library(ape)
tree <- read.tree(file = "~/ASV/CHAP1/BAC/site_1/tree")
tree.rooted <- root(tree, outgroup = 1, resolve.root=TRUE) # Root the tree
```

4.6 Merge into phyloseq object

```
phyloseq_object <- merge_phyloseq(phyloseq_object, tree.rooted)
rank_names(phyloseq_object) # We check the taxonomic rank information
datatable(tax_table(phyloseq_object)) # The table is interactive

#Visualize data
sample_names(phyloseq_object)
rank_names(phyloseq_object)
sample_variables(phyloseq_object)
```

4.7 Clean taxonomy table

Cleaning of taxonomy tables is useful to do at the beginning of the analysis.

```
# Replace sequence IDs with ATACAACTATACG with ASV1 and so on.
taxa_names(phyloseq_object) <- paste0("ASV", seq(ntaxa(phyloseq_object)))

# We will filter our phyloseq object because we only intended to amplify bacterial sequences.
phyloseq_object <- phyloseq_object %>% subset_taxa(domain == "Bacteria") # Change to "Archaea" or "Eukaryota" if needed

tax_table(phyloseq_object)[, colnames(tax_table(phyloseq_object))] <- gsub(tax_table(phyloseq_object)[, colnames(tax_table(phyloseq_object))], pattern = "[a-z]__", replacement = "")

tax_table(phyloseq_object)[tax_table(phyloseq_object)[, "phylum"] == "", "phylum"] <- "Unidentified"
datatable(tax_table(phyloseq_object))

taxa_names(phyloseq_object)[1:5] # print first 5 names
## [1] "ASV1" "ASV2" "ASV3" "ASV4" "ASV5"
```

4.8 Access parts

```
otu_tab <- microbiome::abundances(phyloseq_object) # ASV table
tax_tab <- phyloseq::tax_table(phyloseq_object) # Taxonomy table
```

4.9 Alpha diversity metrics

This part of the script is used to normalize data and calculated the alpha diversity metrics in each sample.

4.9.1 Load packages

```
library(microbiome) # Data analysis and visualisation
library(phyloseq) # Also the basis of data object. Data analysis and visualisation
library(microbiomeutilities) # Some utility tools
library(RColorBrewer) # Nice color options
library(ggpubr) # Publication quality figures, based on ggplot2
library(DT) # Interactive tables in html and markdown
library(data.table) # Alternative to data.frame
library(dplyr) # Data handling
```

4.9.2 Normalization: Median sequencing depth

```
summary(sample_sums(phyloseq_object))
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2972    5466    7531    7724    9186   14032

otu_tab <- t(abundances(phyloseq_object))
p <- vegan::rarecurve(otu_tab, step = 50, label = FALSE,  sample = min(rowSums(otu_tab),   col = "blue", c
ex = 0.6))
```

```
rowSums(otu_tab)
##      RR.01.bac      RR.02.bac      RR.03.bac      RR.04.bac      RR.05.bac
##           3368           8281           7531           6568           7843
##      RR.06.bac      RR.07.bac      RR.08.bac      RR.09.bac      RR.10.bac
##           8858           2972          14032          13078           7353
##      RR.11.bac      RR.12.bac      RR.13.bac      RR.14.bac      RR.15.bac
##           5040           6474           9186          11554           9683
##      RR.16.bac      RR.17.bac      RR.18.bac      RR.19.bac      RR.20.bac
##           8656          13114           5965           3347           3837
## RR.eau.puit.bac
##           5466

# Keep only the samples with more than 1000 sequences (not the case here but will be for site 2, archaea
and eukaryotes. Use the following code :
# phyloseq_object <- subser_samples (phyloseq_object, Sample != "NDDL.07.bac")
Sample_names (phyloseq_object) # See if it worked



Set.seed(9242)  # This will help in reproducing the filtering and normalisation.

## Median sequencing depth
total = median(sample_sums(phyloseq_object))
standf = function(x, t=total) round(t * (x / sum(x)))
phyloseq_object_rar  = transform_sample_counts(phyloseq_object, standf)
 otu_tab_median <- t(abundances(phyloseq_object_rar))
```

```
rowSums(otu_tab_median)
##     RR.01.bac     RR.02.bac     RR.03.bac     RR.04.bac     RR.05.bac
##          7520          7532          7531          7524          7551
##     RR.06.bac     RR.07.bac     RR.08.bac     RR.09.bac     RR.10.bac
##          7535          7530          7535          7531          7492
##     RR.11.bac     RR.12.bac     RR.13.bac     RR.14.bac     RR.15.bac
##          7498          7503          7535          7533          7528
##     RR.16.bac     RR.17.bac     RR.18.bac     RR.19.bac     RR.20.bac
##          7532          7530          7534          7540          7531
## RR.eau.puit.bac
##          7541
print(phyloseq_object_rar) # Check how much data you have now changed
```

4.9.3 Register the rarefied ASV table and taxonomy table

```
write.csv(as.data.frame(phyloseq_object_rar@otu_table), file ="~/ASV/CHAP1/BAC/site_1/asv.table.rarified.S1.bac.csv")
write.csv(as.data.frame(phyloseq_object_rar@tax_table), file ="~/ASV/CHAP1/BAC/site_1/taxonomy.table.rarified.S1.bac.csv")
```

4.9.4 Taxonomic diversity

```
#Table with samples and diversity index
Taxonomic_Div <- alpha(phyloseq_object_rar, index = c("diversity_shannon", "diversity_gini_simpson"))
```

4.9.5 Phylogenetic diversity

```
library(picante)
phyloseq_object_rar_asvtab <- as.data.frame(phyloseq_object_rar@otu_table)
phyloseq_object_rar_tree <- phyloseq_object_rar@phy_tree
# We first need to check if the tree is rooted or not
phyloseq_object_rar@phy_tree
##
## Phylogenetic tree with 3163 tips and 3162 internal nodes.
##
## Tip labels:
##   ASV1, ASV2, ASV3, ASV4, ASV5, ASV6, ...
## Node labels:
##   0.999, 0.983, , 0.785, 0.894, 0.994, ...
##
## Rooted; includes branch lengths.
```

4.9.5.1 mntd index

```
phyloseq_object_rar_tree_dist <- cophenetic(phyloseq_object_rar_tree)
```

```
ses.mntd.result <- ses.mntd(t(phyloseq_object_rar_asvtab), phyloseq_object_rar_tree_dist, null.model= "taxa.labels", abundance.weighted=FALSE, runs=999, iterations = 1000)
```

4.9.5.2 PD index

```
ses.pd.result <-ses.pd(t(phyloseq_object_rar_asvtab), phyloseq_object_rar_tree, runs = 999, iterations = 1000, include.root=TRUE)

ses.pd.result$Sample <- rownames(ses.pd.result)
Taxonomic_Div$Sample <- rownames(Taxonomic_Div) # Add the Sample to diversity pd table
Diversity_tax_PD <- merge(Taxonomic_Div, ses.pd.result, by = "Sample") # Merge these two data frames
ses.mntd.result$Sample <- rownames(ses.mntd.result) # Add the rownames to diversity mntd table
Diversity_metrics <- merge(Diversity_tax_PD, ses.mntd.result, by = "Sample") # Merge with ses.mntd.res ult
colnames(Diversity_metrics) # check the tables
##  [1] "Sample"              "diversity_shannon"     "diversity_gini_simpson"
##  [4] "ntaxa.x"             "pd.obs"                "pd.rand.mean"
##  [7] "pd.rand.sd"          "pd.obs.rank"           "pd.obs.z"
## [10] "pd.obs.p"            "runs.x"                "ntaxa.y"
## [13] "mntd.obs"            "mntd.rand.mean"        "mntd.rand.sd"
## [16] "mntd.obs.rank"       "mntd.obs.z"            "mntd.obs.p"
## [19] "runs.y"
```

4.9.5.3 Clean the diversity table with only the variables of interest

```
Diversity_metrics_clean <-Diversity_metrics[,c("Sample","diversity_shannon","diversity_gini_simpson","pd.obs", "pd.obs.z" ,"mntd.obs","mntd.obs.z")] # Keep sample names, simpson,shannon, PD, ses.PD, MNTD and ses.MNTD
write.csv(as.data.frame(Diversity_metrics_clean), file ="~/ASV/CHAP1/BAC/site_1/diversity.S1.bac.csv") # Register
```

See section 5 to see how to test differences in alpha diversity with spearman correlation, to explore alpha metrics, make the graphics and the Heatmaps.

4.10 Composition

Barplots are a one way of visualising the composition of the samples. We will use the filtered phyloseq object from Set-up and Pre-processing section.

4.10.1 Load packages

```
library(microbiome) # Data analysis and visualisation
library(phyloseq) # Also the basis of data object. Data analysis and visualisation
library(RColorBrewer) # Nice color options
library(ggpubr) # Publication quality figures, based on ggplot2
library(dplyr) # Data handling
```

4.10.2 Barplot counts

```
phyloseq_object_com <- phyloseq_object_rar # We need to set Palette
taxic <- as.data.frame(phyloseq_object_com@tax_table) # This will help in setting large color options
# colourCount = length(unique(taxic$Family))  #define number of variable colors based on number of Fam
ily (change the level accordingly to phylum/class/order)
# getPalette = colorRampPalette(brewer.pal(12, "Paired"))  # Change the palette as well as the number of
 colors will change according to palette.

taxic$OTU <- rownames(taxic) # Add the OTU ids from OTU table into the taxa table at the end.
```

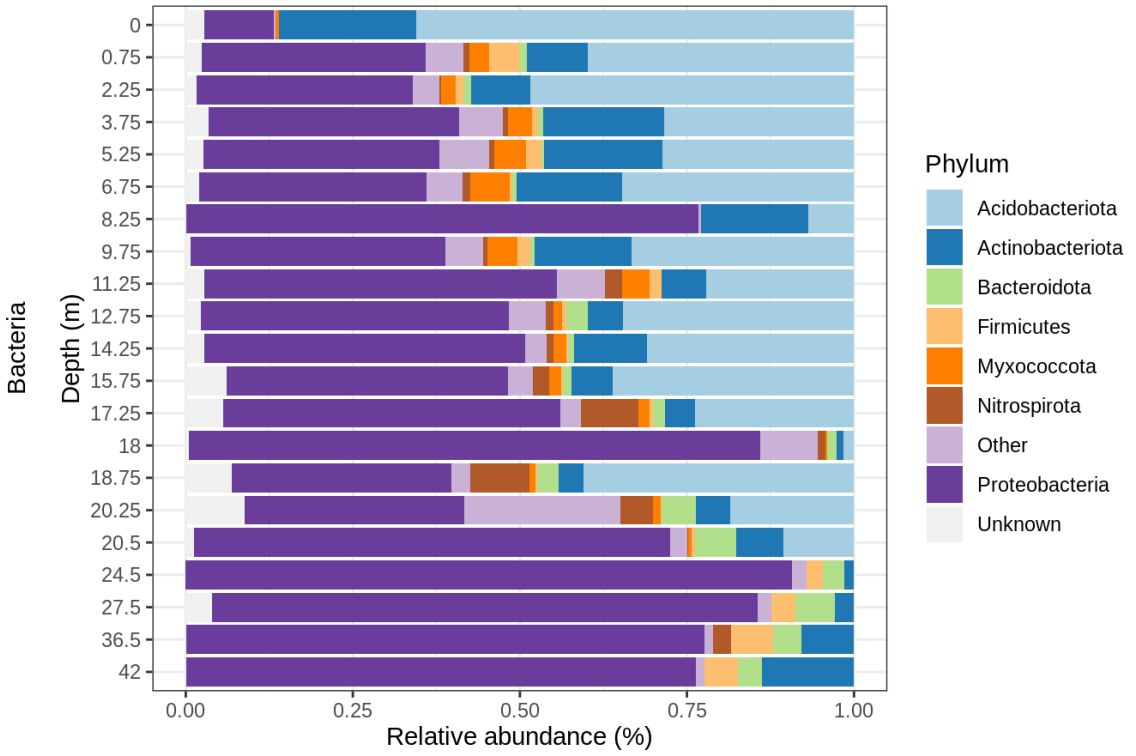
4.10.3 Phylum

```
taxmat <- as.matrix(taxic) # Convert it into a matrix.
new.tax <- tax_table(taxmat) # Convert into phyloseq compatible file.
tax_table(phyloseq_object_com) <- new.tax # incroporate into phyloseq Object
tax_table(phyloseq_object_com)[tax_table(phyloseq_object_com)[, "phylum"] == "", "phylum"] <- "Uncl
assified phylum" # Now edit the unclassified taxa

## Now we need to plot at phylum level, we can do it as follows:
# first remove the phy_tree
phyloseq_object_com@phy_tree <- NULL

# Second merge at phylum level
phyloseq_object_com_phyl <- microbiome::aggregate_top_taxa(phyloseq_object_com, "phylum", top =
8) # The previous pseq object phyloseq_object_com_phyl is only counts.
mycolors =  c("#A6CEE3","#1F78B4","#B2DF8A","#FDBF6F","#FF7F00","#B15928","#CAB2D6", "#6A3D9A
","#F0F0F0")

# Use traqnsform function of microbiome to convert it to rel abun.
phyloseq_object_com_phyl_rel <- microbiome::transform(phyloseq_object_com_phyl, "compositional")
plot.composition.relAbun.phyl <- plot_composition(phyloseq_object_com_phyl_rel,  sample.sort = "Dept
h",  x.label  = "Depth")
plot.composition.relAbun.phyl <- plot.composition.relAbun.phyl + theme(legend.position = "bottom")
plot.composition.relAbun.phyl <- plot.composition.relAbun.phyl + scale_fill_manual("Phylum",values = m
ycolors) + theme_bw()
plot.composition.relAbun.phyl <- plot.composition.relAbun.phyl + theme(axis.text.x = element_text(angl
e = 90))
plot.composition.relAbun.phyl <- plot.composition.relAbun.phyl + ggtitle("Relative abundance") + theme
(legend.title = element_text(size = 18))
```

4.10.3.1 Let ’s customize the Barplot

```
data.com.phyl <- plot.composition.relAbun.phyl$data
p.com.phyl <- ggplot(data.com.phyl, aes(y = factor(xlabel, level = c('42','36.5','27.5','24.5','20.5','20.25','1
8.75','18','17.25','15.75','14.25','12.75','11.25','9.75','8.25','6.75','5.25','3.75','2.25','0.75','0')), x = Abunda
nce, fill = Tax))
```

```
p.com.phyl <- p.com.phyl + geom_bar(position = "stack", stat = "identity")
p.com.phyl <- p.com.phyl + scale_fill_manual("Phylum",values = mycolors)+ theme_bw()
p.com.phyl <- p.com.phyl + ylab("Bacteria\n\nDepth (m)") + xlab ("Relative abundance (%)\n\n(a)")
p.com.phyl
```



(a)

```
### Phylum prevalence (relative abundance)
mean(data.com.phyl[data.com.phyl$Tax == 'Proteobacteria',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Acidobacteriota',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Actinobacteriota',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Other',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Unknown',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Bacteroidota',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Nitrospirota',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Myxococcota',]$Abundance)*100
mean(data.com.phyl[data.com.phyl$Tax == 'Firmicutes',]$Abundance)*100
```

4.10.3.2 Correlation with major phylum

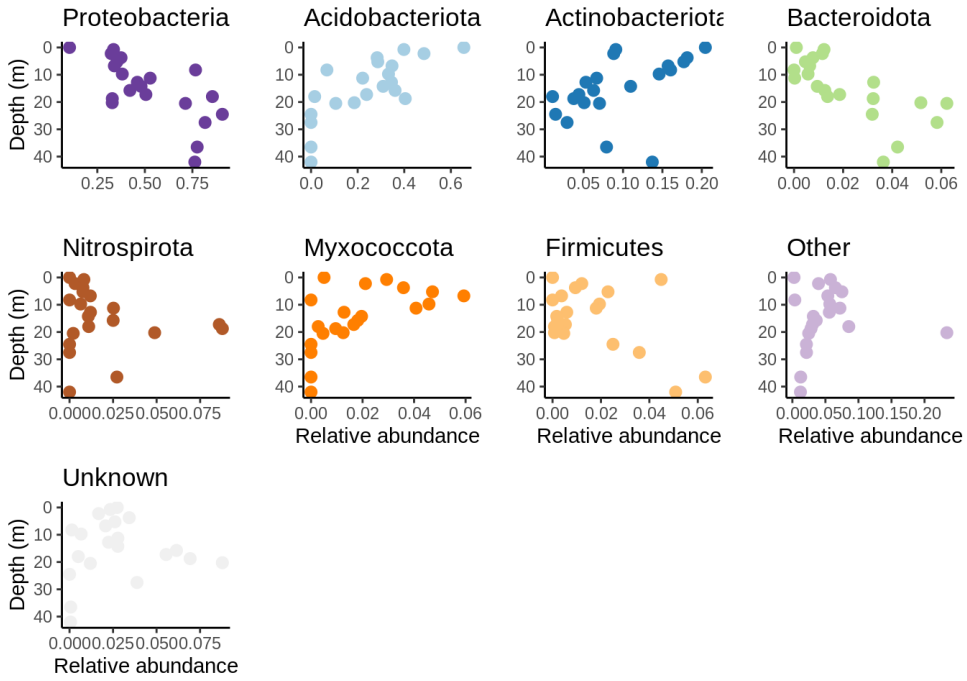
```
## Choose the color with #display.brewer.pal(n = 10, name = 'Paired') and #brewer.pal(n = 10, name = "Paired")

#Proteobacteria
Proteobacteria.depth <- merge(data.com.phyl[data.com.phyl$Tax == 'Proteobacteria',], Abiotic_Factors[, c(1,3)], by = "Sample")
```

```
Proteobacteria.depth.NOGW <- Proteobacteria.depth[Proteobacteria.depth$Sample!="RR.eau.puit.bac",]

cor.test(Proteobacteria.depth.NOGW$Depth, Proteobacteria.depth.NOGW$Abundance ,method = "spearman") #Abundance phyla as a function of depth
Proteobacteria.depth.graph <- ggplot(Proteobacteria.depth, aes(x = Depth, y = Abundance))+
  geom_point(color = "#6A3D9A" , size=2.5) + coord_flip() + scale_x_reverse()+ theme_bw() + xlab("Depth (m)") + ylab("") + theme_classic() + ggtitle("Proteobacteria")
```

Repeat this code (section 4.10.3.2) with the other major phyla and combine plot when it's done.

4.10.3.3 Combine graphs

```
library("RColorBrewer")
library("ggplot2")
library("cowplot")
plot_grid(Proteobacteria.depth.graph, Acidobacteriota.depth.graph, Actinobacteriota.depth.graph, Bacteroidota.depth.graph, Nitrospirota.depth.graph, Myxococcota.depth.graph,Firmicutes.depth.graph,Other.depth.graph, Unknown.depth.graph, ncol = 4, nrow = 3)
```



4.10.4 Genus

```
tax_table(phyloseq_object_com)[tax_table(phyloseq_object_com)[, "genus"] == "", "genus"] <- "Unclassified genus" # Now edit the unclassified taxa


## Now we need to plot at class level, we can do it as follows:
# First remove the phy_tree
```

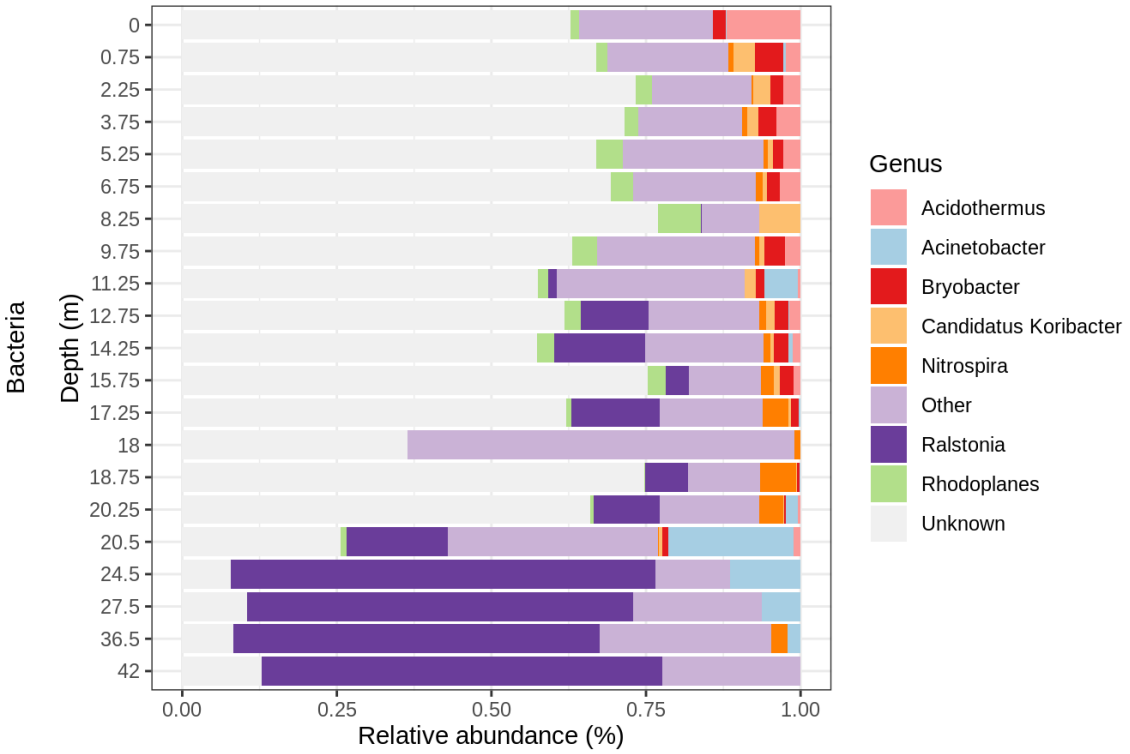
```
phyloseq_object_com@phy_tree <- NULL

# Second merge at phylum level
phyloseq_object_com_genus <- microbiome::aggregate_top_taxa(phyloseq_object_com, "genus", top = 8)
```

4.10.4.1 Barplot relative abundance

```
# The previous pseq object phyloseq_object_com_order is only counts.
mycolors_genus =  c("#FB9A99","#A6CEE3","#E31A1C","#FDBF6F","#FF7F00","#CAB2D6","#6A3D9A","#B2DF8A","#F0F0F0")

# Use traqnsform function of microbiome to convert it to rel abun.
phyloseq_object_com_genus_rel <- microbiome::transform(phyloseq_object_com_genus, "compositional")
plot.composition.relAbun.genus <- plot_composition(phyloseq_object_com_genus_rel, sample.sort = "Depth", x.label = "Depth")
plot.composition.relAbun.genus <- plot.composition.relAbun.genus + theme(legend.position = "bottom")

plot.composition.relAbun.genus <- plot.composition.relAbun.genus + scale_fill_manual("Genus",values =
 mycolors_genus) + theme_bw()
plot.composition.relAbun.genus <- plot.composition.relAbun.genus + theme(axis.text.x = element_text(angle = 90))
plot.composition.relAbun.genus <- plot.composition.relAbun.genus + ggtitle("Relative abundance")+ theme(legend.title = element_text(size = 18))
```

4.10.4.2 Barplot customize

```
data.com.genus <- plot.composition.relAbun.genus$data
p.com.genus <- ggplot(data.com.genus, aes(y = factor(xlabel, level = c('42','36.5','27.5','24.5','20.5','20.25','18.75','18','17.25','15.75','14.25','12.75','11.25','9.75','8.25','6.75','5.25','3.75','2.25','0.75','0')), x = Abundance, fill = Tax))
p.com.genus <- p.com.genus + geom_bar(position = "stack", stat = "identity")
p.com.genus <- p.com.genus + scale_fill_manual("Genus",values = mycolors_genus) + theme_bw()
p.com.genus <- p.com.genus + ylab("Bacteria\n\nDepth (m)") + xlab ("Relative abundance (%)\n\n(a)")
p.com.genus
```



(a)

```
### Genus prevalence
mean(data.com.genus[data.com.genus$Tax == 'Unknown',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Other',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Ralstonia',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Acinetobacter',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Acidothermus',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Bryobacter',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Rhodoplanes',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Candidatus Koribacter',]$Abundance)*100
mean(data.com.genus[data.com.genus$Tax == 'Nitrospira',]$Abundance)*100
```

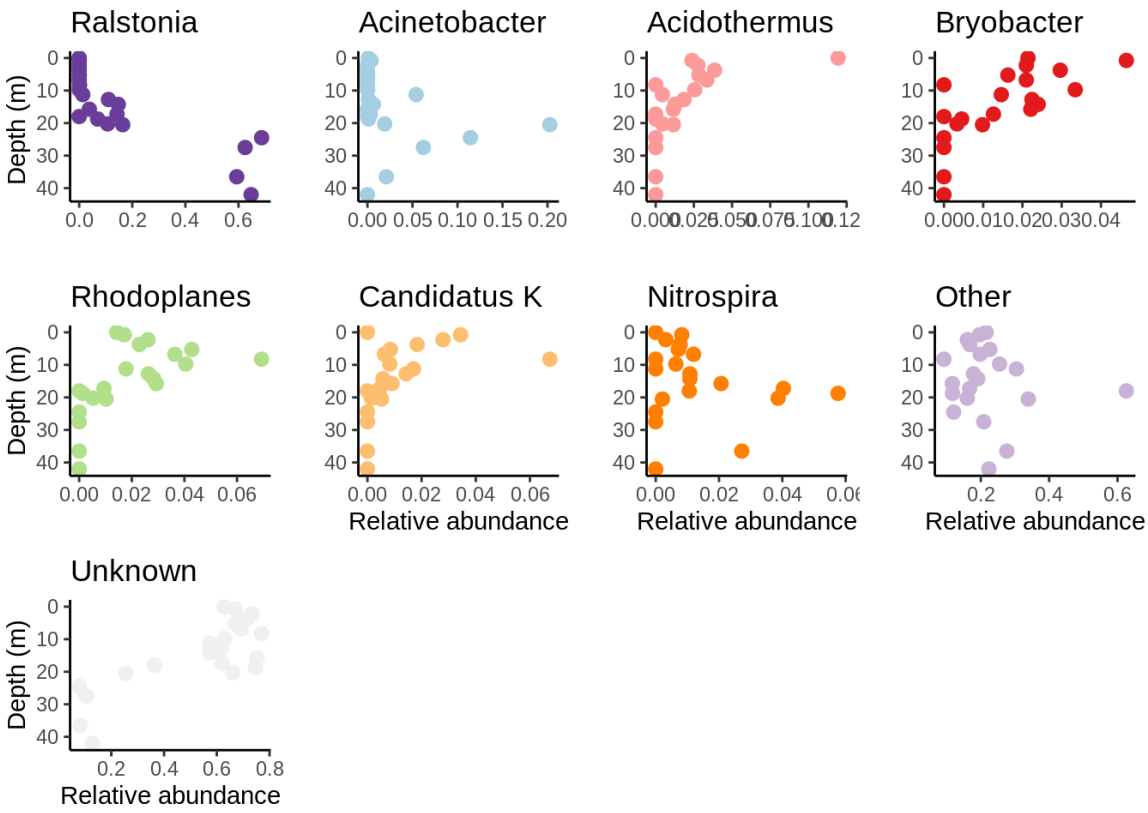
4.10.4.3 Correlation with major genus

Repat the code used in section 4.10.3.2. Select the colors you want to use with :

```
#display.brewer.pal(n = 11, name = 'Paired')
#brewer.pal(n = 11, name = "Paired")
```

4.10.4.4 Combine graphs

```
plot_grid(Ralstonia.gen.depth.graph, Acinetobacter.gen.depth.graph,Acidothermus.gen.depth.graph,Bryobacter.gen.depth.graph, Rhodoplanes.gen.depth.graph,Candidatus.Koribacter.gen.depth.graph, Nitrospira.gen.depth.graph, Other.gen.depth.graph,Unknown.gen.depth.graph, ncol = 4, nrow = 3)
```



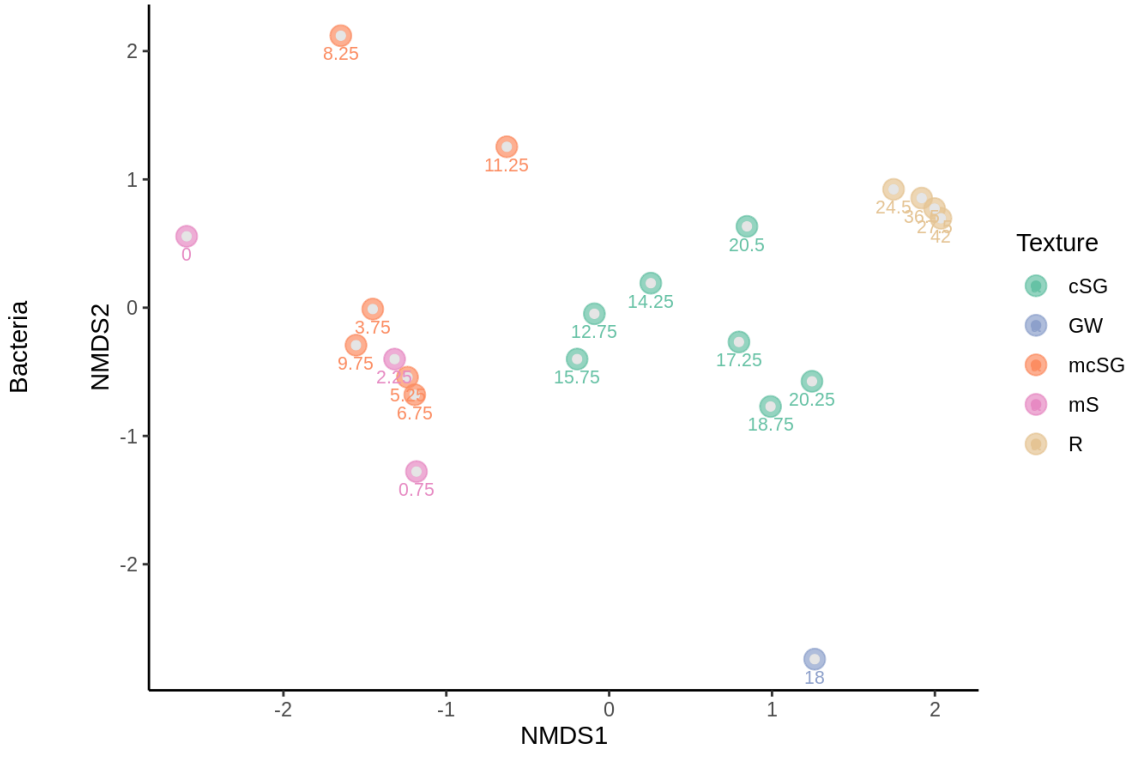
4.11 Beta diversity metrics

```
library(microbiome) # Data analysis and visualisation
library(phyloseq) # Also the basis of data object. Data analysis and visualisation
library(RColorBrewer) # Nice color options
library(ggpubr) # Publication quality figures, based on ggplot2
library(dplyr) # Data handling
summarize_phyloseq(phyloseq_object_rar) # Use this to summarize the normalisez phyloseq object
```

4.11.1 Ordination with Bray-Curtis distance matrix

```
phyloseq_object_rel <- microbiome::transform(phyloseq_object_rar, "compositional") # We used the nor
malised data
set.seed(1)
# Ordinate : http://deneflab.github.io/MicrobeMiseq/demos/mothur_2_phyloseq.html#unconstrained_or
dinations

erie_nmds_bray <- ordinate( physeq = phyloseq_object_rel,  method = "NMDS",  distance = "bray")
nmds_S1_bac_bray <- plot_ordination( physeq = phyloseq_object_rel, ordination = erie_nmds_bray ,
 color = "Texture")+ scale_colour_manual(values = c("#66C2A5","#8DA0CB" ,"#FC8D62","#E78AC3", "#E5
C494")) + geom_point(aes(color = Texture), alpha = 0.7, size = 4) + geom_point(colour = "grey90", size =
1.5) + theme_classic() + geom_text(aes(label = Depth), size = 2.8, vjust = 1.9) + xlab(("NMDS1\n\n(a)")) +
ylab("Bacteria\n\n\nNMDS2")
```

nmds_S1_bac_bray



(a)

4.11.2 Statistically test for beta-diversity

The following test were use to determine correlations between the beta diversy (estimated with Bray-Curtis matrix distance) and the abiotic factors like sedimentary texture (ANOSIM) or Depth (Mantel test).

4.12.2.1 ANOSIM test (Composition and sedimentary texture)

```
phyloseq_object_rel_wt_gwt <- subset_samples(phyloseq_object_rel, Texture != 'GW') # Keep only samp
les to be analyzed
erie_bray <- phyloseq::distance(phyloseq_object_rel_wt_gwt, method = "bray") # Calculate bray curtis di
stance matrix

# Make a data frame from the sample_data
metadf <- data.frame(sample_data(phyloseq_object_rel_wt_gwt))
anosim(erie_bray, metadf$Texture, permutations = 1000)
##
## Call:
## anosim(x = erie_bray, grouping = metadf$Texture, permutations = 1000)
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.6429
```

```
##      Significance: 0.000999
##
## Permutation: free
## Number of permutations: 1000
```

4.12.2.2 Mantel test

```
mineralogy <-metadf[,c(10:25)] # Environmental vector, make sure values are numerical
library(vegan)
dist1 = vegdist(erie_bray)  # Community matrix
dist2 = vegdist (metadf$Depth) # Change Depth with other abiotic factors like pH or change metadf$Depth with mineralogy
mantel(dist1, dist2, method = "spearman", permutations=9999) #9999 permutations
```

5. Alpha diversity metrics

This script was used to make the alpha diversity figures and to calculate the correlations between abiotic factors and alpha diversity metrics. To do so, we used the apha diversity metrics calculated in section 4.9.

5.1 Set the working directory

```
setwd("~/ASV/CHAP1/Alpha_Diversity") # Change to you working directory
```

5.2 Load data

```
Sediment_Characteristics <- read.csv2(file="~/ASV/CHAP1/Alpha_Diversity/Sediment_Characteristics.2.csv", header=TRUE, sep=";", dec=".", row.names = 1) # Sediment Characteristics

# Bacterial diversity
diversity.S1.bac <- read.csv("~/ASV/CHAP1/Alpha_Diversity/diversity.S1.bac.csv", sep=",", row.names = 1)
diversity.S2.bac <- read.csv("~/ASV/CHAP1/Alpha_Diversity/diversity.S2.bac.csv", sep=",", row.names = 1)

# Archaeal diversity
diversity.S1.arc <- read.csv("~/ASV/CHAP1/Alpha_Diversity/diversity.S1.arc.csv", sep=",", row.names = 1)
diversity.S2.arc <- read.csv("~/ASV/CHAP1/Alpha_Diversity/diversity.S2.arc.csv", sep=",", row.names = 1)

#Eucaryotes diversity
diversity.S1.euc <- read.csv("~/ASV/CHAP1/Alpha_Diversity/diversity.S1.euc.csv", sep=",", row.names = 1)
diversity.S2.euc <- read.csv("~/ASV/CHAP1/Alpha_Diversity/diversity.S2.euc.csv", sep=",", row.names = 1)

# The names are not pretty. we can replace them
colnames(diversity.S1.bac) <- c("Sample", "Shannon","Simpson", "PD", "SES.PD","MNTD","SES.MNTD")
colnames(diversity.S2.bac) <- c("Sample", "Shannon","Simpson", "PD", "SES.PD","MNTD","SES.MNTD")
colnames(diversity.S1.arc) <- c("Sample", "Shannon","Simpson", "PD", "SES.PD","MNTD","SES.MNTD")
colnames(diversity.S2.arc) <- c("Sample", "Shannon","Simpson", "PD", "SES.PD","MNTD","SES.MNTD")
colnames(diversity.S1.euc) <- c("Sample", "Shannon","Simpson", "PD", "SES.PD","MNTD","SES.MNTD")
colnames(diversity.S2.euc) <- c("Sample", "Shannon","Simpson", "PD", "SES.PD","MNTD","SES.MNTD")
```

5.3 Load library

```
library("tidyverse")
library("rstatix")
library("ggpubr")
library("gridExtra")
library("cowplot")
library("Hmisc")       # for correlations and p-values
```

```
library("RColorBrewer")  # for color palette
library("gplots")
```

5.4 Prepare data

5.4.1 Prepare Bacteria data

```
diversity.bac <- rbind(diversity.S1.bac, diversity.S2.bac) # Merge these two data frames into one
Sediment_Characteristics$Sample <- rownames(Sediment_Characteristics) # Add the rownames to sedim
ent Characteristics table
Diversity_bac_SC <- merge(Sediment_Characteristics, diversity.bac, by = "Sample")  # Merge Soil_Charact
eristics and diversity.bac into one
```

5.4.2 Prepare Archaea data

```
diversity.arc <- rbind(diversity.S1.arc, diversity.S2.arc) # Merge these two data frames into one
Sediment_Characteristics$Sample <- rownames(Sediment_Characteristics) # Add the rownames to Sedim
ent Characteristics table
Diversity_arc_SC <- merge(Sediment_Characteristics, diversity.arc, by = "Sample") # Merge these two dat
a frames into one
```

5.4.2 Prepare Eukaryota data

```
EUCARYOTES :
diversity.euc <- rbind(diversity.S1.euc, diversity.S2.euc) # Merge these two data frames into one
Sediment_Characteristics$Sample <- rownames(Sediment_Characteristics) # Add the rownames to Soil_C
haracteristics table
Diversity_euc_SC <- merge(Sediment_Characteristics, diversity.euc, by = "Sample") # Merge these two da
ta frames into one
```

5.5 Plot Diversity

5.5.1 Bacteria

```
shannon.bac <- ggplot(Diversity_bac_SC, aes(x = Depth, y = Shannon, colour = Site, shape= Category))+  g
eom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw() + xlab("Bacteri
a\n\nDepth (m)") + ylab("Shannon\n\n(a)") + theme_classic() + scale_shape_manual(values=c(1,8,16,0,8,
15))+ theme(legend.position="none")
pd.bac <- ggplot(Diversity_bac_SC, aes(x = Depth, y = PD, colour = Site,shape= Category))+ geom_point(si
ze=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse()+ theme_bw() + xlab("") + ylab("Faith'PD\n
\n(c)") + theme_classic() + scale_shape_manual(values=c(1,8,16,0,8,15))+ theme(legend.position="none
")
ses.pd.bac <- ggplot(Diversity_bac_SC, aes(x = Depth, y = SES.PD, colour = Site, shape= Category))+  geom
_point(size=2.5) + coord_flip() + scale_x_reverse()+ theme_bw() + xlab(" ") + ylab("SES.PD") + scale_color
_grey() + theme_classic() + scale_shape_manual(values=c(1,8,16,0,8,15)) + theme(legend.position="none
")
```

```
simpson.bac <- ggplot(Diversity_bac_SC, aes(x = Depth, y = Simpson, colour = Site, shape= Category))+ geom_point(size=2.5)  + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+  xlab("") + ylab("Simpson\n\n(b)")+ theme_classic() + scale_shape_manual(values=c(1,8,16,0,8,15))+ theme(legend.position="none")
mntd.bac <- ggplot(Diversity_bac_SC, aes(x = Depth, y = MNTD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+ xlab("") + ylab("MNTD\n\n(d)") + theme_classic()+ scale_shape_manual(values=c(1,8,16,0,8,15))+ theme(legend.position="none")
ses.mntd.bac <- ggplot(Diversity_bac_SC, aes(x = Depth, y = SES.MNTD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+  xlab("  ") + ylab("SES.MNTD") + theme_classic()+ scale_shape_manual(values=c(1,8,16,0,8,15)) + theme(legend.position="none")
```

5.5.2 Archaea

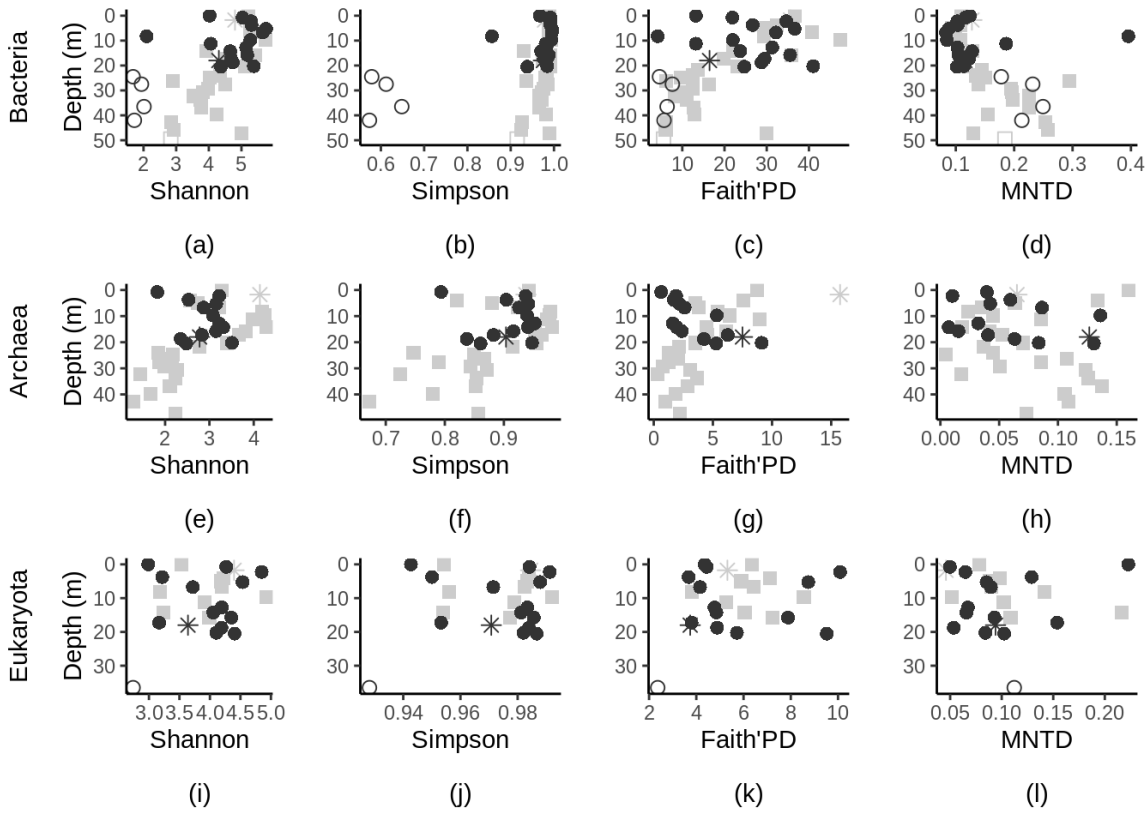
```
shannon.arc <- ggplot(Diversity_arc_SC, aes(x = Depth, y = Shannon, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw() + xlab("Archaea\n\nDepth (m)") + ylab("Shannon\n\n(e)") + theme_classic()+ scale_shape_manual(values=c(8,16,8,15)) + theme(legend.position="none")
pd.arc <- ggplot(Diversity_arc_SC, aes(x = Depth, y = PD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse()+ theme_bw()+  xlab("") + ylab("Faith'PD\n\n(g)") + theme_classic()+ scale_shape_manual(values=c(8,16,8,15)) + theme(legend.position="none")

ses.pd.arc <- ggplot(Diversity_arc_SC, aes(x = Depth, y = SES.PD, colour = Site, shape= Category))+
  geom_point(size=2.5) + coord_flip() + scale_x_reverse()+ theme_bw() + xlab(" ") + ylab("SES.PD") + scale_color_grey() + theme_classic()+ scale_shape_manual(values=c(8,16,8,15)) + theme(legend.position="none")

simpson.arc <- ggplot(Diversity_arc_SC, aes(x = Depth, y = Simpson, colour = Site, shape= Category))+ geom_point(size=2.5)  + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+  xlab("") + ylab("Simpson\n\n(f)")+ theme_classic()+ scale_shape_manual(values=c(8,16,8,15)) + theme(legend.position="none")

mntd.arc <- ggplot(Diversity_arc_SC, aes(x = Depth, y = MNTD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+  xlab("") + ylab("MNTD\n\n(h)") + theme_classic()+ scale_shape_manual(values=c(8,16,8,15))+ theme(legend.position="none")

ses.mntd.arc <- ggplot(Diversity_arc_SC, aes(x = Depth, y = SES.MNTD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+  xlab("") + ylab("SES.MNTD (Archaea)") + theme_classic()+ scale_shape_manual(values=c(8,16,8,15))+ theme(legend.position="none")
```

5.5.3 Eucaryotes

```
shannon.euc <- ggplot(Diversity_euc_SC, aes(x = Depth, y = Shannon, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+ xlab("Eukaryota\n\nDepth (m)") + ylab("Shannon\n\n(i)") + theme_classic()+ scale_shape_manual(values=c(1,8,16,8,15)) + theme(legend.position="none")

pd.euc <- ggplot(Diversity_euc_SC, aes(x = Depth, y = PD, colour = Site,shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse()+ theme_bw()+ xlab("") + ylab("Faith'PD\n\n(k)") + theme_classic()+ scale_shape_manual(values=c(1,8,16,8,15))+ theme(legend.position="none")

ses.pd.euc <- ggplot(Diversity_euc_SC, aes(x = Depth, y = SES.PD, colour = Site, shape= Category))+ geom_point(size=2.5) + coord_flip() + scale_x_reverse()+ theme_bw() + xlab(" ") + ylab("SES.PD") + scale_color_grey() + theme_classic()+ scale_shape_manual(values=c(1,8,16,8,15))+ theme(legend.position="none")

simpson.euc <- ggplot(Diversity_euc_SC, aes(x = Depth, y = Simpson, colour = Site, shape= Category))+ geom_point(size=2.5)  + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+ xlab("") + ylab("Simpson\n\n(j)")+ theme_classic()+ scale_shape_manual(values=c(1,8,16,8,15))+ theme(legend.position="none")

mntd.euc <- ggplot(Diversity_euc_SC, aes(x = Depth, y = MNTD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+  xlab("") + ylab("MNTD\n\n(l)") + theme_classic()+ scale_shape_manual(values=c(1,8,16,8,15))+ theme(legend.position="none")

ses.mntd.euc <- ggplot(Diversity_euc_SC, aes(x = Depth, y = SES.MNTD, colour = Site, shape= Category))+ geom_point(size=2.5) + scale_color_grey() + coord_flip() + scale_x_reverse() + theme_bw()+ xlab("  ") + ylab("SES.MNTD") + theme_classic()+ scale_shape_manual(values=c(1,8,16,8,15))+ theme(legend.position="none")
```

5.6 Combine plots

```
plot_grid(shannon.bac,simpson.bac,  pd.bac,mntd.bac,shannon.arc,simpson.arc,  pd.arc,mntd.arc,shannon.euc,simpson.euc, pd.euc,mntd.euc,  ncol = 4, nrow = 3)
```



5.7 Spearman correlations

Here, we will determine the correlation between the bacterial alpha diversity metrics and the abiotic factors. Script are similar for archaea and eukaryotes.

5.7.1 Prepare data

```
# Merge these two data frames into one
Diversity.S1.SC.bac <- merge(Sediment_Characteristics, diversity.S1.bac, by = "Sample")
Diversity.S2.SC.bac <- merge(Sediment_Characteristics, diversity.S2.bac, by = "Sample")

Diversity.S1.SC.bac <- Diversity.S1.SC.bac[Diversity.S1.SC.bac$Texture!='GW',] # Samples from site 1 and without the groundwater sample for statistical tests
Diversity.S2.SC.bac <- Diversity.S2.SC.bac[Diversity.S2.SC.bac$Texture!='GW',] # Samples from site 1 and without the groundwater sample for statistical tests
```

5.7.2 Test significance

```
# Diversity as a function of Depth/pH/TN/TC/TOC/TIC (SITE 1)
cor.test(Diversity.S1.SC.bac$Depth, Diversity.S1.SC.bac$Shannon, method = "spearman")
cor.test(Diversity.S1.SC.bac$Depth, Diversity.S1.SC.bac$Simpson, method = "spearman")
cor.test(Diversity.S1.SC.bac$Depth, Diversity.S1.SC.bac$PD, method = "spearman")
```

```
cor.test(Diversity.S1.SC.bac$Depth, Diversity.S1.SC.bac$MNTD, method = "spearman")

# Diversity as a function of Depth/pH/TN/TC/TOC/TIC (SITE 2)
cor.test(Diversity.S2.SC.bac$Depth, Diversity.S2.SC.bac$Shannon, method = "spearman")
cor.test(Diversity.S2.SC.bac$Depth, Diversity.S2.SC.bac$Simpson, method = "spearman")
cor.test(Diversity.S2.SC.bac$Depth, Diversity.S2.SC.bac$PD, method = "spearman")
cor.test(Diversity.S2.SC.bac$Depth, Diversity.S2.SC.bac$MNTD, method = "spearman")

# This is an example with depth as abiotic factors. Change depth with pH/TN/TC/TOC/TIC to determine th
e correlation with these abiotic factors
```

6. Biotic interactions

The aim of the script is to create a Heatmap of biotic interactions between the three domains. The Heatmap will be similar to the Figure 3 found in the article: https://doi.org/10.1038/s41598-019-53975-9.

6.1 Load library

```
library(funrar)
library(ggplot2)
library(ggcorrplot)
library(Hmisc)
library(tidyverse)
library(dplyr)
library(plyr)
library(reshape2)
library(car)
library(phyloseq)
library(metagenomeSeq)
```

6.2 Set directory and load data

The data are the ASV normalised in section 4.9.

```
setwd("~/ASV/CHAP1/BIOTIC/S1")
path="~/ASV/CHAP1/BIOTIC/S1"

bac_otu = read.csv(file="asv.table.rarified.S1.bac.csv", header=TRUE, sep=";", dec=".", row.names = 1)
euka_otu = read.csv(file="asv.table.rarified.S1.euc.csv", header=TRUE, sep=";", dec=".",row.names = 1)
arc_otu = read.csv(file="asv.table.rarified.S1.arc.csv", header=TRUE, sep=";", dec=".",row.names = 1)

#Get rid of groundwater samples and surface samples
bac_otu <- bac_otu[,-c(1,21)]
euka_otu <- euka_otu[,-c(1,15)]
arc_otu <- arc_otu[,-c(14)]

colnames(bac_otu) = colnames(bac_otu) %>% str_replace(".bac", "")
colnames(euka_otu) = colnames(euka_otu) %>% str_replace(".euc", "")
colnames(arc_otu) = colnames(arc_otu) %>% str_replace(".arc", "")


rownames(bac_otu) = rownames(bac_otu) %>% str_replace("ASV", "Bac-")
rownames(euka_otu) = rownames(euka_otu) %>% str_replace("ASV", "Euk-")
rownames(arc_otu) = rownames(arc_otu) %>% str_replace("ASV", "Arc-")


# Make a phyloseq object
```

```
bac_data = phyloseq(otu_table(bac_otu, taxa_are_rows = TRUE))
euka_data = phyloseq(otu_table(euka_otu, taxa_are_rows = TRUE))
arc_data = phyloseq(otu_table(arc_otu, taxa_are_rows = TRUE))

# Transform the phyloseq into a data_frame
bac_data = as.data.frame(get_taxa(bac_data))
euka_data = as.data.frame(get_taxa(euka_data))
arc_data = as.data.frame(get_taxa(arc_data))

## Select only the samples that contains the 3 domains
i <- intersect(intersect(colnames(bac_data), colnames(arc_data)),colnames(euka_data))
bac_data <- select(bac_data,all_of(i))
arc_data <- select(arc_data,all_of(i))
euka_data <- select(euka_data,all_of(i))
```

6.3 Transform abundance data in relative abundances

```
bac_data = t(bac_data)
euka_data = t(euka_data)
arc_data = t(arc_data)


bac_data = make_relative(as.matrix(bac_data))
euka_data = make_relative(as.matrix(euka_data))
arc_data = make_relative(as.matrix(arc_data))
```

We focus this study on the interactions between the most abundant bacteria/archaea/eukaryotes. The most abundant bacteria, archaea and eukaryotes will be calculated by adding the relative abundances of the ASVs in each of the samples. Only the 15 most abundant will be selected. The last column which represented the cumulative relative abundance is removed.

```
somme_rel_bac  = as.matrix(colSums(bac_data))
somme_rel_euka = as.matrix(colSums(euka_data))
somme_rel_arc  = as.matrix(colSums(arc_data))

bac_data = t(bac_data)
euka_data = t(euka_data)
arc_data = t(arc_data)


bac  = cbind(bac_data, somme_rel_bac)
euka = cbind(euka_data, somme_rel_euka)
arc  = cbind(arc_data, somme_rel_arc)

# Classified ASV according to their relative abundance
bac  = bac[order(bac[,ncol(bac)], decreasing = TRUE),]
euka = euka[order(euka[,ncol(euka)], decreasing = TRUE),]
```

```
arc  = arc[order(arc[,ncol(arc)], decreasing =TRUE),]

# Keep the 15 most abundant ASV
bac  = bac[1:15, -ncol(bac)]
euka = euka[1:15, -ncol(euka)]
arc  = arc[1:15, -ncol(arc)]
```

6.4 Add correlations between domains

```
bac = as.data.frame(bac)
euka = as.data.frame(euka)
arc = as.data.frame(arc)

bac_arc = bind_rows(bac, arc)
bac_euka = bind_rows(bac, euka)
euka_arc = bind_rows(euka, arc)

# Garder uniquement les échantillons présents dans les 2 tables
```

```
options(scipen = 9999)

# Fonction pour enlever les données répétitives

 get_lower_tri<-function(cormat){ cormat[lower.tri(cormat)] <- NA

   return(cormat) }


# Correlation matrix for Bacteria
bac_cor = rcorr(as.matrix(t(bac)), type = "spearman")
bac_cor_r = get_lower_tri(bac_cor$r)
bac_cor_p = get_lower_tri(bac_cor$P)


bac_p_long = reshape2::melt(bac_cor_p)
bac_r_long = reshape2::melt(bac_cor_r)
bac_r_long$p = bac_p_long$value
bac_r_long$value[is.na(bac_r_long$value)] = 0


# Correlation matrix for Eucaryotes
euka_cor = rcorr(as.matrix(t(euka)), type = "spearman")
euka_cor_r = get_lower_tri(euka_cor$r)
euka_cor_p = get_lower_tri(euka_cor$P)


euka_p_long = reshape2::melt(euka_cor_p)
euka_r_long = reshape2::melt(euka_cor_r)
```

```
euka_r_long$p = euka_p_long$value
euka_r_long$value[is.na(euka_r_long$value)] = 0


# Correlation matrix for Archaea
arc_cor = rcorr(as.matrix(t(arc)), type = "spearman")
arc_cor_r = get_lower_tri(arc_cor$r)

arc_cor_p = get_lower_tri(arc_cor$P)
arc_p_long = reshape2::melt(arc_cor_p)
arc_r_long = reshape2::melt(arc_cor_r)
arc_r_long$p = arc_p_long$value
arc_r_long$value[is.na(arc_r_long$value)] = 0

# Matrice de corrélation de la relation entre bactéries et eucaryotes
bac_euka_cor = rcorr(as.matrix(t(bac_euka)), type = "spearman")
bac_euka_cor_r = bac_euka_cor$r
bac_euka_cor_p = bac_euka_cor$P
bac_euka_cor_n = bac_euka_cor$n

bac_euka_r <- as.matrix(bac_euka_cor_r[16:30,1:15])
bac_euka_p = as.matrix(bac_euka_cor_p[16:30,1:15])

bac_euka_p_long <- reshape2::melt(bac_euka_p)
bac_euka_r_long <- reshape2::melt(bac_euka_r)
bac_euka_r_long$p <- bac_euka_p_long$value

# Matrice de corrélation de la relation entre bactéries et archées
bac_arc_cor = rcorr(as.matrix(t(bac_arc)), type = "spearman")
bac_arc_cor_r = bac_arc_cor$r
bac_arc_cor_p = bac_arc_cor$P
bac_arc_cor_n = bac_arc_cor$n

bac_arc_r <- as.matrix((bac_arc_cor_r[16:30,1:15]))
bac_arc_p = as.matrix(bac_arc_cor_p[16:30,1:15])

bac_arc_p_long <- reshape2::melt(bac_arc_p)
bac_arc_r_long <- reshape2::melt(bac_arc_r)
bac_arc_r_long$p <- bac_arc_p_long$value

# Matrice de corrélation de la relation entre les eucaryotes et les archées
euka_arc_cor = rcorr(as.matrix(t(euka_arc)), type = "spearman")
euka_arc_cor_r = euka_arc_cor$r
euka_arc_cor_p = euka_arc_cor$P
euka_arc_cor_n = euka_arc_cor$n

euka_arc_r <- as.matrix((euka_arc_cor_r[16:30,1:15]))
```

```
euka_arc_p = as.matrix(euka_arc_cor_p[16:30,1:15])
euka_arc_p_long <- reshape2::melt(euka_arc_p)
euka_arc_r_long <- reshape2::melt(euka_arc_r)
euka_arc_r_long$p <- euka_arc_p_long$value
```

6.5 Plot the Heatmaps

6.5.1 Correlation within domain

```
# Graphique pour les corrélations entre les bactéries
stars_bac = cut(bac_r_long$p, breaks=c(-Inf, 0.001, 0.01, 0.05, Inf), label=c("***", "**", "*", ""))
graph_bac = ggplot(aes(x=Var1, y=Var2, fill=value), data=bac_r_long)

fig_bac_S1 = graph_bac + geom_tile() + scale_fill_gradient2(low="mediumblue", mid="white", high="orangered", limits=c(-1, 1), breaks=seq(-1,1,by = 0.50)) +  geom_text(aes(label=stars_bac), color="black", size=5) +  labs(y=NULL, x=NULL, fill="rho") + theme_bw() + theme(axis.text.x=element_text(angle = -45, hjust = 0))
```

```
# Graphique pour les corrélations entre les eucaryotes
stars_euka = cut(euka_r_long$p, breaks=c(-Inf, 0.001, 0.01, 0.05, Inf), label=c("***", "**", "*", ""))

graph_euka = ggplot(aes(x=Var1, y=Var2, fill=value,), data=euka_r_long)

fig_euka_S1 = graph_euka + geom_tile() + scale_fill_gradient2(low="mediumblue", mid="white", high="orangered", limits=c(-1, 1), breaks=seq(-1,1,by = 0.50),) + geom_text(aes(label=stars_euka), color="black", size=5) + labs(y=NULL, x=NULL, fill="rho") + theme_bw() +  theme(axis.text.x=element_text(angle = -45, hjust = 0))
```

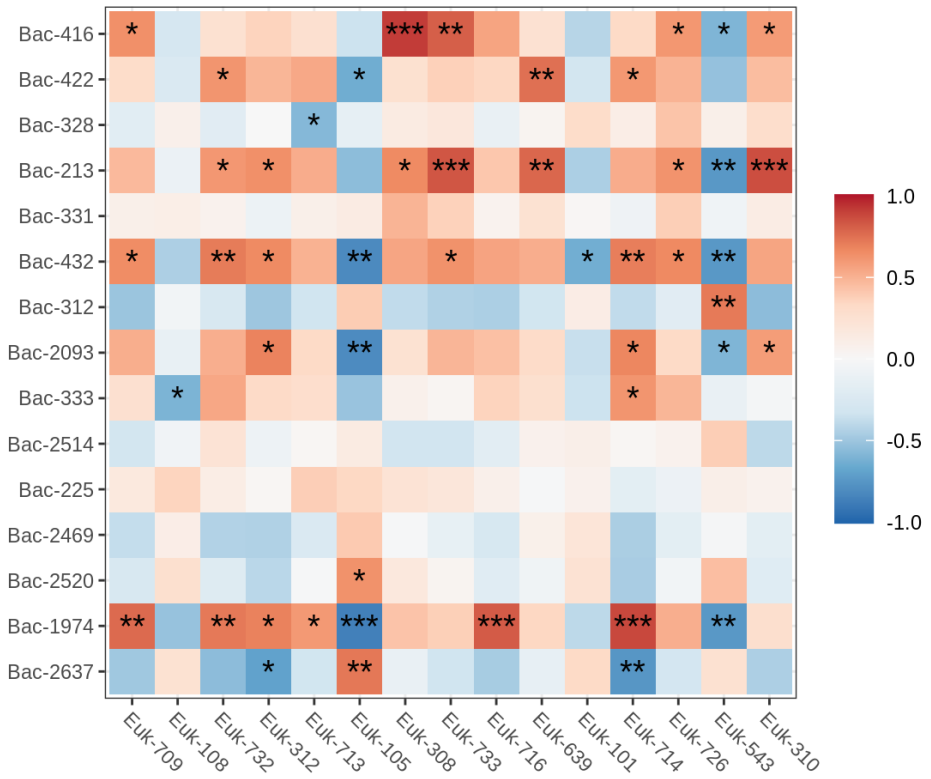
```
# Graphique pour les corrélations entre les archées
stars_arc = cut(arc_r_long$p, breaks=c(-Inf, 0.001, 0.01, 0.05, Inf), label=c("***", "**", "*", ""))
graph_arc = ggplot(aes(x=Var1, y=Var2, fill=value), data=arc_r_long)

fig_arc_S1 = graph_arc + geom_tile() + scale_fill_gradient2(low="mediumblue", mid="white", high="orangered", limits=c(-1, 1), breaks=seq(-1,1,by = 0.50)) + geom_text(aes(label=stars_arc), color="black", size=5) + labs(y=NULL, x=NULL, fill="rho") +  theme_bw() +  theme(axis.text.x=element_text(angle = -45, hjust = 0))
```

6.5.2 Correlation between domains

6.5.2.1 Heatmaps of interactions beetween Bacteria and Eukaryotes

```
# Graphique pour les corrélations entre bactéries et eucaryotes
```

```
stars_bac_euka = cut(bac_euka_r_long$p, breaks=c(-Inf, 0.001, 0.01, 0.05, Inf), label=c("***", "**", "*", "
"))
graph_bac_euka = ggplot(aes(x=Var1, y=Var2, fill=value), data=bac_euka_r_long)
fig_bac_euka_S1 = graph_bac_euka + geom_tile() +  scale_fill_distiller(palette = "RdBu",limits = c(-1,1),
guide = guide_colourbar(nbin=100, draw.ulim = FALSE, draw.llim = FALSE, barheight = 10)) +  geom_text
(aes(label=stars_bac_euka), color="black", size=5) + labs(y=NULL, x=NULL, fill="rho") +  theme_bw() + the
me(axis.text.x=element_text(angle = -45, hjust = 0),legend.title = element_blank())+ coord_fixed() + xlab
("(a)")
fig_bac_euka_S1
```



(a)

6.5.2.2 Heatmaps of interactions beetween Bacteria and Archaea

```
#Graphique pour les corrélations entre bactéries et archées
stars_bac_arc = cut(bac_arc_r_long$p, breaks=c(-Inf, 0.001, 0.01, 0.05, Inf), label=c("***", "**", "*", ""))
graph_bac_arc = ggplot(aes(x=Var1, y=Var2, fill=value), data=bac_arc_r_long)
fig_bac_arc_S1 = graph_bac_arc + geom_tile() + scale_fill_distiller(palette = "RdBu",limits = c(-1,1), guide
= guide_colourbar(nbin=100, draw.ulim = FALSE, draw.llim = FALSE, barheight = 10)) + geom_text(aes(lab
el=stars_bac_arc), color="black", size=5) + labs(y=NULL, x=NULL, fill="rho") +  theme_bw() + theme(axis.t
ext.x=element_text(angle = -45, hjust = 0), legend.title = element_blank())+ coord_fixed() + xlab("(b)")
```

6.5.2.2 Heatmaps of interactions beetween Eukaryotes and Archaea

```
#Graphique pour les corrélations entre eucaryotes et archées
```

```
stars_euka_arc = cut(euka_arc_r_long$p, breaks=c(-Inf, 0.001, 0.01, 0.05, Inf), label=c("***", "**", "*", "
"))
graph_euka_arc = ggplot(aes(x=Var1, y=Var2, fill=value), data=euka_arc_r_long)
fig_euka_arc_S1 = graph_euka_arc + geom_tile() + scale_fill_distiller(palette = "RdBu",limits = c(-1,1), gui
de = guide_colourbar(nbin=100, draw.ulim = FALSE, draw.llim = FALSE, barheight = 10))+  labs(y=NULL, x=
NULL, fill="rho") +  theme_bw() + theme(axis.text.x=element_text(angle = -45, hjust = 0),legend.title = ele
ment_blank())+ coord_fixed() + xlab("(c)")
fig_bac_euka_S1
```

6.6 Combine plots

```
library("gridExtra")
library("cowplot")
library("gplot" )


# Once you are done and you repeat the code for site 2 use the following code to combine plots :

# plot_grid(fig_bac_euka_S1,fig_bac_arc_S1,fig_euka_arc_S1, fig_bac_euka_S2,fig_bac_arc_S2,fig_euka
_arc_S2 common.legend = TRUE, legend = "right", ncol = 3, nrow =
```

7. FEAST Analysis

This is an example with the Feast analysis for Bacteria at site 1 and for the third sample in site 1. In this example the third sample sources are the surface sample and the second samples. The metadata looks like this :

	Env	SourceSink	id
RR.03.bac	2.25	Sink	1
RR.02.bac	0.75	Source	NA
RR.01.bac	0.00	Source	NA

ID 1 is used for the sink and NA for the sources. You have to repeat the script many times for each sample. For Bacteria at site 1, you should have 19 metadata.

7.1 Calculate the source contribution for RR.03 (sink)

```
rm(list = ls())
gc()
source("src.r") # src.r has to be in the same directory (see section 7.2 for more details)
EM_iterations = 1000 # number of EM iterations. default value

# Load sample metadata, doit être en format csv
metadata <- read.csv(file = "metadata.bac.S1.1.csv" ,header = T, sep = ";", row.names = 1) #Change with
metadata 2 to 19

# Load normalised ASV  table (has to be in csv format)
otus <- read.table(file = "count_matrix_bac_S1.csv", header = T, comment = '', check = F, sep = ';', row.na
mes = 1)
otus <- t(as.matrix(otus))

# Extract only those samples in common between the two tables
common.sample.ids <- intersect(rownames(metadata), rownames(otus))
otus <- otus[common.sample.ids,]
metadata <- metadata[common.sample.ids,]
# Double-check that the mapping file and otu table
# had overlapping samples
if(length(common.sample.ids) <= 1) {
 message <- paste(sprintf('Error: there are %d sample ids in common '),
                  'between the metadata file and data table')
 stop(message)}

envs <- metadata$Env

# Extract the source environments and source/sink indices
```

```
train.ix <- which(metadata$SourceSink=='Source')
test.ix <- which(metadata$SourceSink=='Sink')
num_sources <- length(train.ix) #number of sources
COVERAGE =  min(rowSums(otus[c(train.ix, test.ix),]))  # Can be adjusted by the user

# Define sources and sinks
sources <- as.matrix(rarefy(otus[train.ix,], COVERAGE))
sinks <- as.matrix(rarefy(t(as.matrix(otus[test.ix,])), COVERAGE))
print(paste("Number of OTUs in the sink sample = ",length(which(sinks > 0))))
print(paste("Seq depth in the sources and sink samples = ",COVERAGE))
print(paste("The sink is:", envs[test.ix]))

# Estimate source proportions for each sink
FEAST_output<-FEAST(source=sources, sinks = t(sinks), env = envs[train.ix], em_itr = EM_iterations, COV
ERAGE = COVERAGE)
Proportions_est <- FEAST_output$data_prop[,1]
names(Proportions_est) <- c(as.character(envs[train.ix]), "unknown")

print("Source mixing proportions")
## [1] "Source mixing proportions"
Proportions_est
##      0.75         0    unknown
## 0.17236434 0.07091567 0.75671999
```

Here for example, the surface sample (Depth : 0 m) contribute in 7 % in the formation of the bacterial community found in the third sample (Depth : 2.25 m).

7.2 scr.r

scr.r is a code needed in order for the 7.1 script to work.

```
x<-c("vegan", "dplyr", "ggrepel", "doParallel", "foreach" ,"mgcv", "reshape2", "ggplot2", "Rcpp", "RcppAr
madillo")
lapply(x, require, character.only = TRUE)
library("vegan")
library("dplyr")
library("doParallel")
library("foreach")
library("mgcv")
library("reshape2")
library("ggplot2")
library("cowplot")
library("Rcpp")
library("RcppArmadillo")
cppFunction("arma::mat schur(arma::mat& a, arma::mat& b)
        {return(a % b); }", depends="RcppArmadillo")
```

```
change_C<-function(newcov, X){

 X=t(as.matrix(X))
 idx = 1:dim(X)[2]

 if(sum(X) > newcov){

  while(sum(X) > newcov){
   greaterone = X > 1
   samps = 20
   if(samps > length(X[greaterone]))
    samps = length(X[greaterone])
   changeidx = sample(idx[greaterone], samps, replace = F)
   X[changeidx] = X[changeidx] - 1
  }

 }

 if(sum(X) < newcov){

  while(sum(X) < newcov){
   greaterone = X > 1
   samps = 100
   if(samps > length(X[greaterone]))
    samps = length(X[greaterone])
   changeidx = sample(idx[greaterone], samps, replace = F)
   X[changeidx] = X[changeidx] + 1
  }

 }

 return(X)
}

rarefy <- function(x,maxdepth){


 if(is.null(maxdepth)) return(x)

 if(!is.element(class(x), c('matrix', 'data.frame','array')))
  x <- matrix(x,nrow=nrow(x))
 nr <- nrow(x)
 nc <- ncol(x)

 for(i in 1:nrow(x)){
  if(sum(x[i,]) > maxdepth){
```

```
    prev.warn <- options()$warn
    options(warn=-1)
    s <- sample(nc, size=maxdepth, prob=x[i,], replace=T)
    options(warn=prev.warn)
    x[i,] <- hist(s,breaks=seq(.5,nc+.5,1), plot=FALSE)$counts
   }
  }
  return(x)
}

jsdmatrix <- function(x){
  d <- matrix(0,nrow=nrow(x),ncol=nrow(x))
  for(i in 1:(nrow(x)-1)){
    for(j in (i+1):nrow(x)){
      d[i,j] <- jsd(x[i,], x[j,])
      d[j,i] <- d[i,j]
    }
  }
  return(d)
}

jsd <- function(p,q){
  m <- (p + q)/2
  return((kld(p,m) + kld(q,m))/2)
}

kld <- function(p,q){
  nonzero <- p>0 & q>0
  return(sum(p[nonzero] * log2(p[nonzero]/q[nonzero])))
}

h<-function(x) {y <- x[x > 0]; -sum(y * log(y))};
mult_JSD <- function(p,q) {h(q %*% p) - q %*% apply(p, 1, h)}

retrands<-function(V){
  toret<-unlist(lapply(c(V), function(x) runif(1, x+1e-12, x+1e-09)))
  return(toret)
}

getR2<-function(x,y){
  return((cor(x,y))^2)
}

E<-function(alphas, sources){
  nums<-(sapply(1:length(alphas), function(n) Reduce("+", crossprod(as.numeric(alphas[n]),as.numeric(s
ources[[n]])))))
  denom<-(Reduce("+", nums))
```

```
 return(nums/denom)
}

A<-function(alph, XO, raos){
 tmp<-crossprod(alph, XO/raos)
 tmp<-rapply(list(tmp), f=function(x) ifelse(is.nan(x),0,x), how="replace" )
 tmp<-Reduce("+",unlist(tmp))
 return(tmp)
}

M<-function(alphas, sources, sink, observed){

 newalphs<-c()
 rel_sink <-sink/sum(sink)

 if(sum(sources[[1]]) > 1){

  sources <-lapply(sources, function(x) x/(sum(colSums(x))))
 }

 LOs<-lapply(sources, schur, b=rel_sink)
 BOs<-t(mapply(crossprod, x=sources, y=alphas))
 BOs<-split(BOs, seq(nrow(BOs)))
 BOs<-lapply(BOs, as.matrix)
 BOs<-lapply(BOs, t)
 num_list <- list()
 source_new <- list()


 for(i in 1:length(sources)){
  num <- c()
  denom <- c()
  num<-crossprod(alphas[i], (LOs[[i]]/(Reduce("+", BOs))))
  num<-rapply(list(num), f=function(x) ifelse(is.nan(x),0,x), how="replace" ) #replace na with zero
  num_list[[i]]<- num[[1]][1,] + observed[[i]][1,]

  denom <- Reduce("+",unlist(num_list[[i]]))
  source_new[[i]] <- num_list[[i]]/denom
  source_new[[i]][is.na(source_new[[i]])] = 0
 }

 sources = source_new

 newalphs<-c()
#sink<-as.matrix(sink); #src1<-as.matrix(sources[[1]]); src2<-as.matrix(sources[[2]])
 sources<-lapply(sources, t)
 XOs<-lapply(sources,schur, b=rel_sink)
```

```
  AOs<-t(mapply(crossprod, x=sources, y=alphas))
  AOs<-split(AOs, seq(nrow(AOs)))
  AOs<-lapply(AOs, as.matrix)
  AOs<-lapply(AOs, t)
  newAs<-c()
  for(i in 1:length(sources)){
    newA<-crossprod(alphas[i], (XOs[[i]]/(Reduce("+", AOs))))
    newA<-rapply(list(newA), f=function(x) ifelse(is.nan(x),0,x), how="replace" )
    newA<-Reduce("+",unlist(newA))
    newAs<-c(newAs, newA)
  }
  tot<-sum(newAs)
  Results <- list (new_alpha = newAs/(tot), new_sources = sources)
  return(Results) }

do_EM<-function(alphas, sources, observed, sink, iterations){

  curalphas<-alphas
  newalphas<-alphas
  m_guesses<-c(alphas[1])
  for(itr in 1:iterations){

    curalphas<-E(newalphas, sources)
    tmp <- M(alphas = curalphas, sources = sources, sink = sink, observed = observed)
    newalphas <- tmp$new_alpha
    sources <- tmp$new_sources

    m_guesses<-c(m_guesses, newalphas[1])
    if(abs(m_guesses[length(m_guesses)]-m_guesses[length(m_guesses)-1])<=10^-6) break

  }
  toret<-c(newalphas)
  results <- list(toret = toret, sources = sources)

  return(results)
}

M_basic <-function(alphas, sources, sink){
  newalphs<-c()
  XOs<-lapply(sources,schur, b=sink)
  AOs<-t(mapply(crossprod, x=sources, y=alphas))
  AOs<-split(AOs, seq(nrow(AOs)))
  AOs<-lapply(AOs, as.matrix)
  AOs<-lapply(AOs, t)
  newAs<-c()
  for(i in 1:length(sources)){
    newA<-crossprod(alphas[i], (XOs[[i]]/(Reduce("+", AOs))))
```

```
   newA<-rapply(list(newA), f=function(x) ifelse(is.nan(x),0,x), how="replace" )
   newA<-Reduce("+",unlist(newA))
   newAs<-c(newAs, newA)
  }
  tot<-sum(newAs)
  return(newAs/(tot))
}

do_EM_basic<-function(alphas, sources, sink, iterations){
  curalphas<-alphas
  newalphas<-alphas
  m_guesses<-c(alphas[1])
  for(itr in 1:iterations){
   curalphas<-E(newalphas, sources)
   newalphas<-M_basic(curalphas, sources, sink)
   m_guesses<-c(m_guesses, newalphas[1])

   if(abs(m_guesses[length(m_guesses)]-m_guesses[length(m_guesses)-1])<=10^-6) break
  }
  toret<-c(newalphas)
  return(toret)
}

source_process_nounknown <- function(train, envs, rarefaction_depth=1000){

  train <- as.matrix(train)

  # enforce integer data
  if(sum(as.integer(train) != as.numeric(train)) > 0){
   stop('Data must be integral. Consider using "ceiling(datatable)" or ceiling(1000*datatable) to convert fl
oating-point data to integers.')
  }
  envs <- factor(envs)
  train.envs <- sort(unique(levels(envs)))

  # rarefy samples above maxdepth if requested
  if(!is.null(rarefaction_depth) && rarefaction_depth > 0) train <- rarefy(train, rarefaction_depth)

  # get source environment counts
  # sources is nenvs X ntaxa
  X <- t(sapply(split(data.frame(train), envs), colSums))

  rownames(X) <- c(train.envs)
  X <- t(as.matrix(X))

  return(X)
}
```

```
read_pseudo_data <-function(dataset){
 path_to_data<-"../data/"
 if(dataset=="DA"){
   df<-read.table(paste0(path_to_data,"DA_99_T_d10000_date_nan.txt"), fill = NA)
   return(df[complete.cases(df),])
 }else if(dataset=="DB"){
   df<-read.table(paste0(path_to_data,"DB_99_T_d10000_date_nan.txt"), fill = NA)
   return(df[complete.cases(df),])
 }else if (dataset=="F4"){
   df<-read.table(paste0(path_to_data,"F4_99_T_d10000_date_nan.txt"), fill = NA)
   return(df[complete.cases(df),])
 }else{
   df<-read.table(paste0(path_to_data,"M3_99_T_d10000_date_nan.txt"), fill = NA)
   return(df[complete.cases(df),])}
}

create_m <- function(num_sources, n, EPSILON){


 if( n == 1 ){

   index = sample(c(1:num_sources), 1)
   m_1 = runif(min = 0.6, max = 0.9, n = 1)
   resid = 1-m_1
   other_ms = resid/(num_sources-1)
   m = rep(NA, num_sources)
   m[index] = c(m_1)
   m[is.na(m)] = other_ms }

 if( n == 2 ){

   index = sample(c(1:num_sources), 2)
   m_1 = runif(min = 0.1, max = 0.2, n = 1)
   m_2 = runif(min = 0.4, max = 0.5, n = 1)
   resid = 1-(m_1+m_2)
   other_ms = resid/(num_sources-2)
   m = rep(NA, num_sources)
   m[index] = c(m_1, m_2)
   m[is.na(m)] = other_ms }

 if( n == 3 ){

   index = sample(c(1:num_sources), 3)
   m_1 = runif(min = 0.1, max = 0.5, n = 1)
   m_2 = runif(min = 0.2, max = 0.25, n = 1)
   m_3 = runif(min = 0.1, max = 0.15, n = 1)
```

```
    resid = 1-(m_1+m_2+m_3)
    other_ms = runif(min = 0.001, max = resid/(num_sources-3), n = (num_sources-3))
    m = rep(NA, num_sources)
    m[index] = c(m_1, m_2, m_3)
    m[is.na(m)] = other_ms
    m = m/sum(m)

  }
  subsum = 0
  idx = 1:length(m)

  while ((subsum+0.001) < EPSILON){
    tosub = EPSILON - subsum
    tosub = tosub / (num_sources+1)
    mask = m > tosub
    m[mask] = m[mask] - tosub
    subsum = subsum + length(m[mask]) * tosub

  }
  m = c(m,(EPSILON))

  # sum(m)
  return(m)

}

unknown_initialize <- function(sources, sink, n_sources){

  unknown_source = rep(0, length(sink))
  sum_sources = apply(sources, 2, sum)

  unknown_source = c()

  for(j in 1:length(sum_sources)){

    unknown_source[j] = max(sink[j]-sum_sources[j], 0)

  }


  return(unknown_source)

}

unknown_initialize_1 <- function(sources, sink, n_sources){

  unknown_source = rep(0, length(sink))
```

```
 sources_sum = apply(sources, 2 ,sum)

 unknown_source = c()

 for(j in 1:length(sources_sum)){

  unknown_source[j] = max(sink[j]-sources_sum[j], 0) }

 #Select the cor OTUs
 ind_cor = list()
 ind_known_source_abun = c()
 ind_cor_all = which(sources[1,] > 0)

 counter = matrix(0, ncol = dim(sources)[2], nrow =  dim(sources)[1])

 for(j in 1:n_sources){

  ind_cor[[j]] = which(sources[j,] > 0)

  for(k in 1:length(sources[j,])){

   if(sources[j,k] > 0){

    counter[j,k] = counter[j,k]+1 }   } }

 OTU_present_absent = apply(counter, 2, sum)
 ind_cor_all = which(OTU_present_absent >= round(n_sources*0.8))

 if(length(ind_cor_all) > 1){

  cor_abundance = round(apply(sources[,ind_cor_all], 2, median)/2) #take the min abundnace of the 'cor
'
  unknown_source[ind_cor_all] = cor_abundance }


 #keep the sink abundance where there is no known source
 ind_no_known_source_abun = which(sources_sum == 0)

 for(j in 1:length(ind_no_known_source_abun)){

  # unknown_source[ind_no_known_source_abun[j]] = max(runif(n = 1, min = 1, max = 100), sink[ind_no
_known_source_abun[j]])
  unknown_source[ind_no_known_source_abun[j]] = max((sink[ind_no_known_source_abun[j]] - rpois(
n = 1, lambda = 0.5)), 0)  }

 return(unknown_source)}
```

```
unknown__initialize_1 <- function(sources, sink, n_sources){

 unknown_source = rep(0, length(sink))

 #zero all the OTUs with at least 1 known source
 sources_sum = apply(sources, 2 ,sum)
 ind_known_source_abun = which(sources_sum > 0)
 unknown_source[ind_known_source_abun] = 0

 #Select the cor OTUs
 ind_cor = list()
 ind_known_source_abun = c()
 ind_cor_all = which(sources[1,] > 0)

 counter = matrix(0, ncol = dim(sources)[2], nrow =  dim(sources)[1])

 for(j in 1:n_sources){

  ind_cor[[j]] = which(sources[j,] > 0)

  for(k in 1:length(sources[j,])){

   if(sources[j,k] > 0){

    counter[j,k] = counter[j,k]+1 }   } }

 OTU_present_absent = apply(counter, 2, sum)
 ind_cor_all = which(OTU_present_absent >= round(n_sources*0.8))

 if(length(ind_cor_all) > 1){   cor_abundance = apply(sources[,ind_cor_all], 2, median) #take the median
abundnace of the 'cor'
  unknown_source[ind_cor_all] = cor_abundance}

 #keep the sink abundance where there is no known source
 ind_no_known_source_abun = which(sources_sum == 0)

 for(j in 1:length(ind_no_known_source_abun)){
unknown_source[ind_no_known_source_abun[j]] = max( round(sink[ind_no_known_source_abun[j]]+ rn
orm(n = length(sink[ind_no_known_source_abun[j]])), 0) }

 return(unknown_source)}

FEAST <- function(source = sources_data, sinks = sinks, em_itr = 1000, env = rownames(sources_data), in
clude_epsilon = T, COVERAGE, unknown_initialize_flag = 0){

 tmp = source
 test_zeros = apply(tmp, 1, sum)
```

```
 ind_to_use = as.numeric(which(test_zeros > 0))
 ind_zero = as.numeric(which(test_zeros == 0))

 source = tmp[ind_to_use,]
 sinks = sinks

 #####adding support for multiple sources#####
 totalsource<-source
 totalsource<-as.matrix(totalsource)
 sources <- split(totalsource, seq(nrow(totalsource)))
 sources<-lapply(sources, as.matrix)
 dists<-lapply(sources, function(x) x/(sum(colSums(x))))
 totaldist<-t(Reduce("cbind", dists))
 sinks<-matrix(sinks, nrow = 1, ncol = dim(totalsource)[2])

 num_sources = dim(source)[1]
 envs_simulation = c(1:(num_sources))

 source_old = source
 totalsource_old = totalsource

 source_old=lapply(source_old,t)
 source_old<- split(totalsource_old, seq(nrow(totalsource_old)))
 source_old<-lapply(source_old, as.matrix)

 #Creating the unknown source per mixing iteration
 if(include_epsilon == TRUE){

   ##Adding the initial value of the unknown source for CLS and EM
   source_2 = list()
   totalsource_2 = matrix(NA, ncol = dim(totalsource_old)[2], nrow = ( dim(totalsource_old)[1] + 1))

   for(j in 1:num_sources){source_2[[j]] = source_old[[j]]
     totalsource_2[j,] = totalsource_old[j,] }

   #create unknown for each sink i
   sinks_rarefy = rarefy(matrix(sinks, nrow = 1), maxdepth = apply(totalsource_old, 1, sum)[1]) #make

   if(unknown_initialize_flag == 1)
     unknown_source = unknown_initialize_1(sources = totalsource[c(1:num_sources),], sink = as.numeric
(sinks), n_sources = num_sources)

   if(unknown_initialize_flag == 0)
     unknown_source = unknown_initialize(sources = totalsource[c(1:num_sources),], sink = as.numeric(si
nks),  n_sources = num_sources)

   # unknown_source = unknown_source_1 + rpois(n = length(sinks), lambda = 0.5)
```

```
  unknown_source_rarefy = rarefy(matrix(unknown_source, nrow = 1), maxdepth = COVERAGE)
  source_2[[j+1]] = t(unknown_source_rarefy)
  totalsource_2[(j+1),] = t(unknown_source_rarefy)
  totalsource = totalsource_2

  source=lapply(source_2,t)
  # totalsource <- rarefy(x = totalsource, maxdepth = COVERAGE)
  source<- split(totalsource, seq(nrow(totalsource_2)))
  source<-lapply(source_2, as.matrix)

   envs_simulation <- c(1:(num_sources+1)) }

 samps <- source
 samps<-lapply(samps, t)

 observed_samps <- samps
 observed_samps[[(num_sources + 1)]] = t(rep(0, dim(samps[[1]])[2]))

 #Calculate JSD value
 # x <- totalsource[c(1:num_sources),]
 # JSDMatrix <- jsdmatrix(x)
 # JSDMatrix <- JSDMatrix/COVERAGE
 # JS = mean(JSDMatrix[-which(JSDMatrix == 0)])
 # js_values = append(js_values, JS)
 # print(js_values)

 initalphs<-runif(num_sources+1, 0.0, 1.0)
 initalphs=initalphs/Reduce("+", initalphs)
 sink_em = as.matrix(sinks)
 pred_em<-do_EM_basic(alphas=initalphs, sources=samps, sink=sink_em, iterations=em_itr)

 tmp<-do_EM(alphas=initalphs, sources=samps, sink=sink_em, iterations=em_itr, observed=observed_sa
mps)
 pred_emnoise = tmp$toret

 k = 1
 pred_emnoise_all = c()
 pred_em_all = c()

 for(j in 1:length(env)){

  if(j %in% ind_to_use){

    pred_emnoise_all[j] = pred_emnoise[k]
    pred_em_all[j] = pred_em[k]
    k = k+1 }
```

```
 else{

   pred_emnoise_all[j] = 0
   pred_em_all[j] = 0
 } }

pred_emnoise_all[j+1] = pred_emnoise[k]
pred_em_all[j+1] = pred_em[k]


names(pred_emnoise_all) = c(env,"unknown")
names(pred_em_all) = c(env,"unknown")

Results = list(unknown_source = unknown_source, unknown_source_rarefy = unknown_source_rarefy,
           data_prop = data.frame(pred_emnoise_all,pred_em_all))
return(Results)
```

RÉFÉRENCES

Abu-Ashour, J.; Joy, D.M.; Lee, H.; Whiteley, H.R.; Zelin, S. Transport of microorganisms through soil. *Water Air Soil Pollut*. 1994, 75, 141–158.

Agnelli, A.; Ascher, J.; Corti, G.; Ceccherini, M. T.; Nannipieri, P.; Pietramellara, G. Distribution of microbial communities in a forest soil profile investigated by microbial biomass, soil respiration and DGGE of total and extracellular DNA. *Soil Biology and Biochemistry*. 2004, *36*(5), 859-868.

Akob, D. M.; Küsel, K. Where microorganisms meet rocks in the Earth's Critical Zone. *Biogeosciences*. 2011, *8*(12), 3531-3543.

Alain, K.; Callac, N.; Ciobanu, M. C.; Reynaud, Y.; Duthoit, F.; Jebbar, M. DNA extractions from deep subseafloor sediments: novel cryogenic-mill-based procedure and comparison to existing protocols. *Journal of Microbiological Methods*. 2011, *87*(3), 355-362.

Amy, P. S.; Haldeman, D. L.; Ringelberg, D.; Hall, D. H.; Russell, C. Comparison of identification systems for classification of bacteria isolated from water and endolithic habitats within the deep subsurface. *Applied and Environmental Microbiology*. 1992, *58*(10), 3367-3373.

Anacker, B. L.; Harrison, S. P. Historical and ecological controls on phylogenetic diversity in Californian plant communities. *The American Naturalist*. 2012, *180*(2), 257-269.

Andrew, D. R.; Fitak, R. R.; Munguia-Vega, A.; Racolta, A.; Martinson, V. G.; Dontsova, K. Abiotic factors shape microbial diversity in Sonoran Desert soils. *Applied and environmental microbiology*. 2012, *78*(21), 7527-7537.

Antony, C. P.; Cockell, C. S.; Shouche, Y. S. Life in (and on) the rocks. *Journal of biosciences*. 2012, *37*(1), 3-11.

Baker, G. C.; Smith, J. J.; Cowan, D. A. Review and re-analysis of domain-specific 16S primers. *Journal of microbiological methods*. 2003, *55*(3), 541-555.

Baldrian, P.; Kolařík, M.; Štursová, M.; Kopecký, J.; Valášková, V.; Větrovský, T.; Žifčáková, L.; Šnajdr, J.; Rídl, J.; Vlček, C.; Voříšková, J. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME journal*. 2012, *6*(2), 248-258.

Bennett, P. C., Rogers, J. R., Choi, W. J., Hiebert, F. K. Silicates, silicate weathering, and microbial ecology. *Geomicrobiology Journal*. 2001,18(1), 3-19.

Beveridge, T. J., Makin, S. A., Kadurugamuwa, J. L., Li, Z. Interactions between biofilms and the environment. *FEMS Microbiology reviews*. 1997, *20*(3-4), 291-303.

Bomberg, M.; Ahonen, L. Geomicrobes: life in terrestrial deep subsurface. *Frontiers in microbiology*. 2017, *8*, 103.

Borgonie, G.; García-Moyano, A.; Litthauer, D.; Bert, W.; Bester, A.; van Heerden, E.; Moller, C.; Erasmus, M.; Onstott, T.C. Nematoda from the terrestrial deep subsurface of South Africa. *Nature*. 2011, 474, 79–82.

Bräuer, S. L.; Cadillo-Quiroz, H.; Ward, R. J.; Yavitt, J. B.; Zinder, S. H. Methanoregula boonei gen. nov., sp. nov., an acidiphilic methanogen isolated from an acidic peat bog. *International journal of systematic and evolutionary microbiology*. 2011, *61*(1), 45-52.

Butler, B.; Hillier, S. powdR: An R package for quantitative mineralogy using full pattern summation of X-ray powder diffraction data. *Computers & Geosciences*. 2021, 147, 104662.

Callahan, B. J.; Sankaran, K.; Fukuyama, J. A.; McMurdie, P. J.; Holmes, S. P. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*. 2016, *5*.

Cao, P.; Zhang, L. M.; Shen, J. P.; Zheng, Y. M.; Di, H. J.; He, J. Z. Distribution and diversity of archaeal communities in selected Chinese soils. *FEMS Microbiology Ecology*. 2012, *80*(1), 146-158.

Caruso, T.; Schaefer, I.; Monson, F.; Keith, A. M. Oribatid mites show how climate and latitudinal gradients in organic matter can drive large-scale biodiversity patterns of soil communities. *Journal of Biogeography*. 2019, *46*(3), 611-620.

Cavalier-Smith, T.; Chao, E. E.; Thompson, C. E.; Hourihane, S. L. Oikomonas, a distinctive zooflagellate related to chrysomonads. *Archiv für Protistenkunde*. 1996, *146*(3-4), 273-279.

Chen, H.; Yang, Z. K.; Yip, D.; Morris, R. H.; Lebreux, S. J.; Cregger, M. A.; Klingeman, D. M.; Hui, D.; Hettich, R. L.; Wilhelm, S. W.; Wang, G.; Löffler, F. E.; Schadt, C. W. One-time nitrogen fertilization shifts switchgrass soil microbiomes within a context of larger spatial and temporal variation. *Plos one*. 2019, *14*(6), e0211310.

Chu, H.; Sun, H.; Tripathi, B. M.; Adams, J. M.; Huang, R.; Zhang, Y.; Shi, Y. Bacterial community dissimilarity between the surface and subsurface soils equals horizontal differences over several kilometers in the western Tibetan Plateau. *Environmental Microbiology*. 2016, *18*(5), 1523-1533.

Colman, D. R.; Poudel, S.; Stamps, B. W.; Boyd, E. S.; Spear, J. R. The deep, hot biosphere: Twenty-five years of retrospection. *Proceedings of the National Academy of Sciences*. 2017, *114*(27), 6895-6903.

DeLong, E. F. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences*. 1992, *89*(12), 5685-5689.

Direito, S. O.; Marees, A.; Röling, W. F. Sensitive life detection strategies for low-biomass environments: optimizing extraction of nucleic acids adsorbing to terrestrial and Mars analogue minerals. *FEMS microbiology ecology*. 2012, *81*(1), 111-123.

Eckert, D.; Sims, J.T. Recommended soil pH and lime requirement tests. In Recommended Soil Testing Procedures for the Northeastern United States. *Northeast Regional Bulletin*. 1995; *493*, 11-16.

Edwards, K. J.; Becker, K.; Colwell, F.The deep, dark energy biosphere: intraterrestrial life on earth. *Annual review of earth and planetary sciences*. 2012, *40*, 551-568.

Eilers, K. G.; Debenport, S.; Anderson, S.; Fierer, N. Digging deeper to find unique microbial communities: the strong effect of depth on the structure of bacterial and archaeal communities in soil. *Soil Biology and Biochemistry*. 2012, *50*, 58-65.

Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biological conservation*. 1992, *61*(1), 1-10.

Feng, H.; Guo, J.; Wang, W.; Song, X.; Yu, S. Soil depth determines the composition and diversity of bacterial and archaeal communities in a poplar plantation. Forests. 2019, 10(7), 550.

Fierer, N.; Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences.* 2006, *103*(3), 626-631.

Fierer, N.; Schimel, J. P.; Holden, P. A. Variations in microbial community composition through two soil depth profiles. *Soil Biology and Biochemistry*. 2003, *35*(1), 167-176.

Flynn, T. M.; Sanford, R. A.; Ryu, H.; Bethke, C. M.; Levine, A. D.; Ashbolt, N. J.; Santo Domingo, J. W. Functional microbial diversity explains groundwater chemistry in a pristine aquifer. *BMC microbiology*. 2013, *13*(1), 1-15.

Gast, R. J.; Dennett, M. R.; Caron, D. A. Characterization of protistan assemblages in the Ross Sea, Antarctica, by denaturing gradient gel electrophoresis. *Applied and environmental microbiology*. 2004, *70*(4), 2028-2037.

Goberna, M.; Insam, H.; Klammer, S.; Pascual, J. A.; Sanchez, J. Microbial community structure at different depths in disturbed and undisturbed semiarid Mediterranean forest soils. *Microbial Ecology*. 2005, *50*(3), 315-326.

Goldscheider, N.; Hunkeler, D.; Rossi, P. Microbial biocenoses in pristine aquifers and an assessment of investigative methods. *Hydrogeology Journal*. 2006, *14*(6), 926-941.

Government of Canada. Canadian Climate Normals 1981–2010 Station Data. Environ. Resour. 2021a, Climate ID 7033939.

Government of Canada. Canadian Climate Normals 1981–2010 Station Data. Environ. Resour. 2021b, Climate ID 7038975.

Govil, T.; Rathinam, N. K.; Salem, D. R.; Sani, R. K. Taxonomical diversity of extremophiles in the deep biosphere. In *Microbial Diversity in the Genomic Era.* Academic Press. 2019, 631-656.

Griebler, C.; Lueders, T. Microbial biodiversity in groundwater ecosystems. *Freshwater Biology*. 2009, *54*(4), 649-677.

Griebler, C.; Mindl, B.; Slezak, D.; Geiger-Kaiser, M. Distribution patterns of attached and suspended bacteria in pristine and contaminated shallow aquifers studied with an in situ sediment exposure microcosm. *Aquatic microbial ecology*. 2002, *28*(2), 117-129.

Griffiths, R. I.; Whiteley, A. S.; O'Donnell, A. G.; Bailey, M. J. Influence of depth and sampling time on bacterial community structure in an upland grassland soil. *FEMS microbiology ecology*. 2003, *43*(1), 35-43.

Hamarashid, N.H.; Othman, M.A.; Hussain, M.A.H. Effects of soil texture on chemical compositions, microbial populations and carbon mineralization in soil. *Egypt J. Exp. Biol*. 2010, 6, 59-64.

Hansel, C. M.; Fendorf, S.; Jardine, P. M.; Francis, C. A. Changes in bacterial and archaeal community structure and functional diversity along a geochemically variable soil profile. *Applied and environmental microbiology*. 2008, *74*(5), 1620-1633.

Haroon, M. F., Hu, S., Shi, Y., Imelfort, M., Keller, J., Hugenholtz, P., ... & Tyson, G. W. (2013). Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature*, *500*(7464), 567-570.

Hartmann, M.; Lee, S.; Hallam, S. J.; Mohn, W. W. Bacterial, archaeal and eukaryal community structures throughout soil horizons of harvested and naturally disturbed forest stands. *Environmental Microbiology*. 2009, *11*(12), 3045-3062.

Herrmann, M.; Wegner, C. E.; Taubert, M.; Geesink, P.; Lehmann, K.; Yan, L.; Lehmann, R.; Totsche, K. U.; Küsel, K. Predominance of Cand. Patescibacteria in groundwater is caused by their preferential mobilization from soils and flourishing under oligotrophic conditions. Frontiers in microbiology. 2019, 10, 1407.

Ibelings, B. W.; De Bruin, A.; Kagami, M.; Rijkeboer, M.; Brehm, M.; Donk, E. V. Host parasite interactions between freshwater phytoplankton and chytrid fungi (chytridiomycota) 1. *Journal of Phycology. 2004, 40*(3), 437-453.

Jin, Q.; Kirk, M. F. pH as a primary control in environmental microbiology: 1. thermodynamic perspective. *Frontiers in Environmental Science*. 2018, *6*, 21.

Jones, R. T.; Robeson, M. S.; Lauber, C. L.; Hamady, M.; Knight, R.; Fierer, N. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. The ISME journal. 2009, *3*(4), 442-453.

Kang, B.; Bowatte, S.; Hou, F. Soil microbial communities and their relationships to soil properties at different depths in an alpine meadow and desert grassland in the Qilian mountain range of China. *Journal of Arid Environments*. 2021, *184*, 104316.

Kembel, S. W.; Cowan, P. D.; Helmus, M. R.; Cornwell, W. K.; Morlon, H.; Ackerly, D. D.; Blomberg, S.P.; Webb, C. O. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*. 2010, *26*(11), 1463-1464.

Kembel, S. W.; Eisen, J. A.; Pollard, K. S.; Green, J. L. The phylogenetic diversity of metagenomes. *PloS one*. 2011, *6*(8), e23214.

Kieft, T. L.; Murphy, E. M.; Haldeman, D. L.; Amy, P. S.; Bjornstad, B. N.; McDonald, E. V.; Ringelberg, D.B.; White, D.C.; Stair, J.; Griffiths, R. P.; Gsell, T. C.; Holben, W.E.; Boone, D. R. Microbial transport, survival, and succession in a sequence of buried sediments. Microbial ecology. 1998, *36*(3-4), 336-348.

Kim, H. M.; Lee, M. J.; Jung, J. Y.; Hwang, C. Y.; Kim, M.; Ro, H. M.; Chun, J.; Lee, Y. K. Vertical distribution of bacterial community is associated with the degree of soil organic matter decomposition in the active layer of moist acidic tundra. *Journal of Microbiology*. 2016, *54*(11), 713-723

Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glöckner, F. O. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*. 2013, *41*(1), e1-e1.

Kneip, C.; Lockhart, P.; Voß, C.; Maier, U. G. Nitrogen fixation in eukaryotes–new models for symbiosis. *BMC Evolutionary Biology*. 2007, *7*(1), 1-12.

Ko, D.; Yoo, G.; Yun, S. T.; Jun, S. C.; Chung, H. Bacterial and fungal community composition across the soil depth profiles in a fallow field. *Journal of Ecology and Environment*. 2017, *41*(1), 1-10.

Kohlhepp, B.; Lehmann, R.; Seeber, P.; Küsel, K.; Trumbore, S. E.; Totsche, K. U. Pedological and hydrogeological setting and subsurface flow structure of the carbonate-rock CZE Hainich in western Thuringia, Germany. *Hydrol. Earth Syst. Sci. Discuss*. 2016, *3*, 1-32.

LaMontagne, M. G.; Schimel, J. P.; Holden, P. A. Comparison of subsurface and surface soil bacterial communities in California grassland as assessed by terminal restriction fragment length polymorphisms of PCR-amplified 16S rRNA genes. *Microbial Ecology*. 2003, *46*(2), 216-227.

Lazar, C. S.; Lehmann, R.; Stoll, W.; Rosenberger, J.; Totsche, K. U.; Küsel, K. The endolithic bacterial diversity of shallow bedrock ecosystems. *Science of The Total Environment*. 2019, *679*, 35-44.

Lazar, C. S.; Stoll, W.; Lehmann, R.; Herrmann, M.; Schwab, V. F.; Akob, D. M.; Küsel, K. Archaeal diversity and CO2 fixers in carbonate-/siliciclastic-rock groundwater ecosystems. *Archaea*. 2017, *2017.*

Liu, X.; Li, M.; Castelle, C. J.; Probst, A. J.; Zhou, Z.; Pan, J.; Liu, Y.; Banfield, J. F.; Gu, J. D. Insights into the ecology, evolution, and metabolism of the widespread Woesearchaeotal lineages. *Microbiome*. 2018, *6*(1), 1-16.

Li, X.; Wang, H.; Li, X.; Li, X.; Zhang, H. Shifts in bacterial community composition increase with depth in three soil types from paddy fields in China. *Pedobiologia*. 2019, *77*, 150589.

Li, Y.; Adams, J.; Shi, Y.; Wang, H.; He, J.S.; Chu, H. Distinct Soil Microbial Communities in habitats of differing soil water balance on the Tibetan Plateau. *Sci. Rep*. 2017, 7, 46407

Mackenzie, B. W.; Chang, K.; Zoing, M.; Jain, R.; Hoggard, M.; Biswas, K.; Douglas, R. G.; Taylor, M. W. Longitudinal study of the bacterial and fungal microbiota in the human sinuses reveals seasonal and annual changes in diversity. *Scientific reports*. 2019, *9*(1), 1-10.

Mazel, F.; Davies, T.J.; Gallien, L.; Renaud, J.; Groussin, M.; Münkemüller, T.; Thuiller, W. Influence of tree shape and evolutionary time-scale on phylogenetic diversity metrics. *Ecography*. 2016, 39, 913–920.

McMurdie, P. J.; Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*. 2013, 8(4), e61217.

Mondav, R.; Woodcroft, B. J.; Kim, E. H.; McCalley, C. K.; Hodgkins, S. B.; Crill, P. M.; Chanton, J.; Hurst, G.B.; VerBerkmoes, N.C.; Saleska, S.R.; Hugenholtz, P.; Rich, V.I.; Tyson, G. W. Discovery of a novel methanogen prevalent in thawing permafrost. *Nature communications*. 2014, *5*(1), 1-7.

Murali, A.; Bhargava, A.; Wright, E. S. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*. 2018, *6*(1), 1-14.

Muyzer, G.; De Waal, E. C.; Uitterlinden, A. G. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and environmental microbiology*. 1993, *59*(3), 695-700.

Oksanen, J.; Blanchet, F. G.; Kindt, R.; Legendre, P.; Minchin, P. R.; O'hara, R. B.; Simpson, G. L.; Solymos, P.; Steven, H. M.H.; Szoecs, E.; Wagner, H. Package 'vegan'. Community ecology package, version. 2013, 2(9), 1-295.

Parker, N.; Schneegurt, M.; Thi Tu, A-H.; Forster, B.M.; Lister, P. Microbiology : The effects of pH on Microbial Growth. *OpenStax*. 2017, 1301.

Park, J.; Sanford, R. A.; Bethke, C. M. Microbial activity and chemical weathering in the Middendorf aquifer, South Carolina. *Chemical geology*. 2009, 258(3-4), 232-241.

Pereira, M. B.; Wallroth, M.; Jonsson, V.; Kristiansson, E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC genomics*. 2018, *19*(1), 1-17.

Price, M. N.; Dehal, P. S.; Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution. 2009, 26(7), 1641-1650.

Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Pelplies, J.; Glöckner, F. O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*. 2012, 41(D1), D590-D596.

Reigosa, M. J.; Sánchez-Moreiras, A.; González, L. Ecophysiological approach in allelopathy. *Critical reviews in plant sciences*. 1999, *18*(5), 577-608.

Reith, F. Life in the deep subsurface. *Geology*. 2011, *39*(3), 287-288.

Rinke, C.; Rubino, F.; Messer, L. F.; Youssef, N.; Parks, D. H.; Chuvochina, M.; Brown, M.; Jeffries, T.; Tyson. G.W.; Seymour, J.R.; Hugenholtz, P. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *The ISME journal*. 2019, *13*(3), 663-675.

Ritchey, E.L.; McGrath, J.M.; Gehring, D. Determining soil texture by feel. Agric. Nat. Resour. Publ. 2015, 139.

Satyanarayana, T.; Johri, B. N.; Das, S. K. (Eds.). Microbial Diversity in Ecosystem Sustainability and Biotechnological Applications: Volume 1. Microbial Diversity in Normal & Extreme Environments. *Springer*. 2019.

Shen, C.; Liang, W.; Shi, Y.; Lin, X.; Zhang, H.; Wu, X.; Xie, G.; Chain, P.; Grognan, P.; Chu, H. Contrasting elevational diversity patterns between eukaryotic soil microbes and plants. *Ecology*. 2014, *95*(11), 3190-3202.

Shenhav, L.; Thompson, M.; Joseph, T. A.; Briscoe, L.; Furman, O.; Bogumil, D.; Mizrahi, I.; Pe'er, I.; Halperin, E. FEAST: fast expectation-maximization for microbial source tracking. *Nature Methods*. 2019, *16*(7), 627-632.

Smith, H. J.; Zelaya, A. J.; De León, K. B.; Chakraborty, R.; Elias, D. A.; Hazen, T. C.; Arkin, A. P.; Cunningham, A. B.; Fields, M. W. Impact of hydrologic boundaries on microbial planktonic and biofilm communities in shallow terrestrial subsurface environments. *FEMS microbiology ecology*. 2018, *94*(12), fiy191.

Stevens, T. O., McKinley, J. P. Lithoautotrophic microbial ecosystems in deep basalt aquifers. *Science*. 1995, 270(5235), 450-455.

Tripathi, B. M.; Kim, M.; Kim, Y.; Byun, E.; Yang, J. W.; Ahn, J.; Lee, Y. K. Variations in bacterial and archaeal communities along depth profiles of Alaskan soil cores. Scientific reports. 2018, 8(1), 1-11.

Tucker, C. M.; Cadotte, M. W.; Carvalho, S. B.; Davies, T. J.; Ferrier, S.; Fritz, S. A.; Grenyer, R.; Helmus, M. R.; Jin, L.S.; Mooers, A. O.; Pavoine, S.; Purschke, O.; Redding, D. W.; Rosauer, D. F.; Winter, M.; Mazel, F. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*. 2017, *92*(2), 698-715.

Uroz, S., Kelly, L. C., Turpault, M. P., Lepleux, C., Frey-Klett, P. The mineralosphere concept: mineralogical control of the distribution and function of mineral-associated bacterial communities. *Trends in microbiology*. 2015, 23(12), 751-762.

Van de Peer, Y.; De Rijk, P.; Wuyts, J.; Winkelmans, T.; De Wachter, R. The European small subunit ribosomal RNA database. *Nucleic acids research*. 2000, *28*(1), 175-176.

Wan, X.; Gao, Q.; Zhao, J.; Feng, J.; van Nostrand, J. D.; Yang, Y.; Zhou, J. Biogeographic patterns of microbial association networks in paddy soil within Eastern China. *Soil Biology and Biochemistry*. 2020, *142*, 107696.

Webb, C. O.; Ackerly, D. D.; Kembel, S. W. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*. 2008, *24*(18).

Will, C.; Thürmer, A.; Wollherr, A.; Nacke, H.; Herold, N.; Schrumpf, M.; Gutknecht, J.; Wubet, T.; Buscot, F.; Daniel, R. Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Applied and environmental microbiology*. 2010, *76*(20), 6751-6759.

Xu, T.; Chen, X.; Hou, Y.; Zhu, B. Changes in microbial biomass, community composition and diversity, and functioning with soil depth in two alpine ecosystems on the Tibetan plateau. *Plant and Soil*. 2021, *459*(1), 137-153.

Zhou, Z.; Pan, J.; Wang, F.; Gu, J. D.; Li, M. Bathyarchaeota: globally distributed metabolic generalists in anoxic environments. *FEMS microbiology reviews*. 2018, *42*(5), 639-655.