The bacterial communities of sand-like surface soils of the San Rafael Swell (Utah, USA) and the Desert of Maine (USA)

Yang Wang

To cite this version:


HAL Id: tel-01261518
https://tel.archives-ouvertes.fr/tel-01261518
Submitted on 25 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
THESE DE DOCTORAT
DE L’UNIVERSITE PARIS-SACLAY,
préparée à l’Université Paris-Sud

ÉCOLE DOCTORALE N°577
Structure et Dynamique des Systèmes Vivants

Spécialité de doctorat : Sciences de la Vie et de la Santé

Par

Mme Yang WANG

The bacterial communities of sand-like surface soils of the San Rafael Swell (Utah, USA) and the Desert of Maine (USA)

Thèse présentée et soutenue à Orsay, le 23 Novembre 2015

Composition du Jury :

Mme. Marie-Claire Lett, Professeure, Université Strasbourg, Rapporteur
Mme. Corinne Cassier-Chauvat, Directeur de Recherche, CEA, Rapporteur
M. Armel Guyonvarch, Professeur, Université Paris-Sud, Président du Jury
M. Olivier Martin, Directeur de Recherche, INRA, Examinateur
M. Michael DuBow, Professeur, Université Paris-Sud, Directeur de Thèse
Acknowledgements

There are so many persons I would like to thank for helping me in the study and in the life as a Ph.D. student in France. Firstly, I would like to express my sincere gratitude to my supervisor, Pr. Michael S. DuBow, for his continuous support and guidance. His guidance helped me in all the time of research and writing of this thesis, with his patience and immense knowledge.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Roland Mastrippolito and Prof. Olivier Lespinet, for their insightful comments and encouragement.

My study in France was supported by the China Scholarship Council. I am very grateful for the great opportunity they offered me to become a Ph.D. student in the university of Paris-Sud and financial supports in life.

I am grateful to my fellow labmates of the team of Prof. Michael DuBow: Jorge Osman, Nga Nguyen, Gustavo Ribeiro, Christophe Regeard, and Micheline Terruer, for helping me solving the problems in research and their supports in my life in France.

Last but not the least, I would like to thank my family, my parents and my beloved husband for their encouragement all the time, and sharing with me all the happiness and sadness. Without their unconditional love and supports, I could not reach so far.
# Table of contents.

Acknowledgements ........................................................................................................ 3  
List of Figures ................................................................................................................ 7  
List of Tables .................................................................................................................. 8  
List of abbreviations ...................................................................................................... 9  

CHAPTER 1. Introduction ........................................................................................... 10  

1.1. Identification of bacterial diversity and Next Generation Sequencing technology .......................................................... 11  
1.1.1. Introduction on the development of techniques for the identification of bacteria .......................................................... 11  
1.1.2. Introduction to 16S ribosomal RNA ............................................. 14  
1.1.3. Metagenomics and the Next Generation Sequencing techniques ...... 17  
1.1.4. Bioinformatics for Meta-barcode Sequencing ................................... 24  

1.2. Arid Environments ........................................................................................ 30  
1.2.1. Introduction to Arid Environments .................................................... 30  
1.2.2. Characteristics of Drylands and Desert Soils ................................... 34  
1.2.3. Drylands in North America .............................................................. 37  
1.2.4. Analogue Sites for Mars ................................................................. 40  
1.2.5. Desertification ............................................................................ 41  

1.3. Desert Biodiversity ..................................................................................... 43  
1.3.1. Plants and Animals ............................................................................. 43  
1.3.2. Desert Microbiology ........................................................................ 45  
1.3.3. Bacterial Diversity ........................................................................... 48  
1.3.4. Bacteria in Desert-like Environments and Bacteria involved in Weathering .......................................................... 51  

1.4. Relationship of Environmental Factors and Soil Bacterial Diversity ...... 53  
1.4.1. General introduction to the relationships of environmental factors and soil microbial diversity .......................................................... 53  
1.4.2. Environmental factors in Deserts ....................................................... 55
1.4.3. Adaptation of Bacteria to Arid Environments ........................................ 59

1.5. Objective of Thesis .................................................................................. 61

CHAPTER 2: Bacterial Communities of the Desert of Maine ........................ 62

Abstract .......................................................................................................... 64

Introduction ...................................................................................................... 65

Materials and Methods ................................................................................... 67

Results .............................................................................................................. 70

Discussion ........................................................................................................ 73

Acknowledgements .......................................................................................... 78

References ........................................................................................................ 79

Tables ................................................................................................................ 85

Figure Legends ............................................................................................... 89

CHAPTER 3: Bacterial communities of the deserts in Utah ............................ 93

Abstract .......................................................................................................... 95

Introduction ...................................................................................................... 96

Materials and Methods ................................................................................... 99

Results .............................................................................................................. 102

Discussion ........................................................................................................ 107

Acknowledgements .......................................................................................... 112

References ........................................................................................................ 113

Tables ................................................................................................................ 121

Figure Legends ............................................................................................... 127

4. CHAPTER 4: Discussion and perspective .................................................. 132

4.1. Discussion .................................................................................................. 133

4.1.1. Quality Control ................................................................................ 134

4.1.2. Chimera Check ................................................................................. 138

4.1.3. Taxonomic Classification of Sequences ........................................... 142

4.1.4. OTU Clustering ................................................................................ 145

4.2. Perspective ............................................................................................... 147
References

149
List of Figures

Figure 1.1. Nucleic acid-based methods in identifying bacteria ................. 13
Figure 1.2. Distribution of variable regions in 16S rRNA from E. coli (Chuan et al 2014) ................................................................. 14
Figure 1.3. Three-domain system based on taxonomy of 16S rRNA, Carl Woese and George E. Fox, 1977 ................................................................. 15
Figure 1.4. 16S rRNA secondary structure, adapted from Y. Pablo (2014) .... 17
Figure 1.5. The 454 pyrosequencing chemistry (Petrosino et al 2009) ........ 20
Figure 1.6. Process of Pyrosequencing (Margulies et al 2005). .................. 21
Figure 1.7. Process of Illumina Sequencing ............................................. 22
Figure 1.8. Analysis Process of Data from Meta-barcode Sequencing .......... 24
Figure 1.9. Map of Global Distribution of Dry Land. Regions with declined color follows the aridity index (AI) of the UNEP. Arrows show the major intercontinental trajectories for desert dust (Laity 2009) ......................... 31
Figure 1.10. Classification of deserts. The pictures are from three different types of deserts: A. the Arabian Desert; B. the Gobi Desert; C. the Nambi Desert ................................................................. 32
Figure 1.11. The major types of desert landscapes: A. sand desert, B. stone desert, C. rock desert, D. plateau deserts, E. mountain deserts .......... 35
Figure 1.12. Map of North American deserts (Laity 2009) ......................... 37
Figure 4.1. Quality score distribution of raw sequences ............................. 135
Figure 4.2. Comparison of quality score distribution of sequences trimmed using the three different methods ............................................. 136
Figure 4.3. Number of chimeras detected by the Uchime and Decipher .... 139
Figure 4.4. Number of sequences detected by the two programs in the Kumtagh sample ................................................................. 140
List of Tables

Table 1.1. Comparison of different methods to calculate distances between microbial communities .......................................................................................28
Table 1.2. Summary of studies of hot deserts (Makhalanyane et al 2015) ..........33
Table 4.1. The Phred quality score related to the base calling error probabilities.134
Table 4.2. Comparison of different methods for quality control of sequences
(Original number of sequences is 22320). .................................................136
Table 4.3. Diversity result before and after trimming.................................137
Table 4.4. Percentage of potential non-chimeric sequences classified at
different taxonomic levels........................................................................140
Table 4.5. Summary of sequences in each database assigned to different
taxonomic level......................................................................................143
Table 4.6. Classification result of a sample from desert in Utah .................143
Table 4.7. Summary of classification using different classifieres with 1000
sequences randomly selected from gold_database downloaded from
Usearch. ......................................................................................................144
Table 4.8. Number of OTUs clustered using different software at a distance 3%.146
List of abbreviations

ATP: adenosine tri-phosphate
BLAST: Basic Local Alignment Search Tool
FISH: catalyzed reporter deposition fluorescent in situ hybridization
CFU: colony-forming units
DAPI: 4,6-diamidino-2-phenylindole
DGGE: Denaturing Gradient Gel Electrophoresis
DNA: Deoxyribonucleic acid
FISH: fluorescence in-situ hybridization
NGS: next-generation sequencing
OTU: operational taxonomic unit
PAR: photosynthetically active radiation
PCoA: Principal coordinates analysis
PCR: Polymerase Chain Reaction
qPCR: Real time quantitative PCR
RDP: Ribosomal Database Project
rRNA: Ribosomal Ribonucleic Acid
RNA: ribonucleic acid
SRA: Sequence Read Archive
T-RFLP: terminal restriction fragment length polymorphism
UNEP: United Nations Environment Programme
UPGMA: Unweighted Pair Group Method with Arithmetic Mean
UV: UltraViolet
CHAPTER 1. Introduction
1.1. Identification of bacterial diversity and Next Generation Sequencing technology

1.1.1. Introduction on the development of techniques for the identification of bacteria

Since microorganisms were first recognized after the invention of microscopes in the 17th century, scientists have been looking for ways to isolate microorganisms in pure culture and to characterize isolates to differentiate them from each other [1]. Methods that can be used to characterize microbial isolates, as well as for classification of microorganisms, are essential for research in microbiology.

Traditional methods of bacterial identification based on phenotypic differences of organisms relied on cultivation under laboratory conditions, using gram staining, colony morphology differences and biochemical tests. However, these methods of bacterial identification had two major drawbacks. First, they can only be used for organisms that are able to be cultivated in vitro, which is extremely biased, as it selects only the small minority of bacteria that can grow in a laboratory situation [2]. As well, it frequently takes a long time to delineate characteristics of a clone culture, especially for those that grow slowly. Second, some strains exhibit unique phenotypic characteristics that do not fit into patterns that have been classified as marks of any known microbial categories [3]. These limits to databases will often not be suitable for identification and classification of environmental isolates.

The development of molecular biology offered a set of powerful new tools to accelerate the identification of microorganisms, which allows the examination of nucleic acids and detection of small variations within microbial species and even
within individual strains [4,5]. Accurate and definitive microorganism identification has been used in a wide variety of research and applications [6], including disease diagnosis associated with microbial infections, food production, agriculture and environmental studies [7].

Carl Woese, in 1977, first found that the 16S rRNA gene, one of the genes that makes ribosomal RNA, has evolutionary relationships in all prokaryotic organisms, and the distances of the sequences of 16S rRNA between different organisms likely indicates evolutionary distances [8]. With the development of the Polymerase Chain Reaction (PCR) techniques in the last decade of the 20th century and improvements in nucleic acid sequencing techniques, ribosomal RNA (or DNA) sequencing gradually became the basis for classification of microorganisms and the new gold standard for the taxonomy of microorganisms [9,10].

Now, methods of bacterial identification can be broadly sorted into two categories-genotypic techniques based on profiling an organism's genetic material, and phenotypic techniques based on profiling either metabolism or chemical composition of microorganisms [11]. The two methods focus on different characteristics of microorganisms and are frequently combined as complementary approaches in applications of microbiology research.

**Phenotypic methods.** Phenotypic techniques can generate direct functional information, such as activities of certain enzymes, and metabolic activities in specific groups of organisms [12]. They can also offer valuable information on physical and functional activities at the protein level. Phenotypic methods include traditional morphological characteristics, biochemical testing (presence of various enzymes), serological tests (test on specific antibodies, such as ELISA, and Western blotting), Phage typing (susceptibility to various phages), fatty acid profiles, etc.[13]
**Genotypic methods.** Genotypic techniques in general have the advantage over phenotypic methods of being independent of the physiological state of an organism (such as conditions or stresses in growth) [2]. They are based on the profile of a universal component (DNA or RNA), and certain of these fragments or sequences are unique to an individual organism or a group of closely related organism [14]. By comparing the information of different sequences generated by these fragments, organisms can be identified or classified [15]. Genotypic techniques can be classified into two major categories: fingerprint-based and sequence-based, and PCR is generally essential to both methods. The most commonly used nucleic acid-based methods in identifying bacteria are shown in Fig 1.

![Diagram of nucleic acid-based methods in identifying bacteria](image_url)

**Figure 1.1.** Nucleic acid-based methods in identifying bacteria.
1.1.2. Introduction to 16S ribosomal RNA

The 16S small ribosomal RNA gene (16S rRNA) is a component of the 30S small subunit of prokaryotic ribosomes. The genes for ribosomal RNA have evolved as organisms evolved, and the slight changes that have occurred can provide clues as to how closely or distantly various organisms are related. The 16S rRNA genes comprise nine hypervariable regions (V1-V9) that demonstrate considerable sequence diversity among different bacteria, and nine conserved regions (C1-C9) that remain consistent across different bacterial groups [16]. The distribution of these regions in 16S rRNA are shown in Figure 2.

![Figure 2](Figure 2. Distribution of variable regions in 16S rRNA from *E. coli* (Chuan et al 2014)).

- V1: 61-106 bp;
- V2: 121-240 bp;
- V3: 436-500 bp;
- V4: 588-671 bp;
- V5: 734-754 bp;
- V6: 829-857 bp;
- V7: 990-1045 bp;
- V8: 1118-1160 bp;
- V9: 1240-1298 bp

PCR amplification and analysis of the 16S rRNA genes have been widely used as a culture-independent method for documentation of the evolutionary history and taxonomic assignment of individual organisms, as well as in characterization of microbial communities [17]. The 16S rRNA gene has a number of clear advantages that make it optimal as a marker for these types of studies [18]:

i. The length of this gene is convenient for amplification and sequencing, and certain length of its fragments (one or more variable regions) are sufficient for classification of sequences to deep level.

ii. The highly conserved regions can be used for design of universal PCR primers;
iii. The variable regions of this gene allow for accurate taxonomic classification and phylogenetic identification of microbial communities [19];

iv. Lateral transfer of this gene between taxa are rare [19];
v. Since this gene has been widely sequenced in microbial diversity research, there are many reference databases including a large amount of sequences with taxonomic information, which is convenient to assign query sequences to known taxonomic groups and compare community composition across studies [20].

In 1977, Carl Woese and George E. Fox introduced phylogenetic taxonomy of 16S ribosomal RNA and the three-domain system “tree of life” (Figure 3), through analysis of the 16S rRNA sequences [21]. They first discovered the third domain Archaea, and separated it from Bacteria. Over the past two decades, 16S rRNA gene sequencing analysis has made great studies, and has become the “gold standard” for the taxonomy of microorganisms.

**Phylogenetic Tree of Life**

Figure 1.3. Three-domain system based on taxonomy of 16S rRNA, Carl Woese and George E. Fox, 1977
Despite the advantages of the 16s rRNA gene in the classification of bacteria, it is not perfect for some applications. For example, when sequence variation of the 16S rRNA between two microorganisms is very small, distance measurements of this gene may not be able to provide accurate information, such as the presence of similar species in the same genus [22]. The secondary structure of 16S rRNA (Fig. 4) may include valuable information [18], and has not yet been well explored in taxonomy studies. Also, in research about bacterial diversity and community structure using sequence profiles of the 16S rRNA genes, the copy numbers of this gene are regularly assumed to be consistent among different bacterial groups. These differences can affect the estimation of the abundance of different bacteria [23]. In addition, the classification of microorganisms based on 16S rRNA gene sequences relies on a well constructed database. However, current databases, including RDP, Silva, Greengenes and Genbank databases for 16S rRNA sequences, still have limits on the coverage of different taxonomic groups, especially at a deep level [24,25]. Thus multiple approaches, such as the whole-genome molecular techniques and the use of more target genes, combined with the 16S rRNA gene, could offer more accurate information in the classification and identification of microorganisms.
Figure 1.4. 16S rRNA secondary structure, adapted from Y. Pablo (2014)
1.1.3. Metagenomics and the Next Generation Sequencing techniques

Metagenomics is the study of the genetic material recovered directly from environmental samples [26]. With the development of next generation sequencing (NGS, also termed high-throughput sequencing; HTS) after 2004, metagenomics is helping to access the taxonomic and functional composition of microbial communities in any environmental biome, without the need to isolate or culture them in the laboratory [27]. While traditional microbiology relies on cultivated clonal cultures, the vast majority of microbial biodiversity had been missed [1]. Massively parallel sequencing techniques have revolutionized sequencing capabilities, far beyond the electrophoresis-based “first generation” sequencing, and launched the “next-generation” in genomic science. Metagenomics has already been successfully applied in many fields, including the analysis of the microbiome of natural water and soil environments, some extreme physical and chemical environments, food supply chains, animals, and human health. In recent years, to explore the taxonomic complexity, meta-barcode methods are broadly used, which is an amplicon-based approach, based on PCR-targeted sequencing of selected genetic species markers (such as some hypervariable regions of the 16S rRNA gene) [28].

Shotgun sequencing methods are also widely used in metagenomic samples, since shotgun metagenomics can provide information about both which organisms are present and what metabolic processes are possible in the community [29]. DNA sequences are randomly broken up into many small pieces and then reassembled by matching regions of overlap. Many organisms, which may be overlooked using traditional culturing techniques, may be retained as small sequence segments using shotgun sequencing. So to achieve the high coverage of different community members, especially under-represented ones, full sequencing of the genomes of large samples is often required.
Organisms and gene functions of the microbial communities in a specific environment can be obtained by shotgun sequencing of the whole metagenome. However, if we only need to analyses the taxonomic composition and biodiversity assessment of microbial community, an environmental DNA (e-DNA), meta-barcoding approaches can constitute an effective and less expensive solution [28]. It uses universal PCR primers, which are assumed to cover all the species belonging to the explored taxonomic range, and culture-independent sequencing of a selected genetic taxonomic marker (meta-barcode) from a mass collections of organisms or from e-DNA. The PCR products are then sequenced on a next generation sequencer, such as the Roche 454 and Illumina sequencers.

The 16S rRNA gene is commonly used to identify bacteria by meta-barcodes. In the case of recent versions of 454 technology, 16S rRNA gene sequences provides information of bacterial community biodiversity and relative taxa abundance down to the genus level [30].

The first metagenomic studies conducted high-throughput sequencing used massively parallel 454 pyrosequencing [31]. Three other technologies commonly applied to environmental sampling are the Ion Torrent Personal Genome Machine, the Illumina MiSeq or HiSeq and the Applied Biosystems SOLiD systems [32]. These techniques for sequencing DNA generate shorter fragments than traditional Sanger sequencing.

**The Roche 454 Genome Sequencer**

In 2008, 454 Sequencing launched the GS Titanium series reagents for use on the Genome Sequencer FLX instrument. This uses a large-scale parallel pyrosequencing system with the ability to sequence 400-600 million nt per run with 400-500 nt read lengths [33]. The system relies on fixing nebulized and adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion [34]. The DNA
fixed to these beads is then amplified by PCR. The technique is built on 4-enzyme real-time monitoring of DNA synthesis by bioluminescence, using a cascade that, upon nucleotide incorporation, ends in a detectable light signal (bioluminescence). The pyrosequencing chemistry is shown in Figure 5.

Figure 1.5. The 454 pyrosequencing chemistry (Petrosino et al 2009)

Genomic DNA fragments (such as 16S rRNA gene amplicons), are ligated to short adaptors, which provide priming sequences for both PCR amplification and sequencing of the sample-library fragments. Single-stranded template DNA (sstDNA) library is immobilized onto beads. The beads containing a library fragment carry a single sstDNA molecule. The bead-bound library is emulsified with the amplification reagents in a water-in-oil mixture. Each bead is captured within its own micro-reactor where PCR amplification occurs. Sequencing-by-synthesis then occurs by the DNA polymerase-driven generation of inorganic pyrophosphate, resulting in the formation of adenosine triphosphate (ATP) and ATP-dependent conversion of luciferin to oxyluciferin (Fig. 6). The generation of oxyluciferin results in the emission of photons of light, and the amplitude of each signal is directly related to the presence of
one or more nucleosides [35].

The Illumina Genome Analyzer

Illumina sequencing also use sequencing-by-synthesis methods. It begins with the attachment of single stand template DNA (sstDNA) to primers on a slide, and the slide is flooded with nucleotides and DNA polymerase. Four different nucleotides (ATCG) are added, and these nucleotides are fluorescently labelled, with the colour corresponding to the base. They also have a terminator, so that only one base is added at a time. The four bases then compete for binding sites on the sstDNA to be sequenced and non-incorporated molecules are washed away [36]. After each
synthesis, a laser is used to excite the dyes and a photograph of the incorporated base is taken (Figure 7). Illumina sequencing only uses DNA polymerase, instead of multiple enzymes required by pyrosequencing. The current MiSeq platform can yield ~50 million 300 nt reads per run [37].

Ion Torrent Personal Genome Machine

Unlike Illumina and 454, Ion torrent sequencing does not use optical signals. Instead, it exploits the fact that addition of a dNTP to a DNA polymer releases an H+ ion [38]. DNA fragments of approximately 200 bp in length with adaptors are placed onto a bead. Then, DNA strands are amplified on the bead by emulsion PCR, with each bead in a single well of a slide. Like 454, the slide is flooded with a single type of dNTP, along with buffers and polymerase. The pH is detected in each of the wells, as each H+ ion released will decrease the pH. So the sequences of the read are determined by detecting the changes in pH.

Iron Torrent has the same limitation as pyrosequencing, that it is difficult to enumerate
long repeats (homopolymers). The read length in achieved by Ion Torrent semiconductor sequencing is currently 400 nt, and the throughput is currently lower than that of other high-throughput sequencing technologies [39].

Among the three most widely used sequencing techniques, 454 pyrosequencing is broadly applied in meta-barcode analysis because of its longer sequence reads [40].
1.1.4. Bioinformatics for Meta-barcode Sequencing

The massively-parallel sequencing methods are capable of producing millions of reads, which presents a huge challenge for data storage, analyses and other manipulations [41]. Once sequencing is complete, raw sequence data must go through several analysis steps (Figure 8). A generalized data analysis pipeline for meta-barcode sequencing data includes: 1) preprocessing the data to remove adapter sequences and grouping to different libraries according to the specific barcodes, 2) checking the quality score of each base to remove low-quality reads, 3) mapping of the data to a reference database or stand-alone alignment of the sequence reads not relying on a reference database, and remove chimera sequences, 4) assign the sequences to their taxonomic classes, 5) cluster sequences to operational taxonomic unit (OTU) at a specific similarity level (normally 97%), and 6) statistical analysis of microbial community diversity within an environmental library (α-diversity) or between different libraries (β-diversity), based on OTU distribution or phylogenetic distances. Many free online tools and software packages exist to perform the bioinformatics necessary to analyze sequence data.

Figure 1.8. Analysis Process of Data from Meta-barcode Sequencing.
Cleaning of Raw Data

Raw data obtained from a sequencer needs to be grouped by barcodes (for multiple samples), and primers and adapters are removed from each read [42]. Then, the data are reformatted to fasta or fastQ format for further analysis. Several criteria are applied to remove sequence noise [43]: (i) noise length that extends the range of selected marker gene, (ii) more than one mismatch to the primer or barcode sequences, and (iii) the presence of homopolymers of > 8 bp in length. Then, the sequences are checked for quality score by trimming off the ends that normally contain low-quality bases, or filter off low-quality sequences (such as those that contain too many ambiguous bases, average quality scores are low, etc.). Regular pipelines available for these analysis procedures are Mothur [44], Qiime [45], and Galaxy [46].

Chimera sequences can be produced during the PCR process, originating from two or more parental sequences, that can be a large part of NGS sequencing errors. Several programs are used to detect chimeras from the data, using reference databases or de-novo methods, such as UChime [47], Decipher [48], and Chimera Slayer [49]. Since it is difficult to identify chimeras from data generated by massive sequence amplification, sometimes different methods can be combined.

Sequence Classification

Currently, the bioinformatics methods used to assign metagenomics sequences to their taxonomic classes adopt essentially three approaches, similarity-based methods, composition-based methods, and phylogenetic-based methods [28,50].

For the similarity-based methods, even short sequences (about 200 bp) can be classified, and the classification of sequences commonly relies on its comparison to curated collections of sequences in a reference database. So the absence of reference sequences for unknown species, especially for environmental samples, can affect the
assignment accuracy. Currently, the most widely used reference resources available for 16S rRNA gene-based identification of microorganisms are the Greengenes [51], Ribosomal Database Project II (RDP II) [52], and SILVA databases [53].

The composition-based methods use a feature space consisting of 6-8 base subsequences (words) to assign the metagenomic sequences to taxonomically annotated reference sequences [54]. The reference data are pre-treated using different methods, such as the naïve Bayesian classifier and the k-nearest-neighbor algorithms. This approach requires more computational abilities than the similarity-based methods.

The phylogenetic-based methods calculates the phylogenetic distances according to a reference evolutional phylogenetic tree, and assigns the sequences using maximum likelihood, neighbor-joining algorithms, etc. This method requires huge computational demands, that are not available to all researchers. To achieve both accuracy and efficiency in the classification of meta-barcode sequences, some programs combine different algorithms [55].

**Statistical Analysis**

To study the biodiversity of microbial communities, Operational Taxonomical Units (OTU) are commonly used, since it is difficult to calculate on their real species numbers in different communities [56,57]. For meta-barcode sequencing data, not all sequences can be classified at deep taxonomic levels, which makes it more efficient to use OTU methods to reveal the structure of microbial communities [58]. OTUs are clustered based on sequence similarity, and the most accepted similarity is 97%, representing the potential distances among different species, although 95% or 99% are also used [59].

Among different methods adapted to group sequences into different OTU clusters,
Two algorithms are broadly used, picking OTUs based on alignment to reference database (i.e. Mothur and RDP II pipelines), and picking OTUs based on the calculation of sequence distances among each other, such as the Uparse) [60,61]. After OTU tables are generated, further statistical analyses can be performed.

Two biodiversity indices are broadly used to estimate the richness and evenness of a microbial community, Chao 1 and Shannon, based on the distribution of OTU populations in a community [62,63].

**Chao 1 Estimator** calculates the estimated true species diversity of a sample by the equation (The calculations for the bias-corrected Chao1 richness estimator in the program EstimateS).

\[ S_{\text{chao1}} = S_{\text{obs}} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)} \]

where,

- \( S_{\text{chao1}} \) = the estimated richness
- \( S_{\text{obs}} \) = the observed number of species
- \( n_1 \) = the number of OTUs with only one sequence
- \( n_2 \) = the number of OTUs with only two sequences

The **Shannon diversity index** (\( H \)) is another index that is commonly used to characterize species diversity in a community. Shannon's index accounts for both abundance and evenness of the species present.

\[ H_{\text{shannon}} = - \sum_{i=1}^{S_{\text{obs}}} \frac{n_i}{N} \ln \left( \frac{n_i}{N} \right) \]

where,

- \( S_{\text{obs}} \) = the number of observed OTUs
- \( n_i \) = the number of individuals in OTU \( i \)
- \( N \) = the total number of individuals in the community

Other statistical methods, such as rarefaction curves and the Simpson Index, are also used.
Comparisons Between Communities

To investigate the differences between two microbial communities or among multiple libraries based on their composition, distance matrices are calculated using different algorithms, such as Bray-Curtis [64], Jaccard [65], Sorenson [66], weighted and un-weighted Unifrac [67]. OTU-based or phylogenetic-based distance matrices between every pair of community samples are presented in a square matrix. A comparison of different methods is shown in Table 1. Then, the matrix can be plotted using Principal Coordinate Analysis (PCoA) [68], UPMGA-tree (Unweighted Pair Group Method with Arithmetic Mean) or other bioinformatic methods [69]. Pearson correlation coefficients, to explore the relationship between microbial community structure and environmental factors, or between different taxa is also used [70].

Table 1.1. Comparison of different methods to calculate distances between microbial communities

<table>
<thead>
<tr>
<th>Dissimilarity Measures</th>
<th>Species numbers</th>
<th>Phylogenetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence (presence or absence)</td>
<td>Sorensen</td>
<td>Un-weighted UniFrac</td>
</tr>
<tr>
<td>abundance</td>
<td>Bray-Curtis</td>
<td>Weighted UniFrac</td>
</tr>
</tbody>
</table>
List of Data Analysis Platforms and Pipelines for Meta-barcode Sequencing

1. Ribosomal Database Project (RDP II)
   http://rdp.cme.msu.edu/
2. Mothur
   http://www.mothur.org/wiki/Main_Page
3. Qiime
   http://qiime.org/
4. Usearch
   http://www.drive5.com/usearch/
5. SILVA rRNA database project
   http://www.arb-silva.de/
6. Galaxy
   https://galaxyproject.org/
7. MGRAST
   http://metagenomics.anl.gov/
8. Prinseq
   http://prinseq.sourceforge.net/
9. Greengenes
   http://greengenes.lbl.gov/cgi-bin/nph-index.cgi
10. ESPRIT
    http://www.ijbcb.org/ESPRITPIPE/php/onlinetool.php
11. CAMERA
    http://metagenomics.anl.gov/
12. R
    http://www.r-project.org/
1.2. Arid Environments

1.2.1. Introduction to Arid Environments

Hot deserts are regions of land that have little rainfall, where few plants and animals generally exist. Aridity is the dominant climatic factor over about one-third of the land surface of the world, as approximately 7.5% of global land area is classified as extremely arid (hyperarid), 12.5% arid, and about 17.7% semiarid. If dry-subhumid areas (9.9%) are included in the classification, then drylands comprise about 47% of the Earth’s land surface (United Nations Environment Programme, UNEP 1992). In total, about 49 million km² are affected by aridity. Of the total surface area of arid climate, Africa is 36.7%, Asia 31.7%, North America 12%, Australia 10.8%, and South America 8.8% [71,72]. To sum up, the dry areas of the world occupy more land than any other major climatic type [73].

Deserts are considered to be the hyperarid and arid regions, and semiarid and dry-subhumid regions the desert fringes. It is difficult to derive an exact definition of a desert, as aspects of climate (precipitation, evaporation, and temperature), geomorphic features, and flora and fauna, show considerable variation. Although they may share common features such as wide temperature variation, winds, geomorphology, shifting sands, and plant and animal life, they are not components of all arid environments. There are many aridity indexes to measure the level of aridity for a region, an assessment of the extent of drylands broadly applied is aridity index (AI) conducted by Hulme and Marsh (1990) on behalf of UNEP (1992), where

\[ AI = \frac{P}{PET}, \]

P stands for the mean precipitation of a fixed time period, PET is potential evapotranspiration.
An examination of the distribution of deserts based on the AI is shown in Fig 9. The delimitation of the different types of dryland environments by AI values are dry-subhumid (AI = 0.50 - < 0.65), semi-arid (AI = 0.20 - < 0.50), arid (AI = 0.05 - < 0.20), and hyper-arid (AI = < 0.05).

The classification of deserts relies on combinations of the total amount of annual rainfall, temperature, humidity, or other factors. In general, deserts are classified into 3 categories (Fig 10): Hot and dry deserts (such as South and Central America, Africa and Australia), Cold deserts (such as Antarctic, Greenland), Coastal deserts (mostly found on the western edges of a continent, and a famous one is the Atacama Desert in Chile). Table 2 shows a summary of studies of hot deserts [74].

Figure 1.9. Map of Global Distribution of Dry Land. Regions with declined color follows the aridity index (AI) of the UNEP. Arrows show the major intercontinental trajectories for desert dust (Laitty 2009).
Figure 1.10. Classification of deserts. The pictures are from three different types of deserts: A. the Arabian Desert; B. the Gobi Desert; C. the Nambí Desert.
Table 1.2. Summary of studies of hot deserts (Makhalanyane et al 2015)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Country</th>
<th>Name</th>
<th>Approx. size (km²)</th>
<th>Approx. temperature range (°C)</th>
<th>Approx. precipitation (mm/yr)</th>
<th>Average soil pH</th>
<th>Average soil organic carbon (%)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperarid</td>
<td>Chile</td>
<td>Atacama-Sechura</td>
<td>105 000</td>
<td>-5–40</td>
<td>0–20</td>
<td>6.6–9.2</td>
<td>0.1–2.6</td>
<td>[75-77]</td>
</tr>
<tr>
<td></td>
<td>Southwestern</td>
<td>Namib</td>
<td>81 000</td>
<td>5–45</td>
<td>5–100</td>
<td>7.9–8.5</td>
<td>0.1–0.3</td>
<td>[78,79]</td>
</tr>
<tr>
<td></td>
<td>Northern Africa</td>
<td>Sahara</td>
<td>9 100 000</td>
<td>-5–45</td>
<td>5–150</td>
<td>7.6–7.9</td>
<td>0.1–1.2</td>
<td>[80-82]</td>
</tr>
<tr>
<td>Arid</td>
<td>Southern</td>
<td>Gobi</td>
<td>53 000</td>
<td>-20–30</td>
<td>30–100</td>
<td>7.7–10.2</td>
<td>0.1–2.64</td>
<td>[83]</td>
</tr>
<tr>
<td></td>
<td>Mongolia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arabian Peninsula</td>
<td>Arabian</td>
<td>2 300 000</td>
<td>5–40</td>
<td>25-230</td>
<td>7–7.5</td>
<td>NA</td>
<td>[84]</td>
</tr>
<tr>
<td></td>
<td>South Africa</td>
<td>Karoo</td>
<td>395 000</td>
<td>2–40</td>
<td>50–200</td>
<td>6.9–9</td>
<td>0.3–1.3</td>
<td>[85]</td>
</tr>
<tr>
<td></td>
<td>Southwestern USA</td>
<td>Mojave</td>
<td>152 000</td>
<td>-10–50</td>
<td>30–300</td>
<td>7.1–9.4</td>
<td>0.04–0.1</td>
<td>[86]</td>
</tr>
<tr>
<td></td>
<td>Central Australia</td>
<td>Simpson</td>
<td>180 000</td>
<td>5–40</td>
<td>50–400</td>
<td>6.5–7</td>
<td>0.1–0.3</td>
<td>[87,88]</td>
</tr>
<tr>
<td></td>
<td>North Mexico</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arid-semiarid</td>
<td>Chihuahuan</td>
<td>455 000</td>
<td>10–40</td>
<td>70–400</td>
<td>5.9–6.2</td>
<td>0.2–1.9</td>
<td>[89]</td>
</tr>
<tr>
<td></td>
<td>Southern</td>
<td>Sonoran</td>
<td>312 000</td>
<td>-10–50</td>
<td>70–400</td>
<td>5–8.6</td>
<td>0.4–2</td>
<td>[90]</td>
</tr>
<tr>
<td></td>
<td>Southwestern USA</td>
<td>Kalahari</td>
<td>520 000</td>
<td>-10–45</td>
<td>100–250</td>
<td>7.7–8.7</td>
<td>0.1–0.5</td>
<td>[91-93]</td>
</tr>
<tr>
<td></td>
<td>Southern</td>
<td>Negev</td>
<td>13 000</td>
<td>5–40</td>
<td>100–300</td>
<td>7.2–8</td>
<td>0.5–0.7</td>
<td>[94]</td>
</tr>
<tr>
<td></td>
<td>North Mexico</td>
<td>Thar</td>
<td>200 000</td>
<td>4–50</td>
<td>200–300</td>
<td>7.9–8.1</td>
<td>0.3–0.4</td>
<td>[95]</td>
</tr>
<tr>
<td>Semiarid</td>
<td>Southern</td>
<td>Gibson</td>
<td>156 000</td>
<td>6–40</td>
<td>200–400</td>
<td>NA</td>
<td>0.06</td>
<td>[96]</td>
</tr>
<tr>
<td></td>
<td>Northwestern</td>
<td>Great Sandy</td>
<td>285 000</td>
<td>10–40</td>
<td>250–370</td>
<td>5.8–6</td>
<td>0.1–1.1</td>
<td>[96]</td>
</tr>
<tr>
<td></td>
<td>Australia</td>
<td>Tanami</td>
<td>185 000</td>
<td>10–40</td>
<td>300–500</td>
<td>4.9–6.7</td>
<td>0.1–1.4</td>
<td>[97]</td>
</tr>
</tbody>
</table>
1.2.2. Characteristics of Drylands and Desert Soils

**Climate.** The world’s deserts can be divided into three categories (Meigs 1953): hot deserts in tropical and subtropical latitudes, temperate deserts in higher latitudes in continental interiors, and coastal deserts found on the west coasts of continents in tropical latitudes. Hot inland deserts are identified by large temperature variations, and the persistence of high daily temperatures, with maximum temperatures commonly between 45 and 49°C [73]. Temperatures on the soil surface can be considerably higher, as much as 75-80°C [98]. Temperate deserts are characterized by considerable seasonal variations in temperature, and a dependable period of cold temperatures. These deserts are at higher latitudes than hot deserts, some precipitation occurs as snow, and soil moisture is often frozen. They have hot summers counterbalanced by relatively cold winters. In the arid regions of Antarctica, mean winter temperatures may be as low as -30°C [99]. Coastal deserts tend to have relatively low seasonal and diurnal ranges of temperature.

**Precipitation.** Deserts generally receive relatively low amounts of total annual precipitation. Spatial variation in rainfall is high, which leads to the high biodiversity in some desert areas [100]. Rainfall in deserts tends to fall in pulses [101], which can vary considerably in magnitude and timing. For some deserts at coastal areas, where rainfall is very low, fog becomes the major resource of precipitation [102]. The high level of fog in coastal desert fog zones provide habitats extremely favorable for lichen growth [103]. Mountain snow plays an important role in water storage and release in deserts that are in proximity to high-altitude mountain ranges. High runoff from desert slopes may offer sufficient water availability to agriculture and helps dampen the variability of the hydrograph [104].

**Landscapes.** There are five major types of desert landscapes that are commonly recognized: sand deserts, stone deserts, rock deserts, plateau deserts, and mountain
deserts (Fig 11). The sand desert landscape probably accounts for 15-20% of desert land [104], and is thus not as common as often perceived. Stone deserts usually have a gravel surface, which is covered by rocks too large to be carried away by wind or water, and also known as desert pavement. Rock desert landscapes normally have bare rock surfaces. Plateau landscapes are often found in a desert landform of mountain-and-basin deserts. Mountain deserts constitute a landscape form called shield desert, where wind is a more effective force than water, compared with mountain-and basin desert.

![Figure 1.11. The major types of desert landscapes: A. sand desert, B. stone desert, C. rock desert, D. plateau deserts, E. mountain deserts](image)

**Saline Soils.** Soils from desert environments are dominated by a mineral component with low organic matter, but the repeated accumulation of water in certain soils can cause salts to precipitate [105]. When the water table rises to within about 2 m of the ground level, water may begin to rise to the surface by capillary action. Then, dissolved salts will be carried up to the surface, and concentrate in the upper layers of the soil as water is evaporated [106]. Salinity changes the electrochemical balance of soil particles, which is harmful to plant cells, and can also increases soil erosion.

**Radiation.** The desert atmosphere is relatively clear. Clouds are rare, and the water
vapor content is low. The paucity of clouds has several important consequences. Incoming solar energy approaches a maximum in arid regions owing to the lack of cloud cover. Approximately 80% of solar radiation at the top of the atmosphere reaches the surface [73]. The distribution of radiation is characterized as seasonal, normally reaching a peak in June in the Northern Hemisphere and December in the Southern Hemisphere [107].

**Desert Dust.** Mineral dust is the most important export from the world’s arid zones to the global Earth system, and affects atmospheric, oceanic, biological, terrestrial and human processes and systems [108]. It is responsible for suppression of rainfall [109], long-distance microorganisms transport, risks to human health, and agricultural soil erosion and productivity. Dust emission from arid environments in China represents as much as half of the global atmospheric loading of dust, while North America has only one very small zone located in the Great Basin with high values of dust loadings into the atmosphere [110]. The increasing greenhouse gas emissions have brought large changes to climate, and many parameters that control dust emission, such as vegetation, rainfall, soil moisture and surface wind speed are expected to change [111].
1.2.3. Drylands in North America

There are a diverse number of small deserts stretching from southeastern California to western Texas, and from Nevada and Utah to the Mexican states of Sonora, Chihuahua, and Coahuila and much of the peninsula of Baja California (Figure 12). Beyond these areas, semi-arid conditions extend north to eastern Washington, south to the central Mexican plateau, and east to link with the steppes of the High Plains. About 55% of the North American deserts are considered semiarid, 40% arid, and only 5% hyperarid [73].

Figure 1.12. Map of North American deserts (Laity 2009)
Many factors can work to create deserts. The rain shadow effect contributes to the form of North American deserts, which is caused by mountain ranges blocking moisture of Pacific origin in the winter and the Gulf of Mexico in summer. Rain shadow effects are major in the Great Basin and Mojave Deserts, while the effects of high pressure are important in the Chihuahuan and Sonoran Deserts. Increasing precipitation of some regions between July and mid-September over large areas of the southwestern USA and northwestern Mexico is called the North American monsoon [112], which is defined as sites that receive at least 50% of their annual precipitation during the summer period. North American monsoon systems develop in response to atmospheric moisture supplied by nearby warm oceans. The boundary between cold and warm deserts lies across southern Nevada and Utah [73].

The Great Basin, characterized by its high altitude at its northern position, is considered a “cold” desert. Approximately 60% of its precipitation in winter comes in the form of snow, and mean monthly temperatures is below 0°C from December to February. The mean annual temperature is 9°C [113]. Most regions have an elevation > 1200 m, with mountains up to 3000m in height. Mean precipitation varies from 2 to 300 mm among different sites and is evenly distributed throughout the year. Much of Nevada and Utah are in the Great Basin Desert [114].

The Colorado Plateau is a roughly circular area > 300,000 km² passing at higher elevation, that consists of plateaus and isolated mountains of Utah, Colorado, New Mexico, and Arizona [115]. Although the plateau lies mostly above 1500 m, it shows high internal variation [116]. Differential erosion characterizes a landscape dominated by canyons, cuesta scarps, and plains. Dune fields on the Colorado Plateau are not extensive, but have a variety of forms. Wind erosion features of the Colorado Plateau comprise deflation hollows, yardangs, wind-fluted cliffs, and blowouts [73].

The Mojave Desert is the smallest North American desert, and occupies
approximately 140,000 km$^2$ in southeastern California and southern Nevada, with elevations mostly above 1000 m [117]. The late Quaternary climate had a large impact on the surficial stratigraphy of the area [118]. It is roughly rectangular in shape, bounded by the Great Basin Desert to the north, and the Sonoran Desert to the south. It has an annual rainfall from ranging 76 to 102 mm across the desert floor, and reaching about 279 mm with increasing elevation. The Mojave Desert is characterized by numerous mountain ranges, valleys, endorheic basins, salt pans, and seasonal saline lakes. Most of the valleys are internally drained, such that all precipitation that falls within the valley does not eventually flow to the ocean [119].

**The Sonoran Desert** covers approximately 275,000 km$^2$ in large parts of the Southwestern United States in Arizona and California, and of Northwestern Mexico in Sonora, with elevations ranging from below sea level in California to about 1500 m in mountain foothills [120]. The Sonoran Desert is more subtropical than other North American deserts, with summer high temperatures reaching 49°C or more [121].

**The Chihuahuan Desert** occupies approximately 518,000 km$^2$ in southwestern North America, with major parts in northern Mexico, and one forth ranging in western Texas and southern New Mexico. Most sites are located at elevations from 600 to 1,500 m. This desert has more rainfall than other warm desert ecoregions, with precipitation ranging from 150 to 400 mm [122]. The Chihuahuan Desert encompasses one of the most biologically diverse arid regions on Earth [123].
1.2.4. Analogue Sites for Mars

Some arid landscapes contain a range of Mars analogue features, relevant for geology and astrobiology studies. These features include wind erosion, moisture deficits, absence of vegetation, high UV radiation, etc., although surface temperature, atmospheric pressures, gravity and physiochemical composition are very different from that of Mars. The EuroGeoMars 2009 campaign was organized at the Mars Desert Research Station (MDRS) to perform multidisciplinary astrobiology research. MDRS in southeast Utah is situated in a cold arid region with mineralogy and erosion processes comparable to those on Mars [124].

Particular deserts reveal extreme Mars-like surface characteristics, such as the cold Antarctic desert McMurdo Dry Valley which is considered to have the coldest, driest and most oligotrophic soils [125]. Mars is the third largest planet in the solar system. It is often referred to as the "Red Planet" because of the iron oxide prevalent on its surface, which is similar to deserts in Utah, characterized by red-colored hills, soils and sandstones. MDRS is located in a cold arid desert with an temperature of 35°C at midday on the equator and -43 °C during the polar winters [124]. It consists of minerals containing silicon and oxygen, metals, and other elements that typically make up rock. Our ability to study the surface of other planets in the solar system is very limited, and studies on the terrestrial analogues located on Earth will help to build the foundation of terrestrial desert exploration.
1.2.5. Desertification

Desertification - the spread or intensification of land degradation towards greater aridity due to climatic changes or human activities - is occurring at an alarming rate around the world [126]. Dryland degradation results in huge economic losses and directly affects more than 1 billion people who depend on such areas for their livelihood, particularly small farmers (United Nations Environment Programme, UNEP 2012). UNEP estimates that 69% of agricultural drylands in the world are degraded or undergoing desertification.

Arid zones are the most vulnerable areas, and characterized by extreme drought. But drought alone cannot be responsible for desertification. Emanuel et al. (1985) predicted a dramatic increase in global desert lands due to climate changes with a doubling of the atmospheric CO₂ concentrations [127]. Soil salinization, agricultural development in marginal desert lands and housing developments can negatively affect arid environments. Soil salinization reduces soil quality, limits the growth of crops, constrains agricultural productivity, and in severe cases, leads to the abandonment of agricultural soils [128]. High soil salinity occurs naturally in deserts, but poor water management in irrigated areas raises the natural salinity of the soil to the soil surface [129].

Aquifer pumping in desert golf courses reduces the groundwater and increases soil salinity, as well as mineralization and chemical pollution of watercourses [130]. For example, the use of the water from the Colorado River for urban purposes in southern California has resulted in the river no longer reaching the sea. In North America, the replacement of grasslands by woody species with shrubs are particularly negative effects of desertification, making landscapes vulnerable to wind and water erosion [131], and soil erosion results in the loss of biodiversity. Drier conditions linked to increased demand for ground water pumped for agricultural irrigation, particularly in
the central and western US [132], results in the depletion of aquifers [133,134]. Thus, human use can create desert-like conditions in lands that were previously far more productive.

Desertification also contributes to other environmental and social crises, such as the mass migration of people and animals, species loss, climate change and the need for emergency assistance to human populations. It affects both developed and underdeveloped nations, up to 66% of the African continent is threatened by aridity, and almost 40% of land in the continental United States is vulnerable to desertification, estimated by the US Bureau of Land Management [135].

As humans make increasing use of dryland resources, hazards associated with aeolian and fluvial processes will be more intense [136]. It is clear that agricultural activities and water resource use in drylands may result in the potential acceleration of erosion.
1.3. Desert Biodiversity

1.3.1. Plants and Animals

Although it has been often suggested that deserts are relatively simple ecosystems characterized by low biodiversity, some research suggests that deserts are relatively complex and biologically rich [106]. Deserts support various fauna and flora, including terrestrial plants and animals.

Plants have developed various morphological and physiological adaptations to live in the desert environment, classified as drought-escaping, drought-evading, or drought-resisting. There are some principal adaptations: 1) geophytes and other plants have special storage organs, 2) trees and shrubs with deep root systems are able to exploit deep aquifers in dry environments, 3) germinating immediately after the infrequent rains and completing their life cycles before summer heat [106], 4) plants have rapid gas exchange and small leaf surfaces to minimize heat input.

Local geographic factors, such as the mineral composition, nutrient reserve, organic content, and capacity to hold water, can affect the distribution and abundance of plants [73]. Plants can be classified as xerophytes, mesophytes, or phreatophytes according to their water requirements. Xerophytes plants, that can survive and reproduce when water is limited, dominate desert environments. Grasses are the most abundant species of plants in deserts [137]. There are several common desert plants found in desert environments, with different characteristics across the world owing to their different climatic conditions, including barrel cactus, brittle bush, palm trees, jumping cholla, saguaro cactus, etc. Vegetation in deserts can have large effects in different geological processes, including reducing sand transport and changing hydrology and dryland river forms.
Animals have evolved various strategies in order to cope with the extreme conditions in deserts, which can be classified into three categories: behavioral, morphological, and physiological. Morphological adaptations to heat include smaller body sizes and relatively larger surface areas, and light-colored surfaces to reflect radiation. Physiological adaptations are less common. They include dormancy during summer, uric acid as a major nitrogenous waste, the deposition of fat in tails or humps, salt glands that secrete salt without the loss of fluids, and an absence of sweat glands [106].

The Chihuahuan desert is one of the most biologically rich and diverse desert ecoregions in the world, others include the Great Sandy Tanmi Desert of Australia and the Namib-Karoo desert of southern Africa [114]. The high degree of local endemism is the result of the isolating effects of complex basin and range physiography, and dynamic changes in climate over the last 10,000 years [138]. Local species diversity is related to rainfall, more rich in semi-arid zones, while rockiness enhances species diversity because of the presence of many micro-niches.
1.3.2. Desert Microbiology

Arid lands account for the largest terrestrial biome [126], and stresses such as drought, temperature and radiation limit the scale of life extension. Research concerning microbial colonization and dispersion in deserts has been performed to estimate the function of microbial communities from desert sand which may play an important role in soil stability, nutrient cycles and environmental health.

Previous studies have predicted that there may be as many as $10^7$ to $10^9$ unique bacterial species on Earth [139], but with additional sequencing efforts, the species richness may increase. Species richness estimate are significantly higher in non-polar as compared with polar deserts [140]. Hot deserts supported significantly higher abundances of heterotrophic bacteria relative to photoautotrophic bacteria in cold deserts, and implied that productivity is higher in hot deserts and therefore capable of supporting greater biomass and trophic complexity than in cold deserts [141].

In many deserts, small poikilohydric life forms constitute a thin veneer on or within the top few centimeters of most soil and rock surface, which typically contain cyanobacteria, chlorophytes, fungi, heterotrophic bacteria, lichens and mosses [126]. Soil and rock surface communities are widespread and share general similarities in all the hot and cold deserts that have been examined worldwide. *Cyanobacteria* are major N$_2$ fixation hot spots in arid lands that support an array of heterotrophic microorganisms, and normally dominant the soil and rock surface community in arid environments. A study in the Atacama desert showed that the non-cyanobacteria phototrophic bacteria *Chloroflexi* are dominant in the hyper-arid core of the desert [76].

**Biological soil crusts (BSCs)** are specialized communities comprised of mosses, lichens, liverworts, cyanobacteria, and other organisms in the topsoil of terrestrial
environments [142,143]. They play important ecological roles in vegetation and ecological restoration in desert regions through aggregation of soil particles that reduce wind and water erosion, and different crust developmental and successional stages have different ecological functions [144]. BSCs increase infiltration and nitrogen fixation and contribute to local soil organic matter. They may constitute as much as 70% of the cover of biological organisms in a particular community [145]. These organisms are capable of withstanding desiccation, and often equilibrate their activities with soil moisture content [146]. BSCs are very sensitive to destruction by human activities such as grazing, agriculture, construction and outdoor activities [147].

**Epiliths** are observed in different types of arid environments. Lichens and mosses commonly occur on rock surfaces. Biogenic varnishes are well presented, that are associated with microorganisms, covering *Actinobacteria* and *Proteobacteria*.

**Hypoliths** are photosynthetic organisms that live underneath rocks in arid environments. Hypolithic colonization is protected by overlying mineral substrate from incident UV radiation and excessive photosynthetically active radiation (PAR) [148-150]. It has been reported that cyanobacterial hypoliths occurred on quartz in major deserts spanning every continent on Earth [151]. The *Actinobacteria*, *Alphaproteobacteria* and *Gammaproteobacteria* are ubiquitous in all hypoliths, and photoautotrophic cyanobacteria are more abundant in hypoliths of cold deserts [141,150,152]. *Deinococcus* appear to be more abundant in warm and hot deserts but not common in polar hypoliths [76]. Few studies on multi-domain diversity of hypolithic microorganisms in deserts have been performed [153]. A study using quantitative PCR estimated the entire hypolithic communities in Tibetan and Antarctic Dry Valleys, and revealed that eukaryotic and archaeal taxa comprise less than 5% of recoverable phylotypes.
**Endoliths** are organisms (archaea, bacteria, fungi, lichen, algae) that live inside rocks, or in the pores between mineral grains of a rock. Rocks such as sandstone, limestone and weathered granite are normal habitats for endoliths in deserts. Endolithic colonization is widely observed among deserts of all aridity classes, indicating the advantage of adaptation to drought conditions.

**Bioaerosols** are suspensions of airborne particles that contain living organisms. They can disperse with the transportation of desert dust over intercontinental distances. Airborne microorganisms, such as spore-forming bacteria and mitosporic fungi in bioaerosols, pose large risks to human and animal health, with long distance traveling and potential pathogenic abilities.

Microbial communities and their functional structure in deserts are not well studied, and there are many unanswered questions regarding their biology, physiology and ecology. With the development of high-throughput technology, more work can be done using various omics methods such as genomics, proteomics or metabolomics, to fully understanding the microorganisms in deserts.
1.3.3. Bacterial Diversity

Desert biomes have been shown to be remarkably different from other biomes in terms of soil microbial community composition and function [154]. Traditional culture-dependent methods for identification of bacterial diversity can only reveal a very small fraction of the actual bacteria in a community, and the percentage varies from 0.0001% to 15% in different environments [59]. Modern environmental microbiology has been greatly enhanced by the application of molecular genetic technology, which allows the examination of microbial communities through analyses of microbial DNA, RNA and proteins. Many studies have been performed on the bacterial community of soils in arid environments using next generation high-throughput sequencing methods.

The known bacterial diversity on Earth includes approximately 12000 different species (http://www.bacterio.net/-number.html). The unknown bacterial diversity is currently explored using molecular microbiology techniques, and many new bacterial taxa are being submitted to GenBank each year. Early research in the Atacama Desert based on cell numbers reported that cultivable heterotrophic bacteria are present in the less arid region of the Atacama Desert at levels of $10^7$ colony-forming units (CFU) per gram of soil, while only present between $10^2$ to $10^4$ CFU/g of soil in the desert’s core regions [155]. Populations of aerobic bacteria in deserts across the world are reported to vary from < 10/g in the Atacama desert to $1.6 	imes 10^7$/g in soils of Nevada. Sand dunes from the Thar are reported to have a relatively smaller population ($1.5 	imes 10^2$ – $5 	imes 10^4$ /g soil) [156]. Gram-positive spore formers are dominant and the populations do not decline significantly even during summer, and *Actinomycetes* may constitute ~50% of the total microbial bacterial population in desert soils [157].

Using PCR-based biological molecular techniques, microbial research can focus more precisely on the diversity of bacterial communities and their dominant microbes.
Pointing et al. (2008) reported that in the Namib Desert, the majority of 16S rDNA sequences displayed more than 94% homology to members of the Firmicutes (particularly to members of the genus Bacillus), and bacteria belonging to the Bacteroidetes, Planctomycetes, Chloroflexi, and Betaproteobacteria groups were also observed [158]. Connon et al. (2007) found that in the soil of Antarctica, dominant phylotypes were affiliated to the phyla Acidobacteria, Actinobacteria, Bacteroidetes, Proteobacteria, Deinococcus-Thermus, Firmicutes, Cyanobacteria and TM7 [159]. An et al. (2013) studied the bacterial communities in samples from the two largest deserts in Asia, the Taklamaken and Gobi deserts, and found the most dominant phyla are Firmicutes, Proteobacteria, Bacteroidetes and Actinobacteria [160].

Photosynthetic cyanobacteria can be the primary inhabitants in arid environments, and typically live a few millimeters below the surface of translucent rocks, such as quartz, sandstone pebbles, halite and gypsum [161,162]. The capacity to benefit from a sufficient supply of CO₂, N₂ and light to allow photosynthesis and N₂ fixation while being protected from desert-like conditions (high radiation, desiccation, salt stress, etc.), allow these phototrophic communities to be prevalent in deserts. The distribution of cyanobacterial communities in desert pavements present more frequently in the form of patches, and their spatial distribution pattern in different sites are correlated with mean annual precipitation and temperatures [161]. Heterotrophic bacteria also occur widely in desert environments, including Alphaproteobacteria, Actinobacteria, Flexibacteria, Firmicutes, Gemmatimonadetes, Planctomycetes and Deinococcus-Thermus [163]. Members of the CFB group (Cytophaga-Flavobacterium-Bacteroides) were found to be dominant in the hot desert of Tataouine of south Tunisia [164], and in the hyper-arid Taklamakan Desert in China.

Many studies show that members of the Actinobacteria can be dominant in arid environments, and the subclass Actinobacteridae are often found to prevail in desert
soil [165]. Members of the genera *Rubrobacter, Arthrobacter, Thermopolyspora* and *Streptomyces* have been found in both hot and cold deserts [166]. Members of the *Actinobacteria* have wide metabolic and sporulation capacities, as well as multiple UV repair abilities [167].

*Bacteriodes* are also well represented in desert soils. Prestel et al. (2013) reported that samples from Death Valley soils showed a number of phylotypes with high homology to members of the *Flavobacteriales* and to the genus *Adhaeribacter* of the class *Sphingobacteria* [168]. In the Taklamakan Desert, an abundance of *Pontibacter* from the family *Cytophagaceae* were observed [169].

*Proteobacteria* are globally distributed and were thought to be prominent members of desert soil bacterial communities [170]. *Alpha-, Beta- and Gammaproteobacteria* are often linked to soils with higher rates of organic carbon inputs (Lopez et al. 2013). However, several studies have shown that *Proteobacteria* may be functionally important in nutrient-limited arid environments, since some members in this phylum are capable of photosynthesis [171].

Members of the phylum *Gemmatimonadetes* and *Firmicutes* are also widely observed in desert soils, and may be comparatively more abundant than in other biomes [154]. Some genera in the *Firmicutes* phylum, such as *Bacillus* and *Paenibacillus*, are able to form endospores that can facilitate survival under desiccation conditions. Some aerobic taxa in this phylum are characterized by rapid spore germination, non-fastidious growth requirements and short doubling times, which are a nice fit for desert environments.
1.3.4. Bacteria in Desert-like Environments and Bacteria involved in Weathering

Desert-like soil surface features such as desert pavements, surface accumulation of salt, calcium carbonate accumulation, and surface exposure of gypsum materials are manifestations of some kind of land deterioration [172]. This kind of land surface is commonly seen in semi-arid regions outside desert boundaries.

Mineral soil texture in desert-like condition is commonly sandy loam to loam sand. Sandy soils are formed by the weathering of the Earth's surface. Sand is the largest of all soil particle types and more spread apart than the particles of organic or clay soils, and can rarely retain surface water, resulting in less vegetation cover for surface protection. Sandy soils are formed from rock such as shale, granite, quartz and limestone. Sand allows air to freely circulate around it.

Desert-like sandy soils are presented as coastal sand soils, sandy loam in the forests and grassland, and local soil properties shape the bacterial diversity and communities. Coastal sandy soils in general lack three macronutrients: nitrogen, phosphorous and potassium [173], which is reported as the primary limit for growth of vegetation in this soil environment.

Russo et al. (2012) studied the bacterial communities in forests with sandy loam soil textures in the Lambir Hills National Park of Malaysia, which is sandstone-derived, nutrient-depleted and well-drained, and showed that *Proteobacteria* were dominant in sandy loam, while *Acidobacteria* were the most abundant group in clay [174]. They reported that *Actinobacteria, Betaproteobacteria, Clostridia, Bacilli* and *Gammaproteobacteria* were more abundant in sandy loam than in clay. Halliday et al. (2014) studied the bacterial community of dry beach sand in Avalon Bay Beach
(Catalina Island, USA), and reported that the phyla Proteobacteria, Bacteroidetes, Actinobacteria, Planctomycetes and Acidobacteria are dominant [175]. This study also claimed that bacterial communities of beach sand are broadly similar to soil communities at the phylum level and strongly influenced by soil pH and temperature. McHugh et al. (2014) reported that in the semi-arid grasslands of Arizona and New Mexico, bacterial communities were dominated by members the phyla Actinobacteria (53 %), Proteobacteria (16 %), and Acidobacteria (8.7 %) [176].

Microorganisms colonized on the surfaces of mineral soils contribute to precipitation of new minerals and to carbonate production, which plays an important role in the soil environment by contributing to the release of key nutrients. Several bacterial strains from different genera have been found to have mineral-weathering abilities, such as Anabaena, Bradyrhizobium, Burkholderia and Collimonas [177]. It has been reported that surface soil mineral particles appear to be inhabited by different communities: in limestone, the endolithic bacterial communities are comprised of Gram-positive bacteria and Acidobacteria, while the epilithic population are ~50% Proteobacteria [126]. Different primary minerals, such as granite, limestone, apatite, plagioclase, and quartz are colonized by different bacteria [178,179]. Studies of bacterial communities in soils with different mineral composition showed that concentrations of major elements, such as aluminium and calcium, seem to have a significant impact on the structure of the bacterial community [180].
1.4. Relationship of Environmental Factors and Soil Bacterial Diversity

1.4.1. General introduction to the relationships of environmental factors and soil microbial diversity

Soil is a highly complex and important biome possesses immense bio-diversity and a large number of biological processes [181]. Bacteria constitute the largest portion of the biodiversity in soils, and play an important role in maintaining soil processes, which eventually affects the functioning of terrestrial ecosystems [182]. Richness and patterns of microbial diversity are affected by different environmental factors. Recent improvements in techniques have brought a revolution in our understanding of microbial diversity, and allow us to survey the diversity of microbial communities. Many studies have been performed to explore how the changes of environmental physiochemical parameters can affect the microbial diversity and community structures [183]. Several factors, such as pH, total carbon, organic materials, are reported to be able to shape the local microbial community.

Previous studies using isolates from soils based on pure cultures have revealed bacterial diversity within defined isolated taxa [184]. However, the taxa detected by culturing are known to not reflect all the taxa in an environment [185], and molecular methods based on diversity profiling of environmental metagenomics are more useful in the study of bacterial communities. Studies of microbial biogeography using NGS can often provide key insights into environmental tolerances and community structure of microbial taxa, particularly those difficult-to culture taxa that often dominate in natural environments [186].

Soil microbial communities are largely diverse and present geographic and
environmental specificity in the prevalence of different bacterial groups [187]. It is now widely accepted that soil microbial communities are significantly affected by environmental factors at different geographic scales [94]. Geographic location and physicochemical properties are believed to be major factors that affect soil bacterial diversity and community structures [90]. However, it was suggested that when soils are characterized by distinct environment factors, each will likely inhabit a unique microbial community, regardless of the geographic distance between them [74].

A study of microbial community structure from soil samples of the hyper-arid core in the Atacama desert indicated that salt content and water availability significantly correlated with the diversity of microbial communities. Nitrogen can be a limiting factor for biomass production and biological activities in soils of arid environments [75].

Andrew et al. (2012) reported that in Sonoran desert soils, microbial communities are shaped primarily by soil characteristics associated with geographic locations, while rhizosphere associations are secondary factors [187]. Similarly, a study across Israel and the United States found that the bacterial community structures in the phyllosphere of Tamarix trees are driven by climate instead of specific rhizosphere factors [188].

Lozupone et al. (2007) argued that the main environmental determinant of microbial diversity is salinity, rather than extremes of temperature, pH, or other physical and chemical factors represented in their samples [189]. El Hidri et al. (2013) Reported a huge phenotypic and phylogenetic diversity observed in arid and saline systems in southern Tunisia. This study indicated that extremely halotolerant bacteria were the most dominant group and were affiliated to Bacillus, Nesterenkonia, Salinicoccus, and Marinococcus genera [190].
1.4.2. Environmental factors in Deserts

An ubiquitous feature for deserts is the scant, low precipitation level, and they are also characterized by extreme fluctuations in temperature, generally low nutrient status, high levels of incident ultraviolet (UV) radiation and strong winds [170,191]. It has been demonstrated that deterministic factors drive bacterial community structure processes at both global and local scales [192,193]. Here, I summarize some environmental factors that have been reported as important drivers for bacterial communities in desert soil:

1) **Precipitation.** Lack of precipitation is the major feature of arid land, and water availability is considered to be a key limiting factor for all living things. Species richness declines with increasing aridity in deserts [152]. In the most extreme hyper-arid environment, microbial life retreats to isolated “oases” (sheltered niches or under rocks) that are formed as a result of biotic-abiotic relationships between the microorganisms and the available porous and deliquescent mineral substrates, permitting life in landscapes as a refuge from the great extinction pressures. The impact of precipitation on microbial diversity has been reported in many studies of deserts [194]. Some studies claimed that water availability is the primary controlling factor for microbial activity, diversity and community [195]. Gillor et al. (2010) performed bacterial diversity analysis along a large scale of a precipitation gradient using quantitative PCR and terminal restriction fragment length polymorphism (T-RFLP) and showed that, although soil bacterial abundance decreases with precipitation, bacterial diversity is independent of a precipitation gradient, and community composition is unique to each ecosystem [196].

2) **pH.** The effect of pH values are often considered as a main driver for bacterial communities in soil. Lauber et al. (2009) examined bacterial communities in 88
soils from across North and South America using high-throughput sequencing of PCR amplified 16S rDNA genes to investigate the relationships between the pattern of bacterial communities and their environmental factors. These contained ten samples from arid regions such as the Mojave desert, with pH values ranging from 7 to 8.5 [186]. Their results showed that these desert soils possess unique bacterial community structures that are distinguishable from other soil samples, and could be explained by their higher pH values. They also found that overall bacterial community composition was significantly correlated with differences in soil pH, largely driven by changes in the relative abundance of Acidobacteria, Actinobacteria, and Bacteroidetes across the range of soil pH examined. Other studies of microbial diversity in deserts soils illustrated that the impact of soil pH in arid environments may not be as import as moisture or radiation in deserts [197].

3) **UV Radiation.** UV Radiation in the deserts is markedly higher than in other biomes as a result of the low levels of atmospheric water vapor and tree cover, and it is an import factor limiting surface soil biodiversity. The flux of UVA and UVB is currently substantially higher on the Antarctic continent than elsewhere on Earth because of the polar ozone hole depletion and longer day lengths during the austral summer. The lithic environment can provide protection against UV exposure. A high diversity of bacteria associated with hypoliths has been reported in the Antarctic Dry Valleys, located in the polar region and exposed to high summer radiation levels [149].

4) **Salinity.** Saline soils are characterised by high concentrations of salts and by an uneven temporal and spatial water distribution [198], which is commonly found in deserts. The ions responsible for salination are Na⁺, K⁺, Ca²⁺, Mg²⁺ and Cl⁻. A high concentration of salt in soil changes the availability of water and nutrients for microorganisms, and can influent on the size and the activity of soil microbial
biomass. Pointing et al. (2012) suggests that soil salinity is a major factor in community structure in desert ecosystems, and even modest levels of soil salinity may be an important determinant because biological water availability is reduced in the presence of soluble salts [126]. Aislabie et al. (2009) indicate that Actinobacteria and Bacteroidetes were more prevalent in dry alkaline soils and Gammaproteobacteria in dry saline soils [199].

5) **Temperature.** Extreme fluctuations in temperature and variations in diurnal and seasonal cycles also present major challenges in arid lands. Direct stresses are imposed by heat and cold shock and by freeze-thaw cycles, which are common in both hot and cold deserts [200]. In a comparative study of hot and cold deserts in China, it was reported that the presence of specific lineages of Deinococci may be related to mean annual temperature [152].

6) **Substrate mineralogy.** Soil is composed of a mixture of minerals that differ in element and organic compounds. Studies of environments indicate that different minerals in soil may select distinct bacterial communities in their microhabitats [201].

7) **Nutrient availability.** Nitrogen is often regarded as a limiting nutrient in oligotrophic environment like deserts, and it has been reported that microorganism abundances were positively correlated with nitrogen levels. However, Andrew et al. (2012) reported that available carbon, not nitrogen, is a limiting factor in driving local microbial diversity in the Sonoran Desert soils using pyrosequencing methods [90].

8) **Biotic Factors.** Plant communities affect the diversity of soil microorganisms locally through interactions within the rhizosphere, and microbial communities are directly affected by the microenvironment of plant root systems [202]. Early
studies have shown that resource abundance near vegetation can control the heterotrophic bacterial numbers in desert soils [203]. It has been reported that the microbial community structure of desert soils changes with agricultural activities, and long-term farming can induce a drastic shift in the bacterial communities in desert soil [194]. Bacterial communities in agricultural soil showed a higher diversity and a better ecosystem function for plant health, but with a reduction of extremophilic bacteria.

9) Other Factors. Many factors other than the above mentioned can affect the microbial diversity in deserts, such as soil geomorphology, geographic distance, altitude, concentration of heavy metals and etc. Sessitsch et al. (2001) claimed that the soil particle size is one of the abiotic drivers for microbial abundance [204]. Finkel et al. (2012) indicated that in Sonoran Desert soil, the geographic distance across different sample sites was the most important parameter which affected community composition, particularly that of Betaproteobacteria [188]. Seasonal climate factors such as monsoon precipitation in semiarid zones can significantly influence the bacterial population. For example, McHugh et al. (2015) reported that in the semi-arid zones of Arizona and New Mexico, the Firmicutes phylum experienced over a six-fold increase in relative abundance with increasing in response to monsoon precipitation. Conversely, Actinobacteria, the dominant taxa at the site, were reversely correlated to moisture availability [176].
1.4.3. Adaptation of Bacteria to Arid Environments

Microorganisms can adapt to environmental variations much faster than multi-cellular organisms. They are pioneer colonizers that have profound influence on the climate and environments on Earth [205]. Remarkable phylogenetic and metabolic diversity of bacteria, with the ability to develop biofilms, make them adapt and colonize extreme environments not tolerated by other organisms [206]. To survive in the deserts, an efficient metabolic stress response during growth and the ability to transition between active and dormant states are necessary to microorganisms.

It is well known that cells present stress tolerance strategies to avoid moisture, thermal and ultraviolet (UV) stress, but ecological studies are now revealing that bacteria exhibiting adaptation at the community level are also critical to the colonization of desert environments [207]. Desiccation tolerance is mediated both intracellularly by UV-absorbing compounds, and extracellularly by cell walls or polymeric substances, which is an indirect result of coping with stresses, such as osmotic, temperature, and oxygenic stresses [208].

Microorganisms in surface soils of deserts are directly exposed to high environmental stress levels, and possess a range of strategies against potential UV damage such as screening by pigments and damage-repair mechanisms [209]. Extreme ionizing-radiation resistance has been observed in several members of the domains Bacteria and Archaea, among which Deinococcus and Rubrobacter show the highest levels of resistance [210].

Huang et al. (2015) studied the diversity of the radiation-resistant microbes of the hyper-arid Taklamakan desert, and reported that radiation-resistant phylotypes belonged to the genera Knoellia, Lysobacter, Nocardiodes, Paracoccus, Pontibacter, Rufibacter and Microvirga [211]. Cyanobacteria are well presented in a range of hot
and cold deserts, which are important in biogeochemical cycling processes such as C or N utilization and stress response [74]. Studies found that members of the *Cyanobacteria* can maintain photosynthetic metabolism in desert-like conditions such as high radiation, desiccation, and salt stress [212].
1.5. Objective of Thesis

It is now widely accepted that measures must be taken to estimate, record, and conserve microbial diversity, not only to sustain human health but also to enrich the human condition globally through the wise use and conservation of genetic resources of the microbial world [213]. Desert environments, because of the range and severity of environmental factors, are an obvious target for fundamental research on the ecological and evolutionary processes which structure biological communities. They are also important because of the unique species that can occur there.

Arid and semi-arid ecosystems are important to study the potential impact of desertification on the microbial activity and communities structure of arid land [193]. By using NGS technology to study microbial communities, researchers have gained a new appreciation for the dynamics of microbial diversity in specific habitats, the variability of microbial diversity in spatial and temporal scale, and the environmental factors driving this variability [20].

In my thesis research, I used pyrosequencing of 16S rDNA amplicons to analyse the bacterial diversity and communities in the arid regions of San Rafeal Swell of Utah State, as well as a desert-like tourist site – the Desert of Maine, in the USA. We also examined some physicochemical parameters of sample sites to investigate the correlations between bacterial community structure and environmental drivers.
CHAPTER 2: Bacterial Communities of the Desert of Maine
Bacterial communities in the mineral sandy soil from the Desert of Maine (USA) environment

Yang Wang, Jorge R. Osman, Michael S. DuBow*

Laboratoire de Génomique et Biodiversité Microbienne des Biofilms, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Univ. Paris-Sud, Bâtiment 409, 91405 Orsay, France

*Address for correspondence: Laboratoire de Génomique et Biodiversité Microbienne des Biofilms, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Univ. Paris-Sud, Bâtiment 409, 91405 Orsay, France; Tel :+33(0)169154612 ; Fax :+33(0)169157808 ; E-mail : michael.dubow@u-psud.fr

Key Words: Mineral Soil, Bacteria, Biodiversity, 16s rDNA, Pyrosequencing

Running Head: Bacterial Communities of the Desert of Maine
Abstract

The Desert of Maine is a tract of glacial silt, surrounded by a pine forest, in the state of Maine located in the northeastern USA. The soil of the Desert of Maine has a sandy texture with poor water holding abilities, nutrient retention capabilities and a relatively low pH value (pH 5.09). Samples from this site thus present an interesting place to examine the bacterial diversity in mineral sandy loam soils with an acidic pH and low concentrations of organic materials. Two surface sand samples from the Desert of Maine were obtained, and pyrosequencing of PCR amplified 16S rDNA genes from total extracted DNA was used to assess bacterial diversity, community structure and the relative abundance of major bacterial taxa. We found that the soil samples from the Desert of Maine showed high levels of bacterial diversity, with a predominance of members belonging to the Proteobacteria and Actinobacteria phyla. Bacteria from the most abundant genus, Acidiphilium, represent 12.5% of the total 16S rDNA sequences. In total, 1394 OTUs were observed in the two samples, with the number of common OTUs observed in both samples being 668. By comparing our bacterial population results with studies on similar soil environments, we found that the samples contained less Acidobacteria than soils from acid soil forests, and less Firmicutes plus more Proteobacteria than soils from oligotrophic deserts.

Key Words: Mineral Soil, Bacteria, Biodiversity, 16s rDNA, Pyrosequencing
Introduction

Bacteria are integral and diverse components of soil, where their community structure and diversity have been found to be linked to many soil environmental characteristics, such as the physical and chemical properties of the soil (Lauber et al. 2009; Fierer et al. 2012a). Traditional microbial cultivation techniques frequently overlook the majority of microbes present in a sample (Amann et al. 1995), as most bacteria cannot be cultivated under laboratory conditions. Many recent studies have used high-throughput PCR amplified 16S rDNA sequencing to overcome this difficulty to identify the members of a prokaryotic community. The 16S rRNA gene is by far the most widely used genetic marker for phylogenetic and microbial community studies, as it has highly conserved regions that permit effective PCR primer design, and sufficient variable regions to allow for accurate taxonomic and phylogenetic identification of community members. Since this gene has been widely sequenced in microbial diversity surveys, there is a large amount of accumulated 16s rDNA sequence data in databases (Petrosino et al. 2009) such as the Greengenes, Silva and RDP databases. The 16S rDNA gene in bacteria includes a total of 9 hypervariable regions (V1–V9), and the V1-V3 regions have been shown to be effective for bacterial identification (Schloss 2010).

The Desert of Maine is a tract of glacial silt with a surface area of 160,000 m², surrounded by a pine forest, in southern Maine in the northeastern USA. It is not a true desert, as it receives an abundance of precipitation (76-120 cm/year), with a mean annual air temperature of 6-7 °C (http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm, Natural Resources Conservation Service). Although it is a tourist attraction, there are no imported sand nor designer landscaped dunes. This surface area was formed approximately 11,000 years ago, during the end of the last Ice Age of the Pleistocene Period (Bahr and Friedman 2009). The parent material of the soil is sandy glaciofluvial deposits derived
from granite and gneiss. Soil and ground rocks were slowly scraped by glaciers into a sandy substance, forming a layer up to 25 meters deep. Then, over many centuries, surface soils formed a cap, concealing the “desert sand”, and allowing a forest to grow, followed by the subsequent development of agriculture. The glacial “desert” was once covered by a farm, and exposed because of severe soil erosion due to crop rotation mismanagement (http://www.desertofmaine.com).

The soil of the Desert of Maine has a sandy texture with poor water holding abilities, nutrient conservation capabilities and an acidic pH value. Mineral sandy loam soils contain less organic materials, with a basic pH (Griffiths et al. 2011; Crits-Christoph et al. 2013), while low pH soils generally contain more organic materials, such as the soils of forests and some grasslands (Nacke et al. 2011; Shah et al. 2011; Russo et al. 2012). The Desert of Maine presents an interesting example to observe the bacterial populations in a mineral sandy loam with a relatively low pH and concentration of organic materials. Here, we used pyrosequencing of PCR amplified 16S rDNA genes to assess bacterial diversity, community structure and the relative abundance of bacterial taxa within two sites of the surface soil of the Desert of Maine, and compared its bacterial community with those from several other sandy desert environments, as well as from other mineral soils.
Materials and Methods

**Sampling.** Two surface sand samples from the Desert of Maine (Fig. 1) were obtained on September 2, 2011. The two samples were collected by scooping surface sand into 50 ml sterile polyethylene conical centrifuge tubes, in an area cordoned off from tourists and without any measurable rainfall for at least four days. The average air temperature during the week of sampling was 22°C. After collection, the samples were treated as previously described (An et al. 2013). To perform the analyses of selected physicochemical parameters of the sample site soil, samples from the two sites were pooled and sent for analyses using standard methods to the Laboratoire d’Analyses de Sols (Institut National de la Recherche Agronomique, France).

**Sample Preparation.** Total DNA was extracted from each sample using the protocol of An et al. (2013). An aliquot of extracted DNA was adjusted to a final DNA concentration of 15 ng/μl in 1/10 TE buffer (1 mM Tris pH 8; 0.1 mM EDTA) using a NanoVue spectrophotometer (GE Healthcare, Little Chalfont, Buckinghamshire, UK), and the concentration verified by ethidium bromide fluorescence after electrophoresis through a 1% agarose gel in TAE buffer (2 mM Tris-acetate pH 8; 5 mM Na-EDTA). PCR reactions were performed in 25 μl reaction volumes. Each reaction contained one of two different thermostable DNA polymerases and their corresponding reaction buffers, 200 μM of each dNTP, 0.5 μM of each primer and 1 to 10 ng of extracted DNA. The 16S rRNA genes were amplified using the universal bacterial primers for pyrosequencing and covering hypervariable regions V1-V3: primer 27F (A adaptor + GAGTTTGATCMTGGCTCAG) and primer 518R (B adaptor + Mid + WTTACCGCGGCTGCTGG), where A and B represent the adaptors using the 454 Roche FLX Titanium pyrosequencing reaction platform. The Mid sequences are eight nucleotide tags designed for sample identification barcoding according to the 454 protocol. PCR amplification conditions were adapted for the use of two different thermostable DNA polymerases: A) Phusion High-Fidelity DNA Polymerase
(Finnzymes, Finland): 98°C for 2 min, followed by 28 cycles of 98°C for 30 s, 54°C for 20 s and 72°C for 15 s, and a final elongation step at 72°C for 5 min; B) Pfu DNA Polymerase (Fermentas, Canada): 95°C for 2 min, followed by 30 cycles of 95°C for 30 s, 48°C for 30 s and 72°C for 1 min, and a final elongation step at 72°C for 5 min. Each DNA sample was subjected to 3-5 different PCR reactions per DNA polymerase to minimize PCR bias. The PCR products were pooled and subjected to electrophoresis through a 1% agarose gel in TAE buffer. After electrophoresis and visualization of the PCR products by ethidium bromide staining and long wave UV light illumination, NucleoSpin Extract II kits (Macherey-Nagel, Germany) were used to purify the 16S rDNA PCR products. Then, 40 ng of PCR products from each sample were mixed for pyrosequencing, performed using a 454 Roche FLX Titanium Pyrosequencer (Microsynth AG, Switzerland).

**DNA Sequence Data Processing.** The raw DNA sequences were first assigned to each sample via their Mid tag using MOTHUR version 1.33 (Schloss et al. 2009), and reads were removed if at least one of the following criteria was met: (i) length less than 200 nt or longer than 600 nt, (ii) mismatch to the barcode sequences or more than one mismatch to the primer, and (iii) the presence of homopolymers of > 8 bp in length. Adaptor sequences were removed from the sequences using the “Cutadapt” tool (Martin 2011) implemented in the Galaxy server of the Institut de Génétique et Microbiologie (IGM) of the Université Paris-Sud (http://galaxy.igmo rs.u-psud.fr). Then, the sequences were checked for quality scores by ConDeTri version 2.2 (Smeds and Kunstner 2011), using the criteria that 80% of the nucleotides in a sequence have quality scores > 25. We used UCHIME (Edgar et al. 2011), with reference database Greengenes version 2013_May, and Decipher (through web tools available at http://decipher.cee.wisc.edu/) to detect chimera sequences (Wright et al. 2012). Sequences detected as chimeras by both programs were removed from the data sets. The raw sequences have been deposited in the GenBank short-read archive (SRA), with accession number SRP056525.
Taxonomy assignments of the remaining 16S rDNA reads were conducted using RDP II classifier with a confidence threshold cutoff of 80, using the Silva database release 119 reformatted in MOTHUR (database available on the download page at the website of MOTHUR) (Cole et al. 2014; Yilmaz et al. 2014). Sequences classified as Chloroplast, or that could not be classified as belonging to the Bacteria Kingdom, were removed. Diversity analyses were performed using the software package MOTHUR. The relations of the relative abundance of bacterial groups at different taxonomy levels between the two samples were calculated using the Pearson correlation coefficient measure with SPSS Statistics (Version 22). The clean reads were clustered into operational taxonomic units (OTUs) using UPARSE at a cutoff value of 97% sequence identity (Edgar 2010; Edgar 2013). The Chao1 and Shannon indices were calculated to estimate taxon richness and diversity (Schloss and Handelsman 2008). The significance of differences between two bacterial communities was calculated using Libshuff implemented in MOTHUR (Schloss et al. 2004), with 1500 randomly-selected sequences from each sample selected using PANGEA (Giongo et al. 2010).
Results

Chemical and Physical Properties of the Sand Samples

Two areas, separated by 3 m, of the Desert of Maine were sampled on September 2, 2011. The mean chemical and physical properties of the sand samples are shown in Table 1. The mean pH values of the soil at the sampling site was 5.09, indicating an acid soil environment. The levels of total organic carbon and organic material were less than 1 g/kg soil.

Diversity Analyses

The average length of the raw DNA sequences for the two samples were 479 nt (Maine 1) and 480 nt (Maine 2), respectively, while the total number of reads for each sample were 23,405 (Maine 1) and 28,983 (Maine 2), respectively. After bioinformatic cleaning, approximately 95% of the sequences remained (22,320 for Maine 1 and 27,680 for Maine 2). The sequences were further filtered for quality and examined for chimeric sequences, leaving 65% of the total reads (14,776 for Maine 1 and 19,085 for Maine 2). The average length of the sequences after processing were 396 nt for the two samples. The number of reads remaining after each step are presented in Table 2.

The clean sequences were clustered into OTUs at 97% similarity levels, excluding the unclassified sequences at the phylum level, and sequences classified as Chloroplast were removed. In total, 1394 OTUs were observed in the two samples, and the number of common OTUs observed in both samples was 668. The numbers of core taxons (most abundant OTUs with more than 1% of the sequences), were 18 in the Maine 1 sample and 14 in the Maine 2 sample. The core taxons comprised 30% of the bacterial population in the Maine 1 sample and 23% in the Maine 2 sample. We observed no differences between the Shannon diversity indices of the two samples (Table 2).
Classification of DNA Sequences

The sequences of the two samples were classified at 6 taxonomic levels with RDP Classifier, and comprise at least 22 phyla, 41 classes, 76 orders, 115 families and 172 genera, plus a number of unclassified sequences at various taxonomic levels. The distribution of sequences at the phylum level is shown in Fig. 2. Unclassified sequences at the phylum level represent 7.0% of the sequences in Maine 1 and 6.4% in Maine 2. The community structure at the phylum level shows a similar distribution for the two samples (R=0.989 with the Pearson coefficient measure). The significance of the differences between the two bacterial communities was calculated using the Libshuff package in MOTHUR, with a P value < 0.01, indicating that the two bacterial communities do have not the same composition. The predominant phyla of the samples represent members of the *Proteobacteria, Actinobacteria, Chloroflexi, Bacteroidetes*, and *Acidobacteria* phyla, at 39.1%, 18.5%, 10.3%, 6.4%, and 5%, respectively as the average for the two samples, followed by members of the *Cyanobacteria* (3.4%), *Planctomycetes* (2.5%), *Gemmatimonadetes* (2.4%), WD272 (2.0%), *Armatimonadetes* (1.9%), TM7 (0.5%), and *Deinococcus-Thermus* (0.5%) phyla. There are 10 phyla that each comprise < 0.5% of bacterial sequences (*Verrucomicrobia, Chlorobi, Nitrospira, Elusimicrobia, SM2F11, TM6, WCBH1-60, OP11, SHA-109 and Firmicutes*), and these are grouped together as “rare phyla”. The percentage in the two samples assigned to each of these low abundance phyla are shown in Table 3.

Within the phylum *Proteobacteria*, members of the *Alphaproteobacteria* represent the most abundant group in both samples (67.8 % in Maine 1, 63.3% in Maine 2), followed by members of the *Betaproteobacteria* (15.2%, 16.8%), and *Deltaproteobacteria* (12.9%, 14.9%). The bacterial community structure at the Class level (Fig. 1) presents a similar pattern of distribution in the two samples (R=0.959 with the Pearson correlation coefficient measure). The members of the *Sphingobacteriales* family were the predominant members of the *Bacteroidetes*
phylum, with an average of 4.8 % of the total reads for the two samples. Members representing Rhodospirillales and Burkholderiales were the most abundant groups within the Alphaproteobacteria and Betaproteobacteria classes, respectively. The dominant Families in the two samples come from the Acetobacteraceae (16.8 %), Chitinophagaceae (4.6 %) and Oxalobacteraceae (4.3 %).

Among the 172 genera identified in the samples, 69 genera belong to the phylum Proteobacteria, 41 genera belong to the phylum Actinobacteria, 24 genera belong to the phylum Bacteroidetes, and 9 genera belong to the phylum Acidobacteria. The most abundant 20 genera in each sample are shown in Table 4. The similarity between the two bacterial communities at the genus level is 0.975 using the Pearson correlation coefficient measure. Each of these abundant genera account for 0.6%-11.3% of bacteria identified in the Maine 1 sample and 0.8 %-13.6 % in the Maine 2 sample, with the most abundant genus being from Acidiphilium (in the phylum Proteobacteria) for both samples. Members of the genus Crinalium (1.9% in average for our two samples) represent a group of phototrophic bacteria belonging to the phylum Cyanobacteria, and species of this genus have been reported to be highly drought-resistant and to be isolated from coastal sand dunes (Ben de Winder et al. 1990; Ben de Winder and Mur 1994). Members of the genus Arthrobacter are abundant in both samples (1.4% in average for our two samples), and members of this genus are frequently involved in mineral weathering of soil and can secrete large amounts of oxalic acid (Uroz et al. 2009; Frey et al. 2010).
Discussion

In recent years, desert-related environmental effects have been increasing, as global warming and human activities contributing to desertification are increasingly threatening ecosystems around the world. Studies concerning microbial colonization and dispersion in deserts have been performed to estimate the function of microbial communities from desert sand which may play an important role in soil stability, nutrient cycles and environmental health. The Desert of Maine was previously productive agriculturally, and was covered by farm land. The mineral soils were exposed because of severe soil erosion due to crop rotation mismanagement. In this study, we used pyrosequencing of PCR amplified 16S rDNA gene to assess bacterial diversity and community structure of surface soil of the Desert of Maine. An examination of bacterial populations in samples of surface soil from this unique site provides an opportunity to investigate bacterial diversity and community structure in a desert-like environment, and its relation with those from other soils.

Previous studies have revealed that many environmental factors, including pH, the concentration of organic material and that of sodium, can have large effects on the presence and distribution of bacterial community members in soil (Rousk et al. 2010; Griffiths et al. 2011; Centeno et al. 2012). A large proportion of the variance in soil bacterial diversity and community composition appears to be strongly influenced by pH (Fierer et al. 2012b), at local (Osborne et al. 2011) and even continental scales (Lauber et al. 2009). In this study, we compared our data on the distribution of predominant phyla (Fig. 3) with those from studies using 16S rDNA gene pyrosequencing of samples taken from apparently similar soil environments or with similar physiochemical factors. The Desert of Maine is surrounded by a pine forest. Thus, its soil microbiome may be influenced by that of the pine forest environment around it. Uroz et al. (2010) studied the bacterial community of soils from an oak forest of Breuil-Chenue in Morvan (France), and found that members of
Acidobacteria account for about 36% of the total bacterial population in their sample. Shah et al. (2011) examined the sandy, acidic and nutrient-poor soil of a pine barrens region of Long Island (New York, USA), which is also composed of gravel deposited by the withdrawal of glaciers. This soil has a pH value of 4.9, total organic carbon of 20.9 g/kg, and Al and Fe concentration of approximately 0.1 g/100g. Other samples of surface soil from hot or cold deserts were compared with our results: 1) a sand sample (Gobi 1) from the Gobi Desert of Northwestern China (An et al. 2013), which also has a low concentration of organic carbon and organic materials (< 1 g/kg); 2) a sample (Altamira) from the Atacama Desert in Chile (Crits-Christoph et al. 2013); 3) a sample (Upper Wright Valley) from McMurdo Dry Valleys (Lee et al. 2012) in the area of the Antarctic continent (a cold desert). All these studies were performed using pyrosequencing of 16S rDNA amplicons. We used the Pearson correlation coefficient measure to estimate the similarity of the distribution of the predominant phyla among the different sites. The results showed that the distribution of the predominant phyla in both samples from Maine are more close to samples from the two forest soils (R values > 0.7), than to samples from both hot and cold deserts (R values < 0.5). For the Phylum Probeobacteria, there is no significant difference between our samples and these from the other two forest samples, but the relative abundance of this phylum in our samples is greater than that of the desert samples (P < 0.05). The Gobi Desert sample, however, contained a much higher proportion of members of the Firmicutes than the others. Samples from the Atacama Desert and the McMurdo Dry Valleys contain a larger population of Actinoacteria members when compared with other groups (P < 0.05). There is no significant difference in the percentage of Bacteroidetes among the different samples examined here. These results confirm those of earlier studies (Griffiths et al. 2011; Tiao et al. 2012; Crits-Christoph et al. 2013) that suggested that oligotrophic environments with a large mineral component and low levels of organic materials have a large population of Gram positive bacteria, such as those from the Firmicutes and Actinobacteria. Previous studies (Lauber et al. 2009) also indicate that high pH soils typically have a higher relative abundance of
members of the *Actinobacteria* and *Bacteroidetes* phyla, with a lower abundance of *Acidobacteria*, when compared with populations from more acidic soils. However, we did not find significant differences on the percentage of *Acidobacteria* among the different soils that we compared, suggesting that other factors may affect the bacterial community structure of these types of soils.

Lauber et al. (2009) examined bacterial communities in 88 soils from across North and South America using high-throughput sequencing of PCR amplified 16S rDNA genes, and their results showed that, in soils with pH values of 5-6, the dominant phyla were members of the *Acidobacteria*, *Alphaproteobacteria*, *Actinobacteria*, *Bacteroidetes*, and *Beta/Gammaproteobacteria* (Jones et al. 2009). In our samples, which had a similar pH range, we observed a lower proportion of *Acidobacteria* (5.1% vs 29.7%), and a higher proportion of *Actinobacteria* (18.5% vs 8.8%). Our data also reveal a high level (10.3%) of *Chloroflexi* phylum members, which is not commonly found in studies of deserts. However, a study of the Atacama Desert showed that the non-cyanobacteria phototrophic bacteria *Chloroflexi* was dominant in the hyper-arid core of the deserts (Lacap et al. 2011). Previous studies have shown that members of the *Chloroflexi* may play an important role as soil photoautotrophs and contribute to CO₂ uptake in the surface soil (Ley et al. 2006; Freeman et al. 2009). Members of the Family *Acetobacteraceae* were abundant in both our samples, comprising 16.8% of total sequences, on average. Members of this family have been described as nitrogen fixing bacteria able to act in plant growth promotion by a variety of mechanisms (Reis and Teixeira 2015). Koberle et al. (2011) reported a greater proportion of N-fixing bacterial groups in desert soils than farm soils, and suggest that this could be explained by the fact that plant growth promoting bacteria play an important role as a nitrogen donor in soils without compost treatment.

The abundant genera shown in Table 4 demonstrate that the dominant genus in our two samples is *Acidiphilium*, accounting for 11.3% of the bacteria in the Maine 1
sample and 13.6% for the Maine 2 sample. *Acidiphilium* is a genus in the phylum *Proteobacteria*, and many species from this genus are acidophilic bacteria isolated from acidic mineral environments (P. Harrison 1981; Wichlacz et al. 1986), which is consistent with the composition of our sample site. *Acidiphilium spp.* are also involved in the iron cycle, with the function of reducing ferric iron by oxidizing organic matter at low pH (Sanchez-Andrea et al. 2011). In mineral soils, Fe-oxidizing bacteria are well represented (Kan et al. 2011; Wu et al. 2015). Lithotrophs, such as members of *Microcoleus* in the phylum *Cyanobacteria*, are typically the dominant microorganisms in the microbial community in deserts, as well as sulfate-reducing bacteria like *Desulfovibrio*, but these two groups of bacteria do not appear to be abundant in the Desert of Maine soil (Pointing and Belnap 2012). Members of the *Alphaproteobacteria, Acidobacteria* and *Actinobacteria* are ubiquitous in mineral environments, which is consistent with the distribution of bacteria in our samples. We found that the samples contain relatively low levels of mineral weathering bacteria such as members of the *Burkholderia, Agrobacterium* and *Bacillus* genera, and only one abundant genus, *Arthrobacter*, was found to be correlated with mineral weathering (Uroz et al. 2009; Frey et al. 2010; Lepleux et al. 2012).

We also observed that 5% of the total OTUs in the dataset contained >50% of the total sequences, while approximately 80% of the total OTUs were highly diverse and contained <20% of the total sequences. The results of taxonomy assignment of the sequences showed that >30% of the sequences were not able to be classified at the genus level. This situation is frequent in studies of soil bacterial communities using high-throughput DNA sequencing techniques (Schutte et al. 2010; de Gannes et al. 2013; Hartmann et al. 2015).

In this study, we used Pyrosequencing of PCR amplified bacterial 16S rDNA genes to reveal a high degree of bacterial diversity and community structure in two soil samples from the Desert of Maine. This small sand-like environment presents unique
bacterial community patterns when compared with sand from hot deserts, and also presents differences on the abundance of predominant microorganisms within mineral soils.
Acknowledgements

We thank Ginger and Gary Currens from the Desert of Maine for their help in obtaining the sand samples, and all the members of the Laboratoire de Génomique et Biodiversité Microbienne des Biofilms (LGBMB) of the Institute for Integrative Biology of the Cell (I2BC) for interesting discussions, comments and suggestions. YW is a scholarship recipient from the China Scholarship Council (CSC). JO is a scholarship recipient from CONICYT Becas Chile. This work was supported by the Centre National de la Recherche Scientifique (CNRS), France.
References


79


82


Table 1. Average chemical and physical properties of the Desert of Maine soil samples

<table>
<thead>
<tr>
<th>Property</th>
<th>Mean Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organic carbon</td>
<td>0.43 g/kg</td>
</tr>
<tr>
<td>Total nitrogen</td>
<td>&lt;0.02 g/kg</td>
</tr>
<tr>
<td>Organic material</td>
<td>0.75 g/kg</td>
</tr>
<tr>
<td>Aluminium (Al)</td>
<td>4.86 g/100g</td>
</tr>
<tr>
<td>Calcium (Ca)</td>
<td>0.93 g/100g</td>
</tr>
<tr>
<td>Iron (Fe)</td>
<td>1.41 g/100g</td>
</tr>
<tr>
<td>Magnesium (Mg)</td>
<td>0.39 g/100g</td>
</tr>
<tr>
<td>Phosphorous (P₂O₅)</td>
<td>0.08 g/100g</td>
</tr>
<tr>
<td>Potassium (K)</td>
<td>1.97 g/100g</td>
</tr>
<tr>
<td>Sodium (Na)</td>
<td>2.15 g/100g</td>
</tr>
<tr>
<td>Manganese (Mn)</td>
<td>420 mg/kg</td>
</tr>
<tr>
<td>pH</td>
<td>5.09</td>
</tr>
</tbody>
</table>
Table 2. Summary of the number of sequences and diversity indices for each sample

<table>
<thead>
<tr>
<th></th>
<th>Sequence numbers</th>
<th>Diversity Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
<td>After cleaning</td>
</tr>
<tr>
<td><strong>Maine1</strong></td>
<td>23405</td>
<td>14776</td>
</tr>
<tr>
<td><strong>Maine2</strong></td>
<td>28983</td>
<td>19085</td>
</tr>
</tbody>
</table>

*OTUs are clustered at 97% sequence identity.
Table 3. Relative sequence proportions belonging to rare phyla in the Desert of Maine samples

<table>
<thead>
<tr>
<th>Rare phyla</th>
<th>Maine 1</th>
<th>Maine 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP11</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>SHA-109</td>
<td>0.3%</td>
<td>0.5%</td>
</tr>
<tr>
<td>WCHB1-60</td>
<td>0.3%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>0.2%</td>
<td>0.3%</td>
</tr>
<tr>
<td>TM6</td>
<td>&lt;0.1%</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>&lt;0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Elusimicrobia</td>
<td>&lt;0.1%</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>SM2F11</td>
<td>&lt;0.1%</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>Chlorobi</td>
<td>&lt;0.1%</td>
<td>&lt;0.1%</td>
</tr>
<tr>
<td>Nitrospirae</td>
<td>&lt;0.01%</td>
<td>&lt;0.1%</td>
</tr>
</tbody>
</table>
Table 4. Relative abundance of the 20 most abundant bacterial genera in the Desert of Maine samples.

<table>
<thead>
<tr>
<th>Genus</th>
<th>Maine1</th>
<th>Maine2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acidiphilium</em></td>
<td>11.3%</td>
<td>13.6%</td>
</tr>
<tr>
<td><em>Flavisolibacter</em></td>
<td>2.4%</td>
<td>2.7%</td>
</tr>
<tr>
<td><em>Crinalium</em></td>
<td>2.4%</td>
<td>1.3%</td>
</tr>
<tr>
<td><em>Methylobacterium</em></td>
<td>2.2%</td>
<td>1.3%</td>
</tr>
<tr>
<td><em>Noviherbaspirillum</em></td>
<td>0.9%</td>
<td>2.3%</td>
</tr>
<tr>
<td><em>Chthonomonas</em></td>
<td>1.1%</td>
<td>1.6%</td>
</tr>
<tr>
<td><em>Amnibacterium</em></td>
<td>1.6%</td>
<td>1.2%</td>
</tr>
<tr>
<td><em>Arthrobacter</em></td>
<td>1.5%</td>
<td>1.3%</td>
</tr>
<tr>
<td><em>Actinomycetospora</em></td>
<td>1.3%</td>
<td>1.3%</td>
</tr>
<tr>
<td><em>Blastocatella</em></td>
<td>1.1%</td>
<td>1.1%</td>
</tr>
<tr>
<td><em>Anaeromyxobacter</em></td>
<td>0.9%</td>
<td>1.3%</td>
</tr>
<tr>
<td><em>Gemmatimonas</em></td>
<td>1.0%</td>
<td>1.1%</td>
</tr>
<tr>
<td><em>Segetibacter</em></td>
<td>0.7%</td>
<td>1.4%</td>
</tr>
<tr>
<td><em>Acidobacterium</em></td>
<td>1.0%</td>
<td>1.0%</td>
</tr>
<tr>
<td><em>Singulisphaera</em></td>
<td>0.9%</td>
<td>0.7%</td>
</tr>
<tr>
<td><em>Lapillicoccus</em></td>
<td>1.0%</td>
<td>0.6%</td>
</tr>
<tr>
<td><em>Candidatus_Solibacter</em></td>
<td>0.6%</td>
<td>0.9%</td>
</tr>
<tr>
<td><em>Chamaesiphon</em></td>
<td>1.9%</td>
<td>0.5%</td>
</tr>
<tr>
<td><em>Spirosoma</em></td>
<td>0.8%</td>
<td>0.7%</td>
</tr>
<tr>
<td><em>Hymenobacter</em></td>
<td>0.6%</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

*represent genera with > 2 fold differences between the two samples
Figure Legends

Fig. 1. Location of the Desert of Maine in the northeastern USA.

Fig. 2. Relative abundance of bacterial 16S rDNA gene sequences from the two samples at the Phylum (A) and Class (B) levels. See Materials and Methods for details.

Fig. 3. Comparison of the bacterial communities with those from selected other soil samples at the Phylum level.
Relative abundance of bacterial phyla in soil samples from: Two samples of the Desert of Maine; An oak forest in France (Uroz et al. 2010); A pine forest in the USA (soil pH 4.9, total organic carbon 20.9 g/kg, Al and Fe concentration of approximately 0.1 g/100g, Shah et al. 2011); A Gobi desert sand sample in Mongolia (pH 9.8, organic carbon 0.52 g/kg, organic material 0.9 g/kg, An et al. 2013); A sample from the Atacama Desert in Chile (pH 8.9, Ca 0.96 g/100g, Na 0.34 g/100g, Chris-Christoph et al. 2013); A sample from the McMurdo Dry Valley in the Antarctic (pH 7.0, total carbon 0.11 g/kg, Fe 1.39 g/100g, Na 0.35 g/100g, Lee et al. 2012)
Fig 2.
Fig 3.
CHAPTER 3: Bacterial communities of the desert in San Rafael Swell (USA)
Bacterial community of surface soils from the San Rafael Swell

(Utah, USA)

Yang Wang¹, Chloé Jaubert², Jorge R. Osman¹, Tuyet-Nga Nguyen¹, Gustavo R. Fernandes¹, Michael S. DuBow¹*

¹Laboratoire de Génomique et Biodiversité Microbienne des Biofilms, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Univ. Paris-Sud, Bâtiment 409, 91405 Orsay, France
²Laboratoire de Microbiologie Fondamentale et Pathogénicité, UMR-CNRS 5234, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux, France
*Address for correspondence: Laboratoire de Génomique et Biodiversité Microbienne des Biofilms, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, Univ. Paris-Sud, Bâtiment 409, 91405 Orsay, France ; Tel : +33(0)169154612 ; Fax : +33(0)169157808 ; E-mail : michael.dubow@u-psud.fr

Key Words: Deserts, Bacteria, Biodiversity, 16s rDNA, Pyrosequencing

Running Head: Bacterial communities of the deserts in Utah
Abstract

Deserts in Utah has geographic features that resemble the planet Mars, characterized by red-colored hills, soils and sandstones. In this study, we examined the bacterial diversity of surface soil samples from deserts in Utah using pyrosequencing of PCR amplified bacterial 16S rDNA genes. The sample sites cover the Goblin Valley State Park and nearby regions on the Colorado Plateau. We also examined physicochemical parameters of the soil samples to investigate any possible correlations between bacterial community structure and environmental drivers. The predominant phyla of the samples belong to members of the *Proteobacteria, Actinobacteria, Bacteroidetes,* and *Gemmatimonadetes*. The most abundant genera in our samples are *Cesiribacter, Lysobacter, Adhaeribacter, Microvirga* and *Pontibacter*. We found that the relative abundance of Alphaproteobacteria and Gemmatimonadetes are significantly correlated to the environmental factors, such as pH and concentration of organic matters.
Introduction

Deserts are regions of land that have less than 25 centimeters rainfall, and where there is reduced diversity of plants and animals compared to that formed in more temperate zones. Aridity has been a major feature of the Earth’s surface, and arid lands count for the largest terrestrial biome (Pointing and Belnap 2012b). In these environments, stresses such as drought, temperature and radiation are believed to limit the scale of biodiversity (Ward 2009). Research concerning microbial colonization and dispersion in deserts has been performed to estimate the function of microbial communities from desert sand, which may play an important role in soil stability, nutrient cycles and environmental health. Microbial community composition and function in desert biomes have been shown to be remarkably different from other biomes (Fierer et al. 2012). However, the diversity of microbial communities and their functional structure in deserts are still not well studied, and there are many unanswered questions regarding their biology, physiology and ecology (Pointing and Belnap 2012a). With the development of high-throughput sequence and analysis technologies, more work is being done to fully understanding the microorganisms their residing in deserts (van Belkum et al. 2001).

Amplification and analysis of 16S rRNA genes have been broadly used as a culture-independent method for documentation of the evolutionary history and taxonomic assignment of individual organisms, as well as in characterization of microbial communities (Head et al. 1998). More recently, meta-barcode methods are broadly used, which is an amplicon-based approach, based on PCR-targeted sequencing of selected hypervariable regions of the 16S rRNA gene (Bruno et al. 2015).

There are a diverse number of relatively small deserts stretching from southeastern California to western Texas, and from Nevada and Utah to the Mexican states of
Sonora, Chihuahua, and Coahuila and much of the peninsula of Baja California (Wickens 1998). The boundary between cold and warm deserts lies across southern Nevada and Utah. Deserts in Utah have geographic features that resemble Mars, relevant for geology and astrobiology studies (Chan et al. 1998). These features include extensive wind erosion, moisture deficits, absence of vegetation, and high UV radiation among others (Makhalanyane et al. 2015). Also, the iron oxide prevalent on the surface of the Mars is similar to the deserts in Utah, characterized by red-colored hills, soils and sandstones. The Mars Society decided to set up the Mars Desert Research Station (MDRS) in the area as a Mars analog for such reasons (http://mdrs.marssociety.org/).

Bacteria constitute the largest portion of the biodiversity in soils, and play an important role in maintaining soil processes, which affect the functioning of terrestrial ecosystems (Epp et al. 2012). In this study, we examined the bacterial diversity research of surface soil samples from deserts on the Colorado Plateau (Fig. 1). The samples were collected in Goblin Valley (GV), the Little Wild Horse Canyon (LWH) located at south of Goblin Valley, south of the Utah State Route 24 (SR) and Temple Junction road (TJR) near the region of the Goblin Valley State Park, and 200 km southwest of the park near the Muddy Creek road (MCR).

Goblin Valley is a State Park in the USA, and lies on the southeast of the east edge of the San Rafael Swell, which is part of the Colorado Plateau physiographic region. The park is characterized by the presence of hoodoo rocks, referred to as "goblins". Those rock sculptures result from the weathering of Entrada Sandstone. The Entrada Sandstone was deposited in the Jurassic Period around 170 million years ago, and consists of debris eroded from former highlands and redeposited as alternating layers of sandstone, siltstone and shale (Milligan 1999). The distinct shape of these rocks comes from an erosion-resistant layer of hard rock atop softer sandstone which has eroded more quickly. The average daytime temperature in summer is between 32 °C
and 41 °C, and in winter daytime temperatures are above freezing most days but can drop to as low as −12 °C at night with occasional snow. The average annual precipitation is less than 20 cm (http://www.usgs.gov/, Precipitation of the Individual States and of the Conterminous States). During the summer, monsoons can arrive and increase precipitation in some regions between July and mid-September, and consist of at least 50% of the annual precipitation during this period.

The richness and patterns of microbial diversity in soils can be effected by many different environmental factors. Studies of microbial biogeography using metagenomics can often provide key insights into the physiologies, environmental tolerances, and ecological strategies of microbial taxa, particularly those difficult-to-culture taxa that often dominate in natural environments (Lauber et al. 2009). The bacterial community in desert environments can be regarded as a target for fundamental research on ecological and evolutionary processes, as the bacterial diversity in deserts have been found to be more rich than earlier expected (Cary et al. 2010). In this study, we demonstrated the bacterial diversity and community structures of surface soil in the Corolado Plateau in the Utah State using pyrosequencing of 16S rRNA amplicons. We built our pipeline for the analysis of 16S rRNA pyrosequencing data by combining several existing tools of metagenomics. We also examined correlations between certain environmental factors and bacterial diversity in the two deserts.
Materials and Methods

Sampling. To perform the bacterial diversity analysis of surface soils in the San Rafael Swell in Utah (USA), we recovered 18 samples of the soil (sand and soft rock particles) in 5 different regions: Dry river bed in Little Wild Horse Canyon, Goblin Valley State Park, near State Road 23 / I 70 Road, near Muddy Crack Road. All samples were taken at least 91 meters (100 yards) apart. The sampling GPS sites and description are shown in Table 1. The samples were collected by scooping surface sand into 50 ml sterile polyethylene conical centrifuge tubes, in different sites during September 2011. After collection the samples were treated as previously described (An et al. 2013). To perform the analyses of selected physicochemical parameters of the soil samples, the soil from each sites was pooled and sent for analyses, using standard methods, by the Laboratoire d’Analyses de Sols (INRA-ARAAS, France).

Sample Preparation. The total DNA was extracted from each sample using the protocol of An et al. (2013). An aliquot of extracted DNA was adjusted to a final DNA concentration of 15 ng/μl in 1/10 TE buffer (1 mM Tris pH 8; 0.1 mM EDTA) using a NanoVue spectrophotometer (GE Healthcare, Little Chalfont, Buckinghamshire, UK), and the concentration verified by ethidium bromide fluorescence after electrophoresis through a 1% agarose gel in TAE buffer (2 mM Tris-acetate pH 8; 5 mM Na-EDTA). PCR reactions were performed in 25 μl reaction volumes. Each reaction contained one of two different thermostable DNA polymerases and their corresponding reaction buffers, 200 μM of each dNTP, 0.5 μM of each primer and 1 to 10 ng of extracted DNA. The 16S rRNA genes were amplified using the universal bacterial primers for pyrosequencing and covering hypervariable regions V1-V3: primer 27F (A adaptor + GAGTTTGATCMTGGCTCAG) and primer 518R (B adaptor + Mid + WTTACCGCGGCTGCTGG), where A and B represent the adaptors using the 454 Roche FLX Titanium pyrosequencing reaction platform. The Mid sequences are eight nucleotide tags designed for sample identification barcoding according to the 454
protocol. PCR amplification conditions were adapted for the use of two different thermostable DNA polymerases: A) Phusion High-Fidelity DNA Polymerase (Finnzymes, Finland): 98°C for 2 min, followed by 28 cycles of 98°C for 30 secs, 54°C for 20 secs and 72°C for 15 secs, and a final elongation step at 72°C for 5 min; B) Pfu DNA Polymerase (Fermentas, Canada): 95°C for 2 min, followed by 30 cycles of 95°C for 30 secs, 48°C for 30 secs and 72°C for 1 min, and a final elongation step at 72°C for 5 min. Each DNA sample was subjected to 3-5 different PCR reactions per DNA polymerase to minimize PCR bias. The PCR products were pooled and subjected to electrophoresis through a 1% agarose gel in TAE buffer. After electrophoresis and visualization of the PCR products by ethidium bromide staining and long wave UV light illumination, NucleoSpin Extract II kits (Macherey-Nagel, Germany) were used to purify the 16S rDNA PCR products. Then, 40 ng of PCR products from each sample were mixed for pyrosequencing, performed using a 454 Roche FLX Titanium Pyrosequencer (Microsynth AG, Switzerland).

**DNA Sequence Data Processing.** The raw DNA sequences were first assigned to each sample via their Mid tag using MOTHUR version 1.33 (Schloss et al. 2009), and reads were removed if at least one of the following criteria was met: (i) length less than 200 nt or longer than 600 nt, (ii) mismatch to the barcode sequences or more than one mismatch to the primer, and (iii) the presence of homopolymers of > 8 bp in length. Adaptor sequences were removed from the sequences using the “Cutadapt” tool (Martín 2011) implemented in the Galaxy server of the Institut de Génétique et Microbiologie (IGM) of the Université Paris-Sud (http://galaxy.igmors.u-psud.fr). Then, the sequences were checked for quality scores by ConDeTri version 2.2 (Smeds and Kunstner 2011), using the criteria that 80% of the nucleotides in a sequence have quality scores > 25. We used UCHIME (Edgar et al. 2011), with reference database Greengenes version 2013_May, and Decipher (through web tools available at http://decipher.cee.wisc.edu/) to detect chimera sequences (Wright et al. 2012). Sequences detected as chimeras by both programs were removed from the data sets.
The raw sequences have been deposited in the GenBank short-read archive (SRA), under the accession number SRP063276.

Taxonomy assignments of the remaining 16S rDNA reads were conducted using SILVAngs classifier (online server) with a similarity threshold of 90, using the Silva database release 123 (Cole et al. 2014; Yilmaz et al. 2014). Sequences classified as Chloroplast, or that could not be classified as belonging to the Bacteria Kingdom, were removed. Diversity analyses were performed using the software package MOTHUR. Normalization of 16S rDNA sequences based on the copy number of different species were performed using Tax4Fun package in R (Aßhauer et al. 2015).

**Statistical Analysis.** The clean reads were clustered into operational taxonomic units (OTUs) using UPARSE at a cutoff value of 97% sequence identity (Edgar 2010; Edgar 2013). The Chao1 and Shannon indices were calculated to estimate taxon richness and diversity (Schloss and Handelsman 2008). Bray-Curtis distance was used for calculating the distances of multiple samples. β-diversity of samples Principal coordinate analysis (PCoA) was conducted by R to group the microbial communities of different samples. The beta_diversity.py command was used to estimate the beta diversity of bacterial community using Qiime version 1.80 (Caporaso et al. 2010). Pearson correlations between relative abundances of sequences classified on specific phyla of different sample sites and environmental parameters were performed using SPSS version 22.
Results

Chemical properties of the sand samples

The chemical properties of the sand samples are shown in Table 2. For all the sample sites, the pH values of the sand are above 8, indicating an alkaline soil environment, which is normal for hot desert sand samples (Ward 2009). The pH values of the soil at the site of MCR and SRA are 10.1, higher than the other sites in the sample region (pH = 8.2 to 8.8). The chemical component results shows that the concentration of sodium (Na), potassium (K), iron (Fe) and aluminum (Al) were relatively higher in the two sites (MCR and SRA), which present a similar tendency for the pH value in the different sites. These results may indicate that the high concentrations of these metal ions, which come from decomposition of alkaline salt in the soil, contribute or lead to a high alkaline soil environment. On contrast, the concentration of organic materials and organic carbon were relatively lower in the two sites (MCR and SRA) and higher in TJR, which seems to inversely correspond to the concentration of mineral elements (Na, K, Fe, Al).

Sequence cleaning

We obtained 483364 sequences of raw reads in total from different groups of samples by 454 Roche Pyrosequencing (Microsynth AG, Switzerland), with an average length of 480 nt. After bioinformatic cleaning with Mothur, approximately 95% of the sequences remained (460475 reads). The sequences were further filtered for quality and examined for chimeric sequences, leaving 65% of the total reads (312358 reads). The average length of the sequences after processing were approximately 400 nt. The number of reads for each sample are presented in Table 3.

Classification of sequences

In total, 305,167 sequences from the 18 samples are classified into the bacteria domain (sequences classified into Chloroplast are removed from our data). The
sequences of were classified into 6 taxonomic levels by SILVAngs, including 37 phyla, 136 classes, 136 orders, 254 families and 565 genera. Sequences (97%) are classified into 37 phyla, with 13 common phyla for all the samples (Fig. 1a). The predominant phyla of the samples represent members of the Proteobacteria, Actinobacteria, Bacteroidetes, and Gemmatimonadetes, at 28.3%, 25.7%, 19.7%, and 6.5%, as the respective averages for the 18 samples, followed by members of the Acidobacteria (3.9%), Chloroflexi (3.8%), Firmicutes (2.6%), Armatimonadetes (2.3%), Denococcus-Thermus (1.9%), TM7 (1.3%), Saccharibacteria (1.3%), Planctomycetes (1.2%), Cyanobacteria (0.9%), and Verrucomicrobia (0.5%). Other phyla with less than 0.5% of bacterial sequences in total, are presented together as less common phyla.

The sequences assigned to the Proteobacteria phylum, representing 28.3% of the total, fell into the Alpha-, Beta-, and Gamma- three major class. We found that distributions of sequences belonging to the group of Alphaproteobacteria are significant different across sites (P = 0.046), with highest percentage in TJR (23.8% on average) and the lowest percentage in the region SR (5.9% on average). Proteobacteria are the most abundant in TJR_3 and TJR_4 (56.5% and 61.2% respectively), but the primary family in the two sites are different, with Methylobacteriaceae from Alphaproteobacteria dominant in TJR_3 (27.8%) and Xanthomonadaceae from Gammaproteobacteria dominant in TJR_4 (30.4%). Within the Betaproteobacteria class, members of the Burkholderiales family are the most abundant in all our samples.

The Actinobacteria was the second most abundant phylum found in our data, accounting for 25.7% of the total sequences. Three dominant classes within this phylum in our samples belong to the Frankiales, Micrococcales and Acidimicrobiales class. It has been reported that sequences assigned to Actinobacteria are the most abundant (>70%) in the soil sample of the Atacama Desert (Drees et al. 2006). Members of the
Actinobacteria are often found to prevail in desert soil (de la Torre et al. 2003; Neilson et al. 2012), such as *Rubrobacter*, *Arthrobacter*, and *Streptomyces*. However, many taxa belonging to *Actinobacteria* isolated from deserts soils appear to be new species (Mayilraj et al. 2006; Yung et al. 2007; Qin et al. 2009; Makhalanyane et al. 2015). In our study, we found a large proportion (> 50%) of sequences assigned to this phylum cannot be classified to genus level within the order *Frankiales* and *Acidimicrobiales*.

The *Bacteroidetes* was the third most abundant phylum with 19.7% of total sequences, which are commonly present in desert soils. Members of this phylum show optimum growth at high pH values, which is consistent with the alkaline pH seen in most deserts (Lauber et al. 2009). Members from four family were predominant groups in our samples: *Cytophagaceae*, *Flammeovirgaceae*, *Chitinophagacaea* and *Flavobacteriaceae*. Bacteroidetes are most abundant in MCR_3, presenting 64.3% of total sequences. *Firmicutes* are often observed to be more abundant in extreme environments, especially in hot deserts (Andrew et al. 2012; Marasco et al. 2012; Crits-Christoph et al. 2013). An et al. (2013) reported that sequences classified to the phylum Firmicutes account for > 60% of total sequences in a study of surface soils in the Gobi and Takalamaken deserts. We summarized the most abundant sequences classified at class level in Fig. 2.

**Genus**

Within the 565 genera identified among the sequences, 242 genera belong to the phylum *Proteobacteria*, 125 genera belong to the phylum *Actinobactia*, 24 genera belong to the phylum *Bacteroidetes*, and 9 genera belong to the phylum *Acidobacteria*. The most abundant genera in sample sites are shown in Table 4. We found sequences classified to some genera can be predominant and counting for as much as more than 10% of total sequences in some samples. *Microvirga* are most abundant in TJR_3, representing 32.2% of the total sequences. Members of this genus have been reported to be thermophile and members have been isolated from a semi-arid site in Brazil.
(Radl et al. 2014) and thermal aquifer (Radl et al. 2014). Sequences belonging to the \textit{Cesiribacter} genus were predominant in MCR\_3 and SRB\_2, accounting for 30.8\% and 19.9\% of total sequences, respectively, in the two samples. Strains of this genus have been isolated from desert sand in China (Liu et al. 2012). Other predominant genera in a sample with relative abundance > 10\% are \textit{Lysobacter} (25.5\% in TJR\_4), \textit{Salinimicrobium} (12.8\% in TJR\_2) and \textit{Achromobacter} (11.7\% in TJR\_4).

**Normalization of sequences**

To study the real population of bacteria, sequences were normalized base on 16S rRNA gene copy number of different species by Tax4fun. We found that the structure of bacterial community shifts at all the taxonomic level. At the phylum level (Fig. 1b), the relative abundance of each phylum generally reduced since sequences that could not be aligned to the reference database with 16S rRNA copy number will be assigned to unclassified sequences (15.5\% of total sequences at phylum level). However, relative abundance of the Actinobacteria increased from 25.7\% before normalization to 35.3\% after, which indicate the population of the Actinobacteria was under estimated by the method of simply summary of taxonomic classification based on number of sequences.

**OTU-based analysis**

The clean sequences were clustered into OTUs at 97\% similarity, excluding the unclassified sequences at the phylum level and sequences classified as Chloroplast, Eukaryote or Mitochondria. In total, 9908 OTUs were observed in all the samples, ranging from 778 in the Drb\_1 sample to 3595 in the Tjr\_1 sample. The OTU numbers observed in each sample were shown in Table 3. Eleven core taxons (most abundant OTUs with more than 1\% of total sequences) were found, comprising 37\% of the population of all the samples. The number of singleton (OTUs with only one sequence) is 3926, comprising 37\% of the total OTU numbers. Only 49 OTUs were observed in all the 18 samples. For different sample regions, the numbers of common OTUs within a
given sample site were: 599 for DRB (Drb_1 and Drb_2), 323 for GV (GV_1 to GV_5),
170 for MCR (Mcr_1 to Mcr_4), 323 for SR (SR_1 to SR_3), and 238 for TJR (Tjr_1 to
Tjr_4).

The distance of the microbial communities between samples were calculated based on
OTU abundance using a Bray-Curtis measure, and Principal Component Analysis
(PCoA) was used to perceive the distances of the 18 samples (Fig. 3). The distance
between different sites based on the OTU relative abundance are illustrated by a
UPGMA tree (Fig. 4). The tree was generated by merging sequences in each site and
normalized to the same number of reads (12501) in Qiime. The sample sites of GV
and LWH are more close in distance compared with the other sites, and they are also
close in geographic distance. The microbial population of the TJR site has the largest
distance from the other sites, and these may correspond to the difference in
environment parameters, as the concentration of organic carbon and mineral salts are
significantly different at the TJR site (P < 0.05).
Discussion

The deserts in Utah present a range of arid conditions, with high salinity and low concentration of carbon, providing unique environment for microbial communities, and a combination of geochemistry and biodiversity data can be applied to study the environmental factors that may shape the community structure of soils (Crits-Christoph et al. 2013). In this study, we used pyrosequencing amplicons of 16S rRNA genes to analyze the bacterial diversity and communities in the deserts in the Utah State in the USA. We also examined some physicochemical parameters of sample sites to investigate the correlations between bacterial community structure and environmental drivers.

Fierer et al. (2012) reported that soils close to neutral had the highest bacterial diversity levels, compared with very basic (deserts) or acidic soils (rainforests and Arctic tundra) using metagenomic sequencing of total DNA extracted from soils from a wide range of ecosystems. The alkaline conditions may select for taxa most adapted to alkaline growth conditions, and the contribution of pH to bacterial diversity may thus be limited (Finkel et al. 2012). In this study, we used the Pearson’s correlation analysis to illustrate the correlation between bacterial diversity and some environmental factors that may affect the bacterial community in soils (Table 5). We found that the most significant correlation with bacterial richness (Chao 1) is the concentration of organic carbon (P < 0.01), and the correlation between richness and the concentration of mineral elements (Na, K, Al, Fe) are also significant (P < 0.05). The two significant correlation was with the Shannon diversity index and the pH value of soils and the concentration of sodium (P < 0.05). The concentration of Ca, P, Mg and Mn are not found to be significantly correlated with bacterial diversity. However, concerning the limited number of sites in our study (n=5), the correlation between environmental factors and bacterial diversity should be comprehended cautiously.
Some bacterial community studies of surface soils in deserts found that geographical location plays an important role in the microbial communities (Finkel et al. 2011; Finkel et al. 2012; Nemergut et al. 2013). Beta-diversity of bacterial community structure among different samples were performed to test the hypothesis that variability of microbial communities found in desert soils in Utah across different sites is more than the variability within sites. We found the bacterial community in our samples shifts with different sites (Fig.1). However, in the SR, LWH and GV sites, the distribution of bacterial communities are more closely related, and different from MCR and TJR. Although, the MCR site is the most distant geographically from the other sample sites, we did not find a significant difference in the bacterial community structure compared with samples from other sites. The assessment of environmental factors in our study suggests that the concentration of mineral components appears to present a strong correlation with the bacterial community structure (Table 6). In the TJR site, the concentration of organic carbon and mineral salts are significantly different from the other sites (P < 0.05), and displays the largest dissimilarity on the OTUs distribution and relative abundance of some taxonomic groups (Alphaproteobacteria, Acidimicrobiia, Gemmatimonadetes and Deinococci, with P value < 0.05).

Desert environments is a selective environment that has been reported to present bacterial specificity even at the higher taxonomic levels (phylum to family) (Tamames et al. 2010). Salinity is a very important factor in shaping prokaryotic diversity (Li et al. 2013; Canfora et al. 2014; Geyer et al. 2014). In this study, we found that the relative abundance of sequences assigned to the phyla Proteobacteria and Gemmatimonadetes can be significantly related to certain environmental parameters. The relations between environmental factors and Proteobacteria relative abundances were performed using a Pearsons correlation analysis (Table 6). We found that the relative abundance of Proteobacteria was inversely correlated (P < 0.05) with soil pH value (r²= -0.523, with pH ranging from 8.2 to 10.1), as well as the
concentrations of Na, K, Fe, and Al. We also found that the relative abundance of sequences of this phylum was positively correlated with concentration of organic carbon and organic materials (P < 0.05). *Proteobacteria* have been previously reported to be abundant in desert soils (Rasuk et al. 2014; Li et al. 2015).

Sequences related to *Gemmatimonadetes* are commonly found in soils of various environments, including hot desert (Azua-Bustos et al. 2012), pasture (Chim Chan et al. 2008), crop agriculture (Montecchia et al. 2015), forests (Li et al. 2014) and freshwater sediments (Zhang et al. 2015), although only six cultured isolates have been reported (Fawaz 2013). DeBruyn et al. (2011) reported that *Gemmatimonadetes* relative abundances was inversely correlated to moisture in soils, and many *Gemmatimonadetes* phylotypes have higher relative abundances in semiarid and arid soils and deserts (DeBruyn et al. 2011). This also has been reinforced by a study on the distribution of the *Gemmatimonadetes* in agricultural soils, in that members of this phylum prefer dryer soils and tend to be more dependent on moisture availability. In our study, we found *Gemmatimonadetes* relative abundance was significantly different among sample sites (P=0.017), with the highest concentration in SR and MCR and lowest in TJR. Based on the analysis of environmental factors, SR and MCR sites present to be most “extreme” microenvironment compared with other sites, with the lowest concentration of organic materials and highest concentration of mineral elements, as well as the highest pH value (10.1). The relations between environmental factors and *Gemmatimonadetes* relative abundances were performed using the Pearsons correlation analysis (*Table 6*). Earlier studies have showed that relative abundances of *Gemmatimonadetes* are higher in soils with neutral pH versus these in acidic soils (28, 32, 48). We found that *Gemmatimonadetes* was significantly correlated (P < 0.001) with soil pH ($r^2$=0.854), with pH ranges from 8.2 to 10.1. We also found that the relative abundance of sequences of this phylum was significant correlated with the concentration of Na, Fe, K and Al, as well as the concentration of organic carbon and organic materials (P < 0.05). These results indicate that members
of the *Gemmatimonadetes* may present as potential members of biomarkers in soil communities of arid environment characterized by relative higher salinity and lower organic matter.

The percentage of genus level classifiable sequences varies among the phyla we found. For example, 91.4% sequences in the *Bacteroidetes* phylum could be classified to the genus level, while only 37.0% of the *Actinobacteria* phylum are able to be classified to this level. Also, sequences related to the *Acidobacteria* phylum can only rarely be classified to the genus level in our data. These indicated that sequences related to the phyla *Actinobacteria* and *Acidobacteria* in desert soil may present more as unknown species. Direito. *et al.* (2011) studied the microbial diversity near the MDRS in the Utah based on the DGGE profiles of ribosomal RNA genes, and found 239 clones of bacteria (Direito et al. 2011). All the phyla observed in their study have also been detected in our research and we recovered much more different bacterial taxa by using the pyrosequencing method. Sequences assigned to genera that are related to several types of extremophiles or isolated from desert-like environments were observed in all our samples, including radio-resistant, halophilic, thermophilic and endolithic bacteria. *Pontibacter* of the *Bacteroidetes* phylum are found to be abundant in deserts (Zhou et al. 2007; Subhash et al. 2014), and may possess unique abilities to adapt to desert environments (Makhalanyane et al. 2015). In the samples from the desert in Utah we observed that sequences belonging to this genus are abundant (2.2% of total sequences), and most abundant in the sample SRB_2 (7.7% of total sequences). Sequences classifiable to the genera *Rubrobacter, Truepera, Deinococcus* were reported to be radiotolerant (Cox and Battista 2005), and we found sequences classified to these genera to be abundant in the samples in deserts in Utah. Other genera that have been reported to be extremophiles were also observed, including: halophile bacteria such as *Salegentibacter, Halomonas* and *Salinimicrobium, Actinobacteria* (Rossello-Mora et al. 2003; Oren 2015); thermophilic bacteria such as *Thermoleophilum* and *Thermosporothrix* (Andrade et al. 1999); and *Rhodobacter*
known as endolithic bacteria (Stivaletta et al. 2010). Another group of bacteria that are found in these desert environments is *Pseudomonas*, and the members of the Pseudomonas family are believed to play a protective role for bacterial communities in many extreme environments because their ability for biofilm formation (Drenkard and Ausubel 2002; Selenska-Pobell et al. 2002).

In this study, we used Pyrosequencing of PCR amplified bacterial 16S rDNA genes to reveal the bacterial diversity and community structure in surface soil samples from the deserts in Utah, as well as their possible correlations with selected environmental factors. The research of bacterial diversity of soils in the desert in Utah presents a unique opportunity to understand bacterial communities in arid soils, characterized by high salinity and a low concentration of organic matter.
Acknowledgements

We thank all the members of the Laboratoire de Génomique et Biodiversité Microbienne des Biofilms (LGBMB) of the Institute for Integrative Biology of the Cell (I2BC) for interesting discussions, comments and suggestions. This work was supported by the Centre National de la Recherche Scientifique (CNRS), France.
References


community sequencing data. Nat Methods 7:335-336. doi:10.1038/nmeth.f.303


of the Atacama Desert, Chile. Extremophiles 16:553-566. doi:10.1007/s00792-012-0454-z


### Tables

**Table 1.** Information of location of Sample sites.

<table>
<thead>
<tr>
<th>Region</th>
<th>Location Description</th>
<th>Lat/Lon</th>
<th>Sample Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWH</td>
<td>Dry river bed in Little Wild Horse Canyon</td>
<td>38°34'32.18&quot; N 110°46'55.06&quot; W</td>
<td>LHW_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LHW_2</td>
</tr>
<tr>
<td>GVA</td>
<td>Goblin Valley State Park</td>
<td>38°34'30.73&quot; N 110°43'47.24&quot; W</td>
<td>GVA_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GVA_2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GVA_3</td>
</tr>
<tr>
<td>GVB</td>
<td></td>
<td>38°34'12.96&quot; N 110°44'31.89&quot; W</td>
<td>GVB_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GVB_2</td>
</tr>
<tr>
<td>MCR</td>
<td>Near Muddy Creeck Road</td>
<td>37°19'57.49&quot; N 112°41'42.87&quot; W</td>
<td>MCR_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MCR_2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MCR_3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MCR_4</td>
</tr>
<tr>
<td>SRA</td>
<td>Near State Road 24 / I 70 road</td>
<td>38°49'51.46&quot; N 110°22'52.35&quot; W</td>
<td>SRA_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRB</td>
<td></td>
<td>38°52'53.72&quot; N 110°20'8.87&quot; W</td>
<td>SRB_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SRB_2</td>
</tr>
<tr>
<td>TJR</td>
<td>Near Temple Junction Road / Goblin Valley Road</td>
<td>38°39'38.18&quot; N 110°38'46.79&quot; W</td>
<td>TJR_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TJR_2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TJR_3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TJR_4</td>
</tr>
</tbody>
</table>
Table 2. The Physicochemical data of the five soil sample sites

<table>
<thead>
<tr>
<th></th>
<th>LWH</th>
<th>GVA</th>
<th>MCR_1</th>
<th>SRA</th>
<th>TJR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>8.2</td>
<td>8.84</td>
<td>10.1</td>
<td>10.1</td>
<td>8.23</td>
</tr>
<tr>
<td>Ca (g/100g)</td>
<td>11.6</td>
<td>4.14</td>
<td>4.29</td>
<td>0.59</td>
<td>0.32</td>
</tr>
<tr>
<td>Fe (g/100g)</td>
<td>1.49</td>
<td>0.85</td>
<td>1.81</td>
<td>1.85</td>
<td>0.2</td>
</tr>
<tr>
<td>Mg (g/100g)</td>
<td>2.42</td>
<td>0.91</td>
<td>1.53</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>P (g/100g)</td>
<td>0.17</td>
<td>0.08</td>
<td>0.14</td>
<td>0.1</td>
<td>0.02</td>
</tr>
<tr>
<td>Na (g/100g)</td>
<td>0.37</td>
<td>0.55</td>
<td>0.72</td>
<td>0.81</td>
<td>0.08</td>
</tr>
<tr>
<td>K (g/100g)</td>
<td>1.46</td>
<td>1.49</td>
<td>2.26</td>
<td>2.14</td>
<td>0.92</td>
</tr>
<tr>
<td>Mn (mg/kg)</td>
<td>511</td>
<td>272</td>
<td>359</td>
<td>172</td>
<td>29.3</td>
</tr>
<tr>
<td>Al (g/100g)</td>
<td>3.19</td>
<td>3.11</td>
<td>4.8</td>
<td>4.8</td>
<td>1.11</td>
</tr>
<tr>
<td>Organic materials (g/kg)</td>
<td>2.61</td>
<td>1.17</td>
<td>1.15</td>
<td>0.21</td>
<td>3.49</td>
</tr>
<tr>
<td>Organic carbon (g/kg)</td>
<td>1.51</td>
<td>0.68</td>
<td>0.67</td>
<td>0.12</td>
<td>11.7</td>
</tr>
</tbody>
</table>
Table 3. Number of Sequences and Diversity Index of the 18 soil samples

<table>
<thead>
<tr>
<th>Soil sample</th>
<th>Sequence numbers</th>
<th>Diversity indices</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw data</td>
<td>After cleaning</td>
</tr>
<tr>
<td>LWH_1</td>
<td>11632</td>
<td>7572</td>
</tr>
<tr>
<td>LWH_2</td>
<td>20221</td>
<td>13515</td>
</tr>
<tr>
<td>GVA_1</td>
<td>28382</td>
<td>18721</td>
</tr>
<tr>
<td>GVA_2</td>
<td>30479</td>
<td>19357</td>
</tr>
<tr>
<td>GVA_3</td>
<td>28744</td>
<td>18533</td>
</tr>
<tr>
<td>GVB_1</td>
<td>25342</td>
<td>17198</td>
</tr>
<tr>
<td>GVB_2</td>
<td>27282</td>
<td>17258</td>
</tr>
<tr>
<td>MCR_1</td>
<td>21955</td>
<td>14976</td>
</tr>
<tr>
<td>MCR_2</td>
<td>25839</td>
<td>15756</td>
</tr>
<tr>
<td>MCR_3</td>
<td>45693</td>
<td>31143</td>
</tr>
<tr>
<td>MCR_4</td>
<td>19433</td>
<td>13165</td>
</tr>
<tr>
<td>SRA_1</td>
<td>20331</td>
<td>14028</td>
</tr>
<tr>
<td>SRB_1</td>
<td>33407</td>
<td>22134</td>
</tr>
<tr>
<td>SRB_2</td>
<td>37157</td>
<td>25574</td>
</tr>
<tr>
<td>TJR_1</td>
<td>25380</td>
<td>16546</td>
</tr>
<tr>
<td>TJR_2</td>
<td>28836</td>
<td>15976</td>
</tr>
<tr>
<td>TJR_3</td>
<td>19900</td>
<td>12750</td>
</tr>
<tr>
<td>TJR_4</td>
<td>33351</td>
<td>16549</td>
</tr>
</tbody>
</table>
Table 4. The most abundant sequences at the genus level of soil sample sites

<table>
<thead>
<tr>
<th>Genus</th>
<th>GVA</th>
<th>GVB</th>
<th>LWH</th>
<th>MCR</th>
<th>SRA</th>
<th>SRB</th>
<th>TJR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cesiribacter</td>
<td>0.1%</td>
<td>0.6%</td>
<td>1.0%</td>
<td>6.4%</td>
<td>6.0%</td>
<td>7.4%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Lysobacter</td>
<td>0.1%</td>
<td>0.2%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>0.2%</td>
<td>0.2%</td>
<td>10.1%</td>
</tr>
<tr>
<td>Pontibacter</td>
<td>0.9%</td>
<td>1.2%</td>
<td>1.1%</td>
<td>3.0%</td>
<td>3.5%</td>
<td>5.9%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Adhaeribacter</td>
<td>2.8%</td>
<td>1.6%</td>
<td>1.9%</td>
<td>3.1%</td>
<td>1.6%</td>
<td>2.0%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Microvirga</td>
<td>0.2%</td>
<td>0.2%</td>
<td>1.0%</td>
<td>0.4%</td>
<td>0.6%</td>
<td>0.2%</td>
<td>7.5%</td>
</tr>
<tr>
<td>Sphingomonas</td>
<td>1.2%</td>
<td>1.3%</td>
<td>2.4%</td>
<td>2.0%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Segetibacter</td>
<td>4.0%</td>
<td>2.0%</td>
<td>1.2%</td>
<td>2.3%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Flavisolibacter</td>
<td>3.9%</td>
<td>1.3%</td>
<td>0.8%</td>
<td>1.8%</td>
<td>0.2%</td>
<td>0.3%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Massilia</td>
<td>0.9%</td>
<td>3.2%</td>
<td>1.9%</td>
<td>1.4%</td>
<td>5.7%</td>
<td>0.6%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Cytophaga</td>
<td>0.3%</td>
<td>0.2%</td>
<td>0.8%</td>
<td>3.1%</td>
<td>3.2%</td>
<td>2.2%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Bacillus</td>
<td>1.2%</td>
<td>0.5%</td>
<td>0.9%</td>
<td>2.4%</td>
<td>1.1%</td>
<td>0.2%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Truepera</td>
<td>0.5%</td>
<td>0.5%</td>
<td>1.9%</td>
<td>0.6%</td>
<td>4.2%</td>
<td>4.0%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Rubellimicrobium</td>
<td>0.9%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>0.9%</td>
<td>1.3%</td>
<td>1.2%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Arthrobacter</td>
<td>0.6%</td>
<td>1.5%</td>
<td>1.9%</td>
<td>0.7%</td>
<td>3.3%</td>
<td>1.3%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Hymenobacter</td>
<td>1.9%</td>
<td>1.8%</td>
<td>0.8%</td>
<td>1.2%</td>
<td>0.1%</td>
<td>1.2%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Blastocatella</td>
<td>0.8%</td>
<td>0.3%</td>
<td>0.6%</td>
<td>1.2%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Salinimicrobium</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>0.7%</td>
<td>0.0%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Stenotrophomonas</td>
<td>3.3%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>1.4%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Achromobacter</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Acinetobacter</td>
<td>3.0%</td>
<td>0.2%</td>
<td>1.0%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.9%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Rubrobacter</td>
<td>1.8%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.5%</td>
<td>0.9%</td>
<td>1.8%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Euzelbea</td>
<td>0.4%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>1.0%</td>
<td>0.2%</td>
<td>2.4%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Deinococcus</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.8%</td>
<td>1.8%</td>
<td>0.6%</td>
<td>0.8%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Flavobacterium</td>
<td>0.2%</td>
<td>4.0%</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Nocardoides</td>
<td>0.6%</td>
<td>0.6%</td>
<td>1.3%</td>
<td>0.3%</td>
<td>1.1%</td>
<td>0.3%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Streptomyces</td>
<td>0.3%</td>
<td>0.3%</td>
<td>1.1%</td>
<td>0.2%</td>
<td>6.6%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Altererythrobacter</td>
<td>0.2%</td>
<td>0.9%</td>
<td>1.5%</td>
<td>0.2%</td>
<td>2.0%</td>
<td>0.1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Pseudomonas</td>
<td>0.6%</td>
<td>1.5%</td>
<td>0.5%</td>
<td>0.1%</td>
<td>0.9%</td>
<td>0.3%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Blastococcins</td>
<td>0.3%</td>
<td>0.2%</td>
<td>1.1%</td>
<td>0.3%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>1.0%</td>
</tr>
<tr>
<td>Solirubrobacter</td>
<td>1.8%</td>
<td>0.2%</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Gillisia</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Nibribacter</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.4%</td>
<td>0.8%</td>
<td>0.5%</td>
<td>1.0%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Patulibacter</td>
<td>1.2%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Gemmatimonas</td>
<td>0.7%</td>
<td>0.6%</td>
<td>0.4%</td>
<td>0.6%</td>
<td>0.1%</td>
<td>0.2%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Noviherbaspirillum</td>
<td>0.3%</td>
<td>1.0%</td>
<td>0.9%</td>
<td>0.5%</td>
<td>0.1%</td>
<td>0.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Planomicrobium</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
Table 5. Pearson’s correlations between OTU richness and environmental factors

<table>
<thead>
<tr>
<th></th>
<th>PH</th>
<th>Organic carbon</th>
<th>Organic materials</th>
<th>Na</th>
<th>Ca</th>
<th>Fe</th>
<th>Al</th>
<th>P</th>
<th>K</th>
<th>Mg</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTUs</td>
<td>$R^2$</td>
<td>-.781</td>
<td>.912*</td>
<td>.956**</td>
<td>-.966**</td>
<td>-.028</td>
<td>-.852*</td>
<td>-.930*</td>
<td>-.529</td>
<td>-.875*</td>
<td>-.366</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>.060</td>
<td>.016</td>
<td>.005</td>
<td>.004</td>
<td>.482</td>
<td>.033</td>
<td>.011</td>
<td>.180</td>
<td>.026</td>
<td>.272</td>
</tr>
<tr>
<td>Chao1</td>
<td>$R^2$</td>
<td>-.657</td>
<td>.972**</td>
<td>.815*</td>
<td>-.895*</td>
<td>-.356</td>
<td>-.938**</td>
<td>-.938**</td>
<td>-.794</td>
<td>-.865*</td>
<td>-.667</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>.114</td>
<td>.003</td>
<td>.047</td>
<td>.020</td>
<td>.278</td>
<td>.009</td>
<td>.009</td>
<td>.054</td>
<td>.029</td>
<td>.110</td>
</tr>
<tr>
<td>Shannon</td>
<td>$R^2$</td>
<td>-.843*</td>
<td>.547</td>
<td>.879*</td>
<td>-.837*</td>
<td>.449</td>
<td>-.648</td>
<td>-.755</td>
<td>-.102</td>
<td>-.757</td>
<td>.087</td>
</tr>
<tr>
<td></td>
<td>$P$</td>
<td>.036</td>
<td>.170</td>
<td>.025</td>
<td>.039</td>
<td>.224</td>
<td>.119</td>
<td>.070</td>
<td>.435</td>
<td>.069</td>
<td>.445</td>
</tr>
</tbody>
</table>

* with $P$ value < 0.05
** with $P$ value < 0.01
Table 6. Pearson’s correlation between relative abundance of sequences and related environmental factors.

<table>
<thead>
<tr>
<th></th>
<th>pH</th>
<th>Organic carbon</th>
<th>Organic materials</th>
<th>Na</th>
<th>Ca</th>
<th>Fe</th>
<th>Al</th>
<th>P</th>
<th>K</th>
<th>Mg</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proteobacteria</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>-.523*</td>
<td>.595*</td>
<td>.585*</td>
<td>-.623*</td>
<td>-.249</td>
<td>-.569*</td>
<td>-.627*</td>
<td>-.464</td>
<td>-.617*</td>
<td>-.390</td>
<td>-.423</td>
</tr>
<tr>
<td>$P$</td>
<td>.049</td>
<td>.027</td>
<td>.029</td>
<td>.020</td>
<td>.231</td>
<td>.034</td>
<td>.020</td>
<td>.075</td>
<td>.022</td>
<td>.118</td>
<td>.098</td>
</tr>
<tr>
<td><strong>Gemmatimonadetes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.854**</td>
<td>-.751**</td>
<td>-.896**</td>
<td>.899**</td>
<td>.057</td>
<td>.695**</td>
<td>.844**</td>
<td>.426</td>
<td>.856**</td>
<td>.293</td>
<td>.341</td>
</tr>
<tr>
<td>$P$</td>
<td>&lt;.001</td>
<td>.004</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.434</td>
<td>.009</td>
<td>.001</td>
<td>.095</td>
<td>&lt;.001</td>
<td>.191</td>
<td>.152</td>
</tr>
</tbody>
</table>

* with $P$ value < 0.05
** with $P$ value < 0.01
Figure Legends

Fig. 1. Relative abundance of bacterial 16S rDNA gene sequences from the samples at the Phylum level.

Fig. 2. Relative abundance of bacterial 16S rDNA gene sequences from the samples at the Class level.

Fig. 3. PCoA plot of distance of the microbial communities among the 18 samples based on OTU relative abundance using the Bray-Curtis measure.

Fig. 4. The distance between different sites based on the OTU relative abundance illustrated with a UPGMA tree.
Fig 3.
Fig 3.

PCoA

PC1 - Percent variation explained 18.10%

PC2 - Percent variation explained 13.60%
Fig. 4

SRB

SRA

GVA

GVB

LWH

MCR

TJR

0.1
CHAPTER 4: Discussion and perspective
4.1. Discussion

Pyrosequencing is one of the technologies supplanting Sanger sequencing for research on metagenomics and microbial diversity, with large coverage and sampling depth [214]. The application of pyrosequencing 16S rRNA amplicons to study the microbial community is now broadly used in environmental microbial ecology, and microbial diversity in arid environments is found to be much more abundant than previously surmised. The Titanium-based system is one of the most used pyrosequencing platform, and can produce 500 nt per read. However, data generated by pyrosequencing constantly faces the challenge on how to control the quality and subsequent statistical analyses with such large data sets, including raw sequence cleaning, sequence assignment at different taxonomic levels and clustering into Operational Taxonomic Unit (OTU). There are many different software for pyrosequencing data, as well as online pipelines such as the Mothur, Qiime, and Galaxy. It’s crucial to choose the appropriate methods and tools in the analysis of pyrosequencing data, since different tools can generate different results and thus lead to bias on the estimation of diversity and structure of a microbial community [215]. For the data produced from environmental samples, it is particularly difficult to verify the estimation of microbial community diversity because a large proportion of sequences may be unknown or without a close relative in the databases.

In this chapter, I will discuss the tools and methods we have tested for the analyses of our data, including four important steps: i) quality control of raw reads to remove the noise in the sequences; ii) tools to check and remove chimeras in the data; iii) methods to classify sequences to different taxonomic levels and databases used in the taxonomic assignment of sequences; iii) software to cluster the sequences to OTUs.
4.1.1. Quality Control

The 454 pyrosequencing technology is based on sequencing-by-synthesis, and the chemical combination of nucleotide reagents with the template strand in a reaction generates light signal recorded by a camera. Thus the number and type of nucleotides (T, A, C, G) included in each flow of reaction are estimated [216]. The cyclically reaction and corresponding values estimated through light signals forms the basis for base type and per-base quality score calculations [217]. During the process of pyrosequencing, light intensities may not accurately reflect the real homopolymer length, and long homopolymers result in frequent miscalls, including insertions and deletions of nucleotides [218].

There are two major sources of errors that need to be considered in pyrosequencing data: those of the pyrosequencing reactions and those introduced the PCR amplification [43]. In our study, sequences reads generated by pyrosequencing were extracted and treated on FASTA (and QUAL) or FASTQ files. A quality score is assigned to each base as a phred equivalent related to the base calling error probabilities (P):

\[ Q = -10 \log_{10} P \] [219]

Table 4.1. The Phred quality score related to the base calling error probalities.

<table>
<thead>
<tr>
<th>Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
</tbody>
</table>
Downstream analysis of sequences can be affected by low-quality sequences, artificial sequences (primers and barcodes) and sequence contamination [220]. Thus as the first step of cleaning sequences, we filtered sequences with the following criteria: (i) length less than 200 nt or longer than 600 nt, (ii) mismatch with barcode sequences or more than one mismatch to the primer, and (iii) the presence of homopolymers of > 8 bp in length. Then adaptor sequences were also removed from the sequences. These general cleaning criteria are commonly applied in research using pyrosequencing data, which can be performed using many metagenomics analysis tools such as Mothur and Qiime. In this step, 2-3% of total sequences were removed from our raw data.

Raw sequences generated from pyrosequencing normally have low Q score at the 3’ end (Fig 4.1). Thus, truncate sequences to a specific length or trimming off bases under a specific Q score at the 3’ end before filtering low-quality sequences will help conserve a number of sequences as well as remove low-quality bases, with sacrificing the length of sequences.

Figure 4.1. Quality score distribution of raw sequences

We compared various tools to trim the 3’ end and filter low-quality sequences: a) truncate sequences to the same length (400 nt) using Prinseq [220]; b) trim by Q score
(Q = 25) using ConDeTri [221]; c) trim by Q score (25) in sliding windows with Mothur. The result are demonstrated in Table 4.2, using sequences from one of our desert samples.

Table 4.2. Comparison of different methods for quality control of sequences (original number of sequences is 22320).

<table>
<thead>
<tr>
<th></th>
<th>Truncate at 400 nt</th>
<th>Sliding window</th>
<th>ConDeTri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequences</td>
<td>20378</td>
<td>17868</td>
<td>17049</td>
</tr>
<tr>
<td>Average length of reads</td>
<td>389 nt</td>
<td>279 nt</td>
<td>411 nt</td>
</tr>
</tbody>
</table>

Figure 4.2. Comparison of quality score distribution of sequences trimmed using the three different methods.
The results indicated that the method by truncating sequencing to 400 nt with Prinseq can leave the most sequences, but the base quality at the 3’ end is not as good as the other methods. Using sliding windows with the Mothur generated good quality sequences, while the average length of sequences dropped. Trimming the 3’end of sequences with ConDeTri appears to be the best method among the three in terms of number of sequences, average length and base quality.

The quality control step generally removes approximately 10-15% of total sequences from our samples. Sequences with low Q scores can lead to increased false-positive variant calls, resulting in an inaccurate estimation of sequence diversity. For example, we calculated the OTU numbers and Chao 1 index with sequences before and after trimming low quality bases (Table 4.3).

Table 4.3. Diversity result before and after trimming.

<table>
<thead>
<tr>
<th></th>
<th>Number of sequences</th>
<th>OTU</th>
<th>Chao 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before trimming</td>
<td>5000</td>
<td>1277</td>
<td>6393</td>
</tr>
<tr>
<td>After trimming</td>
<td>5000</td>
<td>673</td>
<td>1287</td>
</tr>
</tbody>
</table>

5000 Sequences are randomly selected (using a command in Pangea pipeline) from one of our desert samples (Maine_1).
4.1.2. Chimera Check

During the process of PCR, when incomplete extension occurs in a round, chimeras are generated which can act as a primer for a different sequence in the next round. Chimeras are thus composed of two or more parent sequences. These chimeras need to be detected and removed from data sets by aligning each sequence against a known reference database. We tested two programs for chimera checking that are most frequently used in research with pyrosequencing of 16S rRNA amplicons: UChime and Decipher.

In UChime, the query sequence is divided into four segments, each of which is aligned to a reference database. The best matched parents as candidates are then used to perform a multiple alignment, and chimera sequences will be reported when two candidate parents have an identity closer to (exceed a predetermined threshold) the query sequence than either candidate alone [47]. In Decipher, the query sequence is first assigned to one of the reference phylogenetic groups. Then a set of 30 nt long segments is formed from the sequence to be aligned with those of a reference database. The chimera sequence will be reported when some fragments have few matches within their own reference phylogenetic group but a large number of matches to another reference group [48]. Decipher has a higher rate of detection of short chimeric ranges (100-250 nt) and complex chimeras with multiple parents, while UChime can detect chimeras formed from closely related parents [48].

First, we used a control sample from mixing *E. coli* and *Deinococcus radiodurans* DNA to amplify in PCR. Before chimera checking we followed the cleaning and quality control steps as mentioned above, with 1424 reads remained. The chimeras found in this sample are shown in Fig 4.3. The Greengenes bacterial 16S rRNA database, version 2013_May (OTU_99) was used as reference sequences in UChime,
and the Decipher online version with the RDP database version 2012 was applied in our test. Chimera sequences found in the control samples are 26 for Uchime and 31 for Decipher, and the 26 chimeras observed by Uchime are detected by Decipher as well. Five chimeras are only detected by Decipher.

![Figure 4.3. Number of chimeras detected by the Uchime and Decipher.](image)

We found more chimera sequences detected by Decipher in the control sample, but with more false positive chimeras, then I further tested the two programs using an environmental sample from the a desert in China (Kumtagh). Sequences in the sample were also treated with cleaning step and quality control with ConDeTri, and 27770 sequences remained. The results are shown in Fig 4.4. We found that 1509 sequences were detected as chimeras by both programs, with 22201 sequences reported as good sequences. UChime and Decipher detected an additional 2550 and 1502 chimeras, respectively.
To further test the accuracy of the two programs, we classified sequences in the four parts of Fig 4.4 respectively with RDP II classifier and the Silva Database version 119. Table 4.4 demonstrates the proportion of total sequences in each part classified at different taxonomic levels.

Table 4.4. Percentage of potential non-chimeric sequences classified at different taxonomic levels.

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>U</th>
<th>D</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>63%</td>
<td>86%</td>
<td>89%</td>
<td>99%</td>
</tr>
<tr>
<td>Class</td>
<td>56%</td>
<td>81%</td>
<td>83%</td>
<td>98%</td>
</tr>
<tr>
<td>Order</td>
<td>43%</td>
<td>75%</td>
<td>64%</td>
<td>92%</td>
</tr>
<tr>
<td>Family</td>
<td>35%</td>
<td>53%</td>
<td>46%</td>
<td>78%</td>
</tr>
<tr>
<td>Genus</td>
<td>11%</td>
<td>35%</td>
<td>30%</td>
<td>50%</td>
</tr>
</tbody>
</table>

We found the percentage of classified sequences is more in the non-chimera group (Z) than that of the other groups, and chimera sequences detected by both programs are more poorly classified. We also found at the genus level that Decipher removed most of the sequences classified to Tunicatimonas as chimeras in the Kumtagh sample,
which were in fact good sequences according to Blast results but absent from the reference database in Decipher. Thus, both the programs itself and the reference database used for chimera checking are essential in detecting chimeras. Based on these results, we removed chimeras detected by both programs (UChime and Decipher = X) in our samples for downstream analysis.
4.1.3. Taxonomic Classification of Sequences

Pyrosequencing generates many of 16S rRNA reads per sample. After sequences are cleaned for low quality bases and chimeras, a further classification step using specific 16S rRNA databases is required. There are two major methods to taxonomically classify 16S rRNA fragments: similarity-based tools such as BLAST and alignment-based taxonomic assignment such as RDP classifier [222]. The BLAST based method assigns the query 16S rRNA sequence to the best significant hit to a reference sequence in the database, which is adopted in several metagenomics analysis pipeline, such as MG-RAST [223] and CAMERA [224]. Since the BLAST approach tends to assign the query sequence to known taxonomic groups, it may lead to a large number of sequences remaining unclassified especially for environmental samples, that may belong to novel genera or even new phyla [225]. The RDP Classifier is based on the naïve Bayesian classification of 8-mer words belonging to the query sequence, which classifies the sequence to a similar taxonomic lineage according to the frequencies of 8-mer words identified to a taxonomic lineage in the database [226].

Reference databases are essential to the classification of sequences. There are three major 16S rRNA databases applied in the studies of environmental microbial communities: Greengenes, RDP and Silva. We summarized the taxonomic information of the three 16S rRNA database in Table 4.5 to show the percentage of sequences classified to each taxonomic level in the database. The RDP database has most of their sequences classified to a deep level (genus), but includes the least number of sequences.
Table 4.5. Summary of sequences in each database assigned to different taxonomic level

<table>
<thead>
<tr>
<th></th>
<th>Greengenes (198510)$^a$</th>
<th>RDP (9665)$^b$</th>
<th>Silva 119 (137878)$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of sequences</td>
<td>Percentage</td>
<td>Number of sequences</td>
</tr>
<tr>
<td>phylum</td>
<td>198277</td>
<td>99.9%</td>
<td>9665</td>
</tr>
<tr>
<td>class</td>
<td>196138</td>
<td>98.8%</td>
<td>9665</td>
</tr>
<tr>
<td>order</td>
<td>186786</td>
<td>94.1%</td>
<td>9496</td>
</tr>
<tr>
<td>family</td>
<td>153246</td>
<td>77.2%</td>
<td>9040</td>
</tr>
<tr>
<td>genus</td>
<td>91122</td>
<td>45.9%</td>
<td>8833</td>
</tr>
</tbody>
</table>

$^a$Greengenes database version 2013_May, 99_otus, with 198510 bacterial sequences

$^b$RDP database version 2012, with 9665 bacterial sequences

$^c$Silva database version 119, 99_non-redundant, with 137878 bacterial sequences.

We also tested the three databases using RDP Classifier with a sample from desert in Utah (Table 4.6). The Silva database_v119 presented better results with a higher proportion of sequences classified to the genus level, while the Greengenes database appeared to be better at higher taxonomic level (Phylum to Order).

Table 4.6. Classification result of a sample from desert in Utah

<table>
<thead>
<tr>
<th></th>
<th>Percentage of classified sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Silva</td>
</tr>
<tr>
<td>Phylum</td>
<td>89.4%</td>
</tr>
<tr>
<td>Class</td>
<td>87.3%</td>
</tr>
<tr>
<td>Order</td>
<td>82.2%</td>
</tr>
<tr>
<td>Family</td>
<td>73.1%</td>
</tr>
<tr>
<td>Genus</td>
<td>55.4%</td>
</tr>
</tbody>
</table>
The RDP Classifier is currently the most popular software used in the classification of 16S rRNA sequences in taxonomic studies of microbial communities [222]. To test the different classifier, we randomly selected 1000 sequences from the Gold_database (full length of 16S rDNA, non-chimera) downloaded in the Usearch website, whose genomic information is well characterized. Three different classifiers tested in our study are RDP II Classifier, Uclust implemented in Qiime and Crest (online version). The results were shown in Table 4.7, and RDP II presented better taxonomic classifications at each level.

Table 4.7. Summary of classification using different classifiers with 1000 sequences randomly selected from gold_database downloaded from Usearch.

<table>
<thead>
<tr>
<th></th>
<th>RDP II</th>
<th>Uclust</th>
<th>Crest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>100.00%</td>
<td>99.90%</td>
<td>99.90%</td>
</tr>
<tr>
<td>Class</td>
<td>100.00%</td>
<td>99.80%</td>
<td>93.90%</td>
</tr>
<tr>
<td>Order</td>
<td>99.70%</td>
<td>99.30%</td>
<td>85.90%</td>
</tr>
<tr>
<td>Family</td>
<td>98.40%</td>
<td>81.10%</td>
<td>64.60%</td>
</tr>
<tr>
<td>Genus</td>
<td>96.20%</td>
<td>74.30%</td>
<td>35.80%</td>
</tr>
</tbody>
</table>

Based on these results, the RDP II Classifier with the Silva 16S rRNA database v_119 database were generally applied for the classification of sequences in our samples.
4.1.4. OTU Clustering

Since the majority of microbial species and their 16S rRNA genes have not been taxonomically classified, diversity estimations using pyrosequencing data have to be frequently measured as the number of OTUs in a sample. OTUs are defined by applying a distance threshold corresponding approximately to a specific taxonomic level: 1-3% are typically used for species, 5% for genera, 15% for classes, etc. [227-230]. The number of OTUs are broadly used in microbial diversity studies to reflect the number of species in a sample. Further analyses are also based on the number of OTUs, such as the diversity Chao 1 and Shannon indices, as well as multiple sample comparisons on the microbial community structure.

To cluster 16S rDNA sequences to OTUs, the major methods can be categorized as taxonomic dependent or independent algorithms [231]. The taxonomic dependent methods such as RDP and Mothur, align sequences to a template database to calculate the distance among each other. In contrast, taxonomic independent methods compare sequences against each other to form a distance matrix, like Clustal and MUSCLE, or construct consensus sequences representing each cluster based on a greedy algorithm such as CD-HIT and Uclust.

In our study, we compared four different programs that have been used with an control sample (1393 sequences after quality control and removing chimeras) containing only two species (from E. coli and Deinococcus. radiodurans), including Uparse, ESPRIT-Tree, RDP and Mothur pipelines. Uparse is based on the CD-HIT greedy heuristic algorithm that can estimate the similarity between two sequences without performing multiple alignments of all pairs of sequences [60]. We chose representative software of each category to test: Uparse and ESPRIT-Tree for taxonomic independent algorithms; RDP and Mothur for taxonomic dependent
algorithms (Table 4.8).

Table 4.8. Number of OTUs clustered using different software at a distance 3%.

<table>
<thead>
<tr>
<th></th>
<th>Number of OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uparse</td>
<td>2</td>
</tr>
<tr>
<td>ESPRIT-Tree</td>
<td>3</td>
</tr>
<tr>
<td>RDP</td>
<td>4</td>
</tr>
<tr>
<td>Mothur</td>
<td>18</td>
</tr>
</tbody>
</table>

The algorithm implemented in the Mothur pipeline generated 18 OTUs, much more than the real number of species in the control sample (two species). We also extracted the representative sequences in each OTU and used BLAST against the NCBI 16S rRNA database. We found that the two representative sequences in Uparse are correctly assigned to the two species. Thus Uparse was chosen for the downstream analysis of our sample in deserts.
4.2. Perspective

In this study, we demonstrated the bacterial diversity and community structures of surface soil in deserts in the Utah State and the Desert of Maine using pyrosequencing of 16S rDNA amplicons. We built our pipeline for the analysis of 16S rDNA pyrosequencing data by combining several existing metagenomics tools. We also searched for correlations between environmental factors and bacterial diversity in the two deserts. Further studies stand be performed to reveal the microbial community and its ecological function in deserts:

a) Some special bacterial groups, such as members of *Gemmatimonadetes* and *Alphaproteobacteria*, whose population we found significantly correlated to environmental factors in desert soils (pH and concentration of organic matters), and some extremophile genera observed in samples, can be selected for further study using qPCR to more precisely quantify their presence in many desert soils.

b) Omics-based technologies, such as sequencing of total DNA and RNA in sample sites have the potential to yield major advances in our understanding of the functional capacity of microbial communities and their adaptive potential [232]. This research can help us to understand how microbial communities function in deserts their contribution to biogeochemical processes.

c) Other microorganisms, such as viruses, fungi and Archaea play important roles in the deserts. NGS technologies can be applied to study their community structures. There are 18S rDNA sequence database available for eukaryote community analysis. It should be noted that the reference database for an Archaeal community study needs to be applied with caution, since the taxonomic classification of Archaeal 16S rDNA sequences are not as well organized as
those in bacterial 16S rDNA sequences.
References


66. Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr 5: 34.


79. F.D. Eckardtldt, , K. Soderbergb, c, L.J. Coope, A.A. Mullera, K.J. Vickeryd, R.D.


100. Noy-Meir I (1979) Structure and function of desert ecosystems. Israel Journal of
101. Webb RH, Fenstermaker, L.F., Heaton, J.S., Hughson, D.L., McDonald, E.V.,
Sustainability.
102. Chapin FS, Sala, Osvaldo E., Huber-Sannwald, Elisabeth (2001) Global
Biodiversity in a Changing Environment.
103. Rundel PW (1978) Ecological Relationships of Desert Fog Zone Lichens. The
Bryologist 81: 277-293.
University Press.
University Press.
Geomorphology: Process, Form and Change in Drylands, 3rd Edition: John
Wiley & Sons.
109. Hui Wanching Jacquie, Cook Benjamin I., Ravi Sujith, Fuentes José D., Paolo
spring dust storm frequency of northern China in relation to antecedent and
concurrent wind, precipitation, vegetation, and soil moisture conditions.
of Climate Change.
American Meteorological Society 78: 2197-2213.


124. Direito SOLE, Pascale; Marees, Andries; Staats, Martijn; Foing, Bernard; Röling, Wilfred F. M. (2011) A wide variety of putative extremophiles and large
beta-diversity at the Mars Desert Research Station (Utah). Special issue: Astrobiology field research in Moon/Mars analogue environments 10: 16.


Microbiology Ecology 21: 121-130.


desert of Tataouine: an extreme environment that hosts a wide diversity of
microorganisms and radiotolerant bacteria. Environmental Microbiology 8:
514-525.
Evidence of species recruitment and development of hot desert hypolithic
Streptomyces bullii sp. nov., isolated from a hyper-arid Atacama Desert soil.
Antonie van Leeuwenhoek 103: 367-373.
9: 791-802.
bacteriophages from a Mesquite Flats site of the Death Valley desert. Antonie
van Leeuwenhoek 103: 1329-1341.
169. Subhash Y, Sasikala C, Ramana Ch V (2014) Pontibacter ruber sp. nov. and
Pontibacter deserti sp. nov., isolated from the desert. Int J Syst Evol Microbiol
64: 1006-1011.
Whole-Genome Sequences of an Aerobic Anoxic Phototroph,
Blastomonas sp. Strain AAP53, Isolated from a Freshwater Desert Lake in
Inner Mongolia, China. Genome Announcements 1.
accumulation of salt, calcium carbonate accumulation, and surface exposure.
104: 288.


Conservation 183: 4-18.


Titre : La diversité bactérienne dans les sols de surface de San Rafael Swell (Utah, USA) et le Desert de Maine (USA)

Synthèse en français :


Le désert du Maine situé dans le nord-est Etats-Unis est une étendue de boue glaciaire, entourée par une forêt de pins. Nous avons observé que les échantillons de sol provenant du désert du Maine présentent une diversité bactérienne singulière, avec une prédominance de *Proteobacteria* et *Actinobacteria*. Les bactéries du genre le plus abondant, *Acidiphilium*, représentent 12,5% du total des séquences d'ADNr 16S.

Le Désert de l'Utah présente des caractéristiques géographiques qui ressemblent à Mars. En effet il est caractérisé par la présence de collines de couleur rouge et de sols constitués de grès. Les phylums prédominants sont les *Proteobacteria, Actinobacteria, Bacteroidetes* et *Gemmatimonadetes*. Les genres les plus abondants dans nos échantillons sont *Cesiribacter, Lysobacter, Adhaeribacter, Microvirga* et *Pontibacter*. Mais de façon notable, il semble que l'abondance relative des *Alphaproteobacteria* et des *Gemmatimonadetes* est significativement corrélée aux certains facteurs environnementaux des sols, par exemple de pH et des concentration des matières organiques.
Titre : La diversité bactérienne dans les sols de surface de San Rafael Swell (Utah, USA) et le Desert de Maine (USA)

Mots clés : Diversité bactérienne, Désert, Pyroséquençage


Le désert du Maine situé dans le nord-est Etats-Unis est une étendue de boue glaciaire, entourée par une forêt de pins. Le sol de ce désert possède les caractéristiques d’un sable avec de très faibles capacités de rétention d'eau, d'une rétention des éléments nutritifs, ainsi qu’une valeur de pH relativement faible (pH 5,09). Les échantillons provenant de ce site présentent donc des propriétés particulièrement intéressantes à étudier en lieu avec la diversité bactérienne. Deux échantillons de sable de la surface du désert du Maine ont été obtenus, et le pyroséquençage des gènes d'ADNr 16S obtenus après amplification par PCR à partir de l'ADN total extrait a été utilisé pour évaluer la diversité bactérienne, la structure de la communauté bactérienne et l'abondance relative des principaux taxons. Nous avons observé que les échantillons de sol provenant du désert du Maine présentent une diversité bactérienne singulière, avec une prédominance de Proteobacteria et Actinobacteria. Les bactéries du genre le plus abondant, Acidiphilium, représentent 12,5% du total des séquences d'ADNr 16S. Au total, 1 394 OTU ont été comptabilisées. En comparant les résultats de notre population
bactérienne avec des études portant sur des sols avec caractéristiques similaires, nous avons constaté que les échantillons du Maine contiennent une faible diversité du phylum *Acidobacteria* que les sols acides des certains forêts, et moins de *Firmicutes* ainsi que plus de *Proteobacteria* que les sols des déserts oligotrophes.

Le Désert de l'Utah présente des caractéristiques géographiques qui ressemblent à Mars. En effet il est caractérisé par la présence de collines de couleur rouge et de sols constitués de grès. Les sites d'échantillonnage couvrent le Goblin Valley State Park et autour, notamment sur le plateau du Colorado. Avec des approches similaires à ceux utilisés pour le désert du Maine, des corrélations entre facteurs environnementaux (paramètres physico-chimiques) et diversité de structure des communautés bactériennes obtenus, ont été étudiées. Les phylums prédominants sont les *Proteobacteria, Actinobacteria, Bacteroidetes* et *Gemmatimonadetes*. Les genres les plus abondants dans nos échantillons sont *Cesiribacter, Lysobacter, Adhaeribacter, Microvirga* et *Pontibacter*. Mais de façon notable, il semble que l'abondance relative des *Alphaproteobacteria* et des *Gemmatimonadetes* est significativement corrélée aux certains facteurs environnementaux des sols, par exemple de pH et des concentration des matières organiques.
Title: The bacterial communities of sand-like surface soils of the San Rafael Swell (Utah, USA) and the Desert of Maine (USA)

Keywords: Bacterial diversity, Deserts, Pyrosequencing

Abstract: Aridity is the dominant climatic factor over approximately 30% of the land surface of the world. Research concerning microbial populations in two U.S. deserts has been performed to determine the diversity of these bacteria. Pyrosequencing-based profiling of 16S rRNA amplicons from surface soils of sand samples can provide key insights into the structure of bacterial communities and their diversity. In this study, we demonstrated the bacterial diversity and community structures of surface soil in the Colorado Plateau in the Utah State and the Desert of Maine using pyrosequencing of 16S rRNA amplicons. We built our pipeline for the analysis of 16S rRNA pyrosequencing data by combining several existing tools of metagenomics. We also examined correlations between certain environmental factors and bacterial diversity in the two deserts.

The Desert of Maine is a tract of glacial silt, surrounded by a pine forest, in the state of Maine located in the northeastern USA. The soil of the Desert of Maine has a sandy texture with poor water holding abilities, nutrient retention capabilities and a relatively low pH value (pH 5.09). Samples from this site thus present an interesting place to examine the bacterial diversity in mineral sandy loam soils with an acidic pH and low concentrations of organic materials. Two surface sand samples from the Desert of Maine were obtained, and pyrosequencing of PCR amplified 16S rDNA genes from total extracted DNA was used to assess bacterial diversity, community structure and the relative abundance of major bacterial taxa. We found that the soil samples from the Desert of Maine showed high levels of bacterial diversity, with a predominance of members belonging to the Proteobacteria and Actinobacteria phyla. Bacteria from the most abundant genus, Acidiphilium, represent 12.5% of the total 16S rDNA sequences. In total, 1394 OTUs were observed in the two samples, with the
number of common OTUs observed in both samples being 668. By comparing our bacterial population results with studies on related soil environments, we found that the samples contained less *Acidobacteria* than soils from acid soil forests, and less *Firmicutes* plus more *Proteobacteria* than soils from oligotrophic deserts.

Deserts in Utah has geographic features that resemble Mars, characterized by red-colored hills, soils and sandstones. Our sample sites cover the Goblin Valley State Park and nearby regions on the Colorado Plateau. We also examined physicochemical parameters of soil from the sample sites to investigate correlations between bacterial community structure and environmental drivers. The predominant phyla of the samples represent members of the *Proteobacteria, Actinobacteria, Bacteroidetes,* and *Gemmatimonadetes.* The most abundant genera in our samples are *Cesiribacter, Lysobacter, Adhaeribacter, Microvirga* and *Pontibacter.* We found that the relative abundance of Alphaproteobacteria and Gemmatimonadetes are significantly correlated to some environmental factors of soils, such as pH and concentration of organic matters.