*Author:*
**Nash, Maisie V**

*Title:*
**Metagenomic insights into microbial communities in proglacial landscapes**

# Metagenomic insights into microbial communities in proglacial landscapes

Maisie Victoria Nash

A thesis submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Science

School of Geographical Sciences

April 2019

University of BRISTOL

Word count: 46,231

# Abstract

Environmental DNA analysis using metagenomics can provide an insight into the taxonomy and functional potential of microbial communities ex *situ*, without the need for culturing or DNA amplification. However, metagenomics has had limited application to environmental microbial ecology, in particular, to microbial communities in proglacial regions. This thesis aims to contribute to the body of literature on environmental metagenomics through evaluating assemblers for soil microbial ecologists, and subsequently applying metagenomics to investigate microbial communities in proglacial environments.

Assembly of metagenome sequencing reads can improve sequence alignment to taxonomic and functional databases, thereby improving ecological conclusions. However, limited guidance is available for assembler choice by microbial ecologists. The first study in this thesis compares assemblers for soil metagenome data, demonstrating the importance of assembler evaluation and parameterization. The guidance produced was applied to investigate microbial communities in proglacial regions, including fjords and forefields. Proglacial forefields present a unique opportunity to understand microbial colonization in land exposed by glacier retreat. Here, metagenomics was used to investigate microbial diversity and functional potential during forefield succession, alongside comparing the diversity of nitrogen-fixing bacteria between Arctic forefields. This work contributes to our understanding of Arctic microbial ecology, which has significance given the continued exposure of forefield soils during global warming. In addition, metagenomics was used to investigate microbial communities in oligotrophic, dark, saline fjord waters, fed by glacial meltwater. This work highlights the potential of metagenomics to understand uncultured microbial samples and demonstrate areas for further analysis, such as targeting novel genomes.

This thesis has contributed to the literature on metagenomics by providing methodological guidance for microbial ecologists, alongside enhancing understanding of microbial diversity in proglacial regions. It is hoped that this work will inspire others to use metagenomics to explore uncultured microbial samples and to target further analysis or exploration for unique genomes.

# Acknowledgements

I would like to thank the following people who have provided guidance and support over the course of my PhD. Firstly, I would like to thank my advisors, Patricia Sánchez-Baracaldo who has been a fantastic collaborator and a huge source of support, Alex Anesio for providing me with the opportunity to enter the world of microbiology, and Gary Barker for pushing me to become a better bioinformatician. Thank you to NERC for funding my project through a $GW4^+$ DTP scholarship and the PISCES project for supporting my fieldwork. Many thanks to my cohort of PhD students and the Browns community, in particular Nathan Christmas for being my technical help desk, and Steve Chuter, Ale Urra and Claire Donnelly for supporting me through the ups and downs. I would like to thank my fieldwork team, including Jon Hawkings, Rebecca Huggett, Rory Burford, Alex Beaton, Helena Pryer, Hong Chin and all those on the PISCES project. Thank you to my team and coaches at CrossFit 605 who have kept my competitive spirit alive and motivated me to improve in all aspects of my life. I would like to thank my best friend, Alice Haworth and my sisters, Amy and Libby, for supporting me through my crazy pursuit in science. Finally, I would like to thank my parents, Trevor and Julie, without their support I would never have made it this far, and their work ethic and positive attitude inspires me every day.

# Authors declaration

I declare that the work was carried out in accordance with the requirements of the University’s regulations and Code of Practice for Research Degree programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate’s own work. Work done in collaboration with, or the assistance of others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: ……………………………………………………………. Date:…………………

# Table of Contents

# List of Figures

# List of Tables

# Key terms

| | |
|---|---|
| **Allochthonous** | Originating outside the location it is identified in |
| **Anoxic** | Without oxygen |
| **Assembly bubbles** | Two de Bruijn assembly paths that begin and end at the point on the graph |
| **Assembly completeness** | How much of the input DNA (reads) are incorporated in the output assembly |
| **Assembly contiguity** | Low long DNA contigs are in an assembly |
| **Assembly coverage** | The percentage of organisms, genes or sequences captured by sequencing or assembly |
| **Autochthonous** | Formed in situ |
| **Autotroph** | An organism which can produce complex organic compounds from inorganic material |
| **Bioinformatics** | The computational analysis of complex genetic data e.g. from genomes or metagenomes |
| **Chemo-lithotroph** | An organism which can use inorganic substances for growth |
| **Chemo-organotroph** | An organism which obtains its energy from organic molecules |
| **Chronosequence** | A set of soil sites which share similar attributes but differ in soil development due to age |
| **Contig** | A consensus set of DNA reads, merged into a continuous sequence using overlaps |
| **Diazotroph** | A microorganism which can fix dinitrogen ($N_2$) to ammonia ($NH_3$) |
| **Facultative anaerobes** | An organism which can respire anaerobically or aerobically |
| **Forefield** | Land immediately in front of a glacier terminus, exposed by ice retreat |
| **Heterotroph** | An organism which uses fixed organic substances for growth |
| **Homopolymer regions** | Repeat regions in genomic DNA |
| **Hypoxic** | Deprived of oxygen |
| **K-mer** | A short DNA sequence that reads are fragmented in to prior to assembly |
| **Metagenome** | Environmental genetic material (DNA) sourced directly from a sample, without amplification |
| **Methanogen** | Microbes which produce methane ($CH_4$) during anaerobic respiration |
| **N50** | The length (N) over which 50% of contigs are above this length |
| **Oligotrophic** | Very low nutrient |
| **Phototroph** | An organism which utilizes energy from the sun to create organic molecules |
| **Phylogeny** | The evolutionary history of a group of organisms and their relationship to one another |
| **Primary succession** | Gradual development of land through microbial and plant colonization |
| **Proglacial** | The area in front of a glacier terminus (ocean, lake, soil) |
| **Proteome (proteomics)** | The entire set of proteins expressed by a genome or metagenome |
| **Read coverage** | The frequency a sequenced DNA read appears in a metagenome |
| **Scaffold** | A non-contiguous set of contigs and gaps, formed to create a draft genome |
| **Taxonomy** | The classification of organisms |
| **Transcriptome** | All the mRNA expressed by a genome or metagenome |

# Abbreviations

| | |
|---|---|
| **16s rRNA** | 16s ribosomal ribonucleic acid |
| **ANOVA** | Analysis of variance |
| **AOA** | Ammonia oxidizing archaea |
| **AS** | Alignment score |
| **ATP** | Adenosine triphosphate |
| **BLAST (n&p)** | Basic local alignment search tool (nucleotide & protein) |
| **BP** | Base pairs |
| **CTD sensor** | Conductivity, temperature and depth sensor |
| **ddNTP's** | Dideoxynucleotides triphosphates |
| **DNA** | Deoxyribonucleic acid |
| **dsrAB gene** | Dissimilatory sulfite reductase |
| **EPS** | Exopolymeric substances |
| **INDELS** | Insertions and deletions |
| **LCA** | Last common ancestor |
| **ML** | Midtre Lovénbreen glacier |
| **NCBI** | National Centre for Biotechnology Information |
| **NGS** | Next generation sequencing |
| **Nif genes** | Nitrogen fixation genes |
| **NPP** | Net primary productivity |
| **OLC** | Overlap layout consensus |
| **OTU** | Operational taxonomic unit |
| **PCR** | Polymerase chain reaction |
| **PP** | Primary productivity |
| **RB** | Rabots glacier |
| **RL** | Russell glacier |
| **SD** | Standard deviation |
| **ST** | Storglaciären glacier |
| **TN** | Total nitrogen |
| **TOC** | Total organic carbon |
| **UNSECO** | United Nations Educational, Scientific and Cultural Organization |

# Chapter 1: Literature Review

## 1.1 Introduction to glacial systems

At present, it is estimated that ice sheets and glaciers account for ~10% of global land coverage (16 million km$^2$) (Knight, 1999). These systems comprise of subglacial (beneath glacier), englacial (within glacier), supraglacial (glacier surface) and proglacial (in front of glacier) environments (Benn and Evans, 2014). Ice sheets and glaciers have been the focus of scientific research over recent years due to ongoing ice melt and glacier retreat with warming global temperatures (IPCC, 2013). The importance of these regions with regard to meltwater fluxes, sea level rise, sediment and nutrient transport and global surface albedo have been documented (Tranter *et al.,* 2002; Hood *et al.,* 2009; Wadham *et al.,* 2010; Wadham *et al.,* 2013; Dutton *et al.,* 2015). The importance of glaciers and ice sheets in global biogeochemical cycles has also been noted, through supplying limiting nutrients to downstream fjord and ocean ecosystems, such as iron, carbon and nitrogen, and thereby stimulating productivity in near shore regions (Tranter *et al.,* 1994; Stratham *et al.,* 2008; Lawson *et al.,* 2014; Hawkings *et al.,* 2015).

However, the study of microbial communities, in both glaciers and ice sheets has only come to the forefront of research in recent years (Hodson*,* 2006; Hodson *et al.,* 2008; Anesio *et al.,* 2009). The potential implications of these communities on global biogeochemical cycles, such as carbon, nitrogen and methane, are substantial (Anesio *et al.,* 2009; Boyd *et al.,* 2010; Stibal *et al.,* 2012; Telling *et al.,* 2012; Wadham *et al.,* 2012). Most research on microbial communities in glacial habitats has focused on supraglacial environments, such as cryoconite holes and subglacial sediments (Anesio *et al.,* 2009; Boyd *et al.,* 2010; Telling *et al.*, 2010). However, limited research has investigated microbial community diversity and function in proglacial regions, such as forefields and fjords (Duc *et al.,* 2009; Zumsteg *et al.,* 2013). Thus far, these regions have been shown to harbour diverse, active communities, and may therefore have vital roles in local biogeochemical cycles (Duc *et al.,* 2009). Additionally, these microbial communities may be influenced by warming temperatures and increased glacial melt, in line with climate change (Alison and Tresder, 2008; Rinnan *et al.,* 2009).

## 1.2 Microbial communities in glacial systems

In early work, glaciers were considered purely abiotic systems, due to the low temperatures, high UV exposure and the oligotrophic nature of the glacial environments (Collins, 1979; Raiswell, 1984). However, recent developments have identified active and diverse microbial communities in supraglacial, subglacial and proglacial habitats (Skidmore *et al.,* 2000; Sigler and Zeyer, 2002; Bhatia *et al.,* 2006; Stibal *et al.,* 2006; Malard and Pearce, 2018). A wide range of organisms have been identified, such as diazotrophs, methanogens, heterotrophs and chemolithoautotrophs (Skidmore *et al.,* 2000; Christner *et al.,* 2008; Hodson *et al.,* 2008; Stibal *et al.,* 2012). The adaptation of these organisms to survive in such extreme conditions, and their importance in both local and global biogeochemical cycles has sparked a wave of research interest (Mindl *et al.,* 2007; Wadham *et al.,* 2013). These organisms can often be found living across environmental gradients, for example at stages of increasing soil development in glacier forefields (Bradley *et al.,* 2014) or along salinity gradients in fjords draining glacial meltwater (Gutiérrez *et al.,* 2015). Consequently, it is interesting to understand the diversity of these communities, how they contribute to biogeochemical cycles, and how this varies over environmental gradients. Investigating microbial communities in proglacial regions is particularly important, as these may be modified with climate change, for example, increased glacial meltwater fluxes may modify salinity balances in fjords (Davila *et al.,* 2002). Here, focus is placed on proglacial microbial communities in forefield soils and fjord sediments (Figure 1.1).

## 1.3 Proglacial forefields and microbial communities

The proglacial forefield of land terminating glaciers facilitates the growth of a diverse range of microorganisms, due to the presence of proglacial rivers, lakes, soils and vegetation, all of which harbour distinct, active communities (Liu *et al.,* 2006; Duc *et al.,* 2009; Reddy *et al.,* 2009; Zumsteg *et al.,* 2013; Figure 1.1). Proglacial rivers are important for the export of labile organic matter and nutrients to downstream ecosystems (Hood & Scott, 2008; Hood *et al.,* 2009; Hawkings *et al.,* 2014; Hopwood *et al.,* 2014).

Proglacial soils have been the focus of an abundance of plant-based research, in particular, looking at the succession of plant species on newly exposed soil, following glacier retreat (Frenot *et al.,* 1998; Hodkinson *et al.,* 2003; Breen and Levesque, 2006). However, more recently, these soils have been the focus of microbiologists, investigating how microbial communities are initially established, and their subsequent role in soil, nutrient and plant development (Schutte *et al.,* 2009; Göransson *et al.,* 2011; Wojcik *et al.,* 2019). These microbial communities are present across an environmental gradient of soil succession, and therefore may be diverse in taxonomy and function across the proglacial region (Bradley *et al.,* 2014). The following section will discuss the research surrounding proglacial soil succession in depth, and in particular, its application within microbiology.

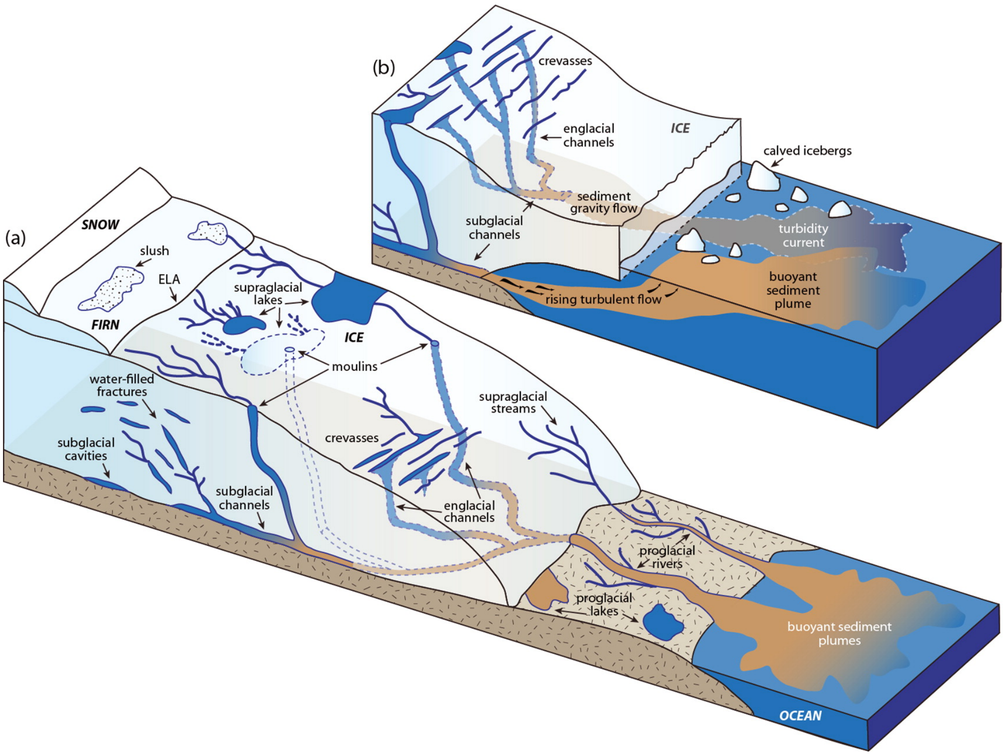


***Figure 1.1****: Schematic of glacier forefields (a) land terminating glacier with forefield soils and (b) marine terminating glacier with proglacial fjord. Source: Chu et al., (2014).*

### 1.3.1 Using forefields to investigate microbial succession

Terrestrial soils, exposed following glacial retreat, pose an interesting opportunity to study initial soil colonisation and succession (Edwards and Cook, 2015). As these soils have been beneath ice for thousands of years, soil development is limited, alongside the absence of established flora or microbial communities (Tscherko *et al.,* 2003). Furthermore, low temperatures and slow weathering rates mean that soil development occurs over longer timescales, and therefore the successional pathways are more identifiable (Bradley *et al.,* 2014). Consequently, it is possible to use these soils to investigate how microbial communities first colonise, and subsequently advance over time, whilst modifying soil development (Schulz *et al.,* 2013). This is typically carried out by utilizing a chronosequence approach, whereby a space-for-time substitution is used, moving away from the glacier terminus along a perpendicular transect (Tscherko *et al.,* 2003). The bulk of chronosequence-based studies have focused on the succession of soil structure and plant matter, whilst limited attention has been paid to the advance of microbial communities (Edwards and Cook, 2015; Matthews and Vater, 2015).

Whilst chronosequence studies are fundamental to our current understanding of soil succession, it is important to acknowledge that chronosequences only show a record for a single location and time point (Fastie, 1995). Forefields are heterogeneous, with many micro-environments, soil structure disparities, differences in water and nutrient availability, and subsequently variations in floral and faunal diversity, all occurring within a single foreland (Duc *et al.,* 2009). Furthermore, these environments may change with seasonal environmental modifications or disturbances (Fastie, 1995; Bradley *et al.,* 2015). Consequently, considerations must be taken when extrapolating information obtained from a single chronosequence to other environments.

### 1.3.2 Plant succession and soil development

Investigating soil development is key to understanding how terrestrial environments change over time, the driving factors behind this, and the causes of heterogeneous succession both within and between environments (Hodkinson *et al.,* 2003). Soil development is mediated by key environmental factors, such as the local bedrock and topography, the climate, the time available, alongside both plant and microbial communities (Schulz *et al.,* 2013). Cold environments, which exhibit slow rock weathering rates, are preferentially used for the study of soil succession, due to the slow (and therefore traceable) developments over chronosequences (D'Amico *et al.,* 2014; Bradley *et al.,* 2015). Much of the research on succession has been focused on plant establishment and development (Frenot *et al.,*1998; Hodkinson *et al.,* 2003).

The key trends identified with soil succession are an increase in nutrients, organic carbon and reduction in pH, in line with the development of vascular plants (Ohtonen *et al.,* 1999; Strauss *et al.,* 2009; Knelman *et al.,* 2012). Plants have been deemed as fundamental in enhancing soil stabilisation, modifying the biogeochemical properties of soils (i.e. building nutrient and organic matter pools through litter and root exudates) and subsequently facilitating the colonisation of higher plant communities (D'Amico *et al.,* 2014; Matthews and Vater, 2015).

Soil and vegetation development do not occur at the same rate, nor follow the same successional pattern between different environments (Hodkinson *et al.,* 2003). This relates to the initial soil conditions, local climatic factors and acting disturbances (Moreau *et al.,* 2008; Sattin *et al.*, 2009). Some plant communities may be rapidly established, whilst for others, typically in cold, nutrient limited environments, plant free periods exceeding 50 years may be identified (Sattin *et al.,* 2009).

### 1.3.3 Microbial Succession in glacial forefields

Microbial succession in newly exposed soils is a more recent subject of interest, in comparison to the study of soil structure and plant colonisation (as reviewed in Bradley *et al.,* 2014). Microbial communities play an important role in developing soil structure and biogeochemical cycles, therefore influencing soil physicochemical characteristics and vegetation (Tscherko *et al.,* 2003; Hahn and Quideau, 2013). Microbial communities often act as the primary colonisers of forefield soils exposed by glacier retreat (Schmidt *et al.,* 2008). Examining this initial colonisation is key for our understanding of how life may first be established and may act as analogous to our understanding of extra-terrestrial life (Bradley *et al.,* 2015). Furthermore, initial colonisation, and subsequent succession is important for highlighting the pathways to fertile soil, the key factors involved in the heterogeneity between environments and if these may be modified with global climate change (Schulz *et al.,* 2013). Much of the research surrounding the importance of initial microbial colonisers has been based on glacial forefields, in locations such as the Damma Glacier, Switzerland and Midtre Lovénbreen glacier, Svalbard (Mindl *et al.,* 2007; Schmidt *et al.,* 2008; Schulz *et al.,* 2013).

The pioneer microbial colonisers may be significant in the development of soil nutrient stocks, regulating soil pH and promoting stability for plant colonisation through the release of EPS (Sattin *et al.,* 2009; Schulz *et al.,* 2013; Wojcik *et al.,* 2019). This may facilitate the colonisation of further microbial communities and higher plant life, which depend on the presence of labile nutrient pools and a degree of soil stability (Chapin *et al.,* 1994; Hahn and Quideau, 2013). The release of nutrients is particularly significant in extreme locations, such as glacial forefields, whereby initial soils are likely to be oligotrophic (Chapin *et al.,* 1994). However, the significance of the initial colonizers in developing labile nutrient pools has been debated (Nicol *et al.,* 2005). Jumpponen (2003) identified aeolian deposition as the most significant source of nutrients to newly exposed soils, alongside nutrients present in the initial soils themselves, whilst suggesting that microbial communities were dormant (Jumpponen, 2003). However, further research has identified the active contribution of initial microbial colonisers to nutrient stocks during the first 20 years of colonisation (Schmidt *et al.,* 2008). The action of photoautotrophic cyanobacteria and free living diazotrophs can build carbon and nitrogen pools in the soils, however aeolian deposition may also be a significant contribution to nutrient stocks (Nicol *et al.,* 2005; Schmidt *et al.,* 2008).

Furthermore, the composition of the initial coloniser communities has been the subject of debate (Kastovska *et al.,* 2005). One research body identifies autotrophic organisms as the dominant entities following soil exposure, facilitating the build-up of nutrients and organic carbon and the subsequent colonisation by heterotrophs and higher plants (Hodkinson *et al.,* 2002). However, more recently, studies have indicated the presence of heterotrophic communities, prior to the establishment of autotrophs (Tscherko *et al.,* 2003; Bardgett *et al.,* 2007). These heterotrophs can deplete the ancient autochthonous organic matter and nutrients that are already present in the exposed soils or use allochthonous nutrients received by aeolian deposition (Kastovska *et al.,* 2005; Bardgett *et al.,* 2007). Following the depletion of these stocks, photoautotrophic and diazotrohic organisms gain a competitive advantage and therefore become prevalent (Chapin *et al.,* 1994; Sattin *et al.,* 2009).

It has been shown that microbial communities in forefields become more abundant, active and diverse with succession, due to the enhanced availability of organic matter, nutrients, and moisture (Kastovska *et al.,* 2005; Nicol *et al.,* 2005; Mindl *et al.,* 2007; Schütte *et al.,* 2010). However, some studies do not show this trend, which may be attributed to an increase in competition between microbes (Sigler and Zeyer, 2002; Zumsteg *et al.,* 2013). Distinct shifts in community composition and functional diversity have been identified, indicating that the environmental factors present at each stage of succession are effective in selecting the dominant microbial communities (Tscherko *et al.,* 2003; Nemergut *et al.,* 2007; Frey *et al.,* 2013). These environmental factors include nutrient and organic matter availability, disturbances, water flow pathways, soil pH and minerology (Sakata Bekku *et al.,* 2004; Frey *et al.,* 2013). Consequently, whilst these factors drive changes along the chronosequence within a forefield, they also stimulate differences between forefields (Sigler and Zeyer, 2002; Liu *et al.,* 2012). Research by Schütte *et al.,* (2010) identified significant differences between bacterial communities in two adjacent glaciers, which share bedrock and dominant climate. These differences were attributed to UV exposure and moisture availability, largely in relation to topography, alongside the presence of vegetation (Schütte *et al.,* 2010). This is supported by Meola *et al.,* (2014), who show that environmental conditions are key drivers between successional differences, however stresses the importance of soil mineralogical properties in mediating bacterial community composition (Meola *et al.,* 2014).

More recently, microbial succession in glacial forefields has been represented by biogeochemical models, with the aim of delineating the complex nature of microbial dynamics and their role in soil development (Bradley *et al.,* 2014). Bradley *et al.,* (2015) presents the mathematical model SHIMMER, with the aim of simulating the initial microbial colonisation, prior to plant establishment. Whilst there are key uncertainties, such as the degree of aeolian deposition, organic matter bioavailability and microbial activity constants, models such as SHIMMER provide new insights, supporting field studies on microbial succession (Bradley *et al.,* 2015). Furthermore, incorporating biogeochemical models will become increasingly important for understanding modifications with global climate change (Bradley *et al.,* 2014).

The above discussion has highlighted the importance of environmental variables in controlling the rate, abundance and diversity of microbial succession (Göransson *et al.,* 2011). These factors include the soil pH, oxygen, salinity, temperature, soil moisture, organic matter and nutrients, disturbances and competition (Hodkinson *et al.,* 2003; Mindl *et al.,* 2007). In particular, studies have noted nitrogen limitation as prevalent in newly exposed glacial soils (Kastovska *et al.,* 2005; Mindl *et al.,* 2007). Whilst initial soils are likely to be oligotrophic, nutrient concentrations following microbial succession are likely to be a result of the in-situ biotic and abiotic nutrient cycling, alongside the prevalence of aeolian deposition (Kastovska *et al.,* 2005; Schmidt *et al.,* 2008). Consequently, the following section will discuss nitrogen dynamics and the role of soil microbial communities during primary succession.

### 1.3.4 Microbial nitrogen cycling in glacial forefields

Nitrogen is a fundamental nutrient in forefield soils, required by plants and microbes for protein synthesis (Treseder, 2008). Bioavailable nitrogen, which can be readily assimilated, consists of organic nitrogen (ON), ammonia, nitrite ($NO_2^-$) and nitrate ($NO_3^-$) (Barber, 1995; Bremner, 1965). The main sources of bioavailable nitrogen to forefield soils are from bedrock weathering, nitrogen-fixing microorganisms (diazotrophs), allochthonous deposition, degradation of organic material and in washing of snowmelt (Bradley *et al.,* 2015). Typical nitrogen concentrations range between 0.2 – 2 mg $g^{-1}$ in newly exposed soils, with studies observing a general increase with soil development (Strauss *et al.,* 2009; Bradley *et al.,* 2014).

Nitrogen-fixing microorganisms are often shown to be prevalent in the initial pioneer communities (Duc *et al.,* 2009; Strauss *et al.,* 2009). If nitrogen levels in newly exposed soils are limited, nitrogen fixers obtain a competitive advantage, and readily colonise (Ohtonen *et al.,* 1999; Strauss *et al.,* 2009). This nitrogen fixation builds the bioavailable soil nitrogen stocks and facilitates the succession of higher microorganisms and vascular plants (Bradley *et al.,* 2015). Nemergut *et al.,* (2007) exemplifies this, utilising a chronosequence approach in three un-vegetated Peruvian successional soils. Low initial nitrogen concentrations were developed by nitrogen-fixing microbes, which increase the habitability of the soils for plant colonisation (Nemergut *et al.,* 2007). Studies have identified a plenitude of nitrogen fixation, mineralisation and denitrification associated genes in glacial forefields (Deiglamayer *et al.,* 2006; Duc *et al.,* 2009). Duc *et al.,* (2009) built upon this, and identified active nitrogen fixation was possible in forefield soils, through using assays sourced from Damma glacier soils (Duc *et al.,* 2009). This therefore indicates the importance of free-living diazotrophs to nitrogen cycling, prior to the establishment of plants (Duc *et al.,* 2009).

However, additional sources, alongside nitrogen fixation have been identified as significant contributors to forefield nitrogen stocks (Bradley *et al.,* 2014). Brankatschk *et al.,* (2011) suggests that allochthonous deposition and remineralisation of overridden organic matter contributes significantly more than microbial communities to soil nitrogen stocks, supported by the limited number of nifH genes (encoding nitrogen fixation), found in Damma Glacier soils (Brankatschk *et al.,* 2011). A significant increase in soil nitrogen may also be observed following the colonisation of soils by vascular plants, attributed to nitrogen from symbiotic root associated nitrogen fixers, and degraded plant litter (Bradley *et al.,* 2014).

Alongside nitrogen fixation, other components of the microbial nitrogen cycle have been identified in forefield soils (Brankatschk *et al.,* 2011). Kandeler *et al.,* (2006) identified an increase in denitrification-associated gene copies with soil development, which may be related to anoxic conditions from high moisture, plant dominated soils (Schluz *et al.,* 2011). Furthermore, Brankatschk *et al.,* (2011) identified nitrogen mineralisation, from decomposing organic matter as the most significant component of the nitrogen cycle in forefield soils. The findings also indicated low nitrogen fixation, nitrification and denitrification rates initially, however increased with soil age, related to increases in bioavailable nitrogen stocks (Brankatschk *et al.,* 2011). Importantly, gene copy numbers were high throughout, and even during periods of low activity, highlighting the mismatch between gene abundance and activity (Brankatschk *et al.,* 2011).

Consequently, nitrogen dynamics within forefield soils are complex, and likely to vary both within and between locations, largely in relation to the prevalent environmental characteristics (Deiglamayer *et al.,* 2006).  For example, in areas which nitrogen is restricted in the bedrock mineralogy, such as Damma glacier, Switzerland, the soils rely on nitrogen fixation and aeolian deposition as the primary sources (Schulz *et al.,* 2013). However, in areas whereby nitrogen is readily sourced in initial soils from bedrock weathering, may depend less on microbial and allochthonous sources (Bradley *et al.,* 2014).

### 1.3.5 Carbon, phosphorous and sulfur in glacial forefields

Nutrient and organic matter inputs to glacial forefields can come from allochthonous sources, such as supraglacial or subglacial runoff, bird and mammal droppings and aeolian deposition (Zumsteg *et al.,* 2012; Bradley *et al.,* 2014). Additionally, in situ microbial activity can provide autochthonous sources of carbon, nitrogen, phosphate and sulfur (Hahn and Quideau, 2013).

Carbon is fundamental to microbial and plant growth, providing the backbone for biological molecules. A range of carbon compounds can be utilized by microbes, such as $CO_2$ which is fixed by autotrophic bacteria and plants, organic carbon molecules which are bioavailable for heterotrophic organisms, wind-blown hydrocarbons and ancient carbon from overridden soils (Hodkinson e*t al.,* 2002; Guelland *et al.,* 2013). The total organic carbon (TOC) content of forefield soils typically ranges between 0.1 – 40 mg $g^{-1}$ and increases with soil succession (Guelland *et al*., 2013; Bradley *et al.,* 2014). Consequently, TOC content in forefields has been shown to increase in relation to soil age, attributed to the development of microbial communities and vegetation (Conen *et al.,* 2007). Autotrophic microbes have been proposed as key facilitators for the build-up of organic carbon in forefield soils through $CO_2$ fixation, particularly in newly exposed soils (Strauss *et al.,* 2012). However, TOC can be supplied by aeolian deposition of materials such as soot or may already be present in soils in the form of ancient organic carbon, exposed by glacier retreat (Stibal *et al.,* 2008; Guelland *et al.,* 2013).

Phosphorous is a key limiting nutrient for microbes and plants, used in the synthesis of nucleic acids and adenosine triphosphate (ATP). Phosphorous is commonly sourced from weathering of underlying bedrock in forefields and therefore soil minerology can be a key control on microbial growth potential in initial soils (Egli *et al.,* 2012; Bradley *et al.,* 2014). Similarly, to TOC, phosphorous content in forefield soils has been shown to increase with succession, ranging between 2 – 8 ug $g^{-1}$ (Bradley *et al.,* 2014). Furthermore, whilst not a required nutrient for all microbial growth, sulfur is cycled in soils through microbial redox reactions (Koltz *et al.,* 2011). Sulfur oxidation is used by microbes for energy and is coupled to the reduction of oxygen or nitrate, in aerobic or anaerobic conditions, respectively (Wainwright, 1978). The process oxidises elemental sulfur ($S^0$) or reduced sulfur such as sulfide ($H_2S$), to sulfate ($SO_4^{2-}$) (Wainwright, 1978). Conversely, $SO_4^{2-}$ can be used as a terminal electron acceptor for microbial oxidation of organic matter, in the absence of oxygen, forming reduced $H_2S$ (Widdel and Hansen, 1992). Evidence for microbial sulfide oxidation has been found across glacier forefields (Szynkiewicz *et al.,* 2013; Wolicka *et al.,* 2014). Bedrock minerology and subsequent weathering is a key source of sulfide to the soil microbial sulfur cycle (Synkiewicz *et al.,* 2013).

## 1.4 Proglacial fjords and microbial communities

### 1.4.1 Fjord systems

Proglacial fjords are at the interface between glacial meltwater and the open ocean. Glaciers supply freshwater, sediment and nutrients to saline nearshore waters (Davila *et al.,* 2002; Lawson *et al.,* 2014; Gutiérrez *et al.,* 2015; Hawkings *et al.,* 2015; Figure 1.1). This freshwater flux often stratifies the water column, creating a surface freshwater lens on top of dense saline waters (Prado-Fiedler, 2009). Fjord systems draining glacier ice fields, such as those in Southern Chile, are hotspots of primary productivity (Iriarte *et al.,* 2007). These regions have been shown to have high rates of primary productivity, diverse ecosystem structures and to harbour unique surface and benthic communities (Iriarte *et al.,* 2007). Fjords may be subject to gradients of meltwater influence, with reduced salinity, increased nutrients and suspended sediment close to the glacier terminus (Davila *et al.,* 2002; Cowton *et al.,* 2012; Lee *et al.,* 2013; Wadham *et al.,* 2013). The biological implications of these gradients have yet to be fully understood. Bacterial community composition and function are likely to modify along these gradients, in line with salinity, light and nutrient content (Gutiérrez *et al.,* 2015). Changes to bacterial communities may have impacts on local biogeochemical cycles, and the wider food chain (Meerhoff *et al.,* 2013).

Understanding the influence of glaciers on microbial communities is key in the context of global climate change. Glaciers, such as those in the Patagonian ice fields, are retreating and therefore proglacial fjords and in-situ microbial communities are likely to be influenced by short term increases in ice melt (Iriarte *et al.,* 2014). By investigating and understanding microbial communities in current fjord systems, the future impacts of rising freshwater fluxes can be better understood.

### 1.4.2 The influence of glacial meltwater

When glacial meltwater drains through fjord systems, both vertical and horizontal environmental gradients in nutrients, light, salinity and sediment can be created (Iriarte *et al.,* 2014). This is due to the mixing of saline ocean water, and fresh glacial meltwater, and subsequent stratification of the water column (Iriarte *et al.,* 2014). In Chilean fjords, high nutrient, saline, Sub-Antarctic water flows at depth, below a freshwater layer of glacial meltwater (Aracena *et al.,* 2011). Mixing between the layers is crucial to distribute nutrients from deep water masses (Aracena *et al.,* 2011). Environmental oxygen gradients may occur due to the stratified water column, and deep hypoxic zones may be created due to the lack of replenishment of oxygen from surface waters (Silva and Vargas, 2014). Distinct gradients in nutrient content are probable due to the stratification, with surface waters likely to be low in nutrient content due to lack of mixing with deeper high nutrient bottom waters (González *et al.,* 2013). This has been exemplified in the saline Sub-Antarctic bottom waters in Chilean fjords, which are high in both nitrogen and phosphate compared to surface waters (González *et al.,* 2013). Consequently, glacial runoff may be a key source of nutrients, such as iron, silicon, phosphate and nitrogen to nutrient limited surface waters in Chilean fjords (Hood *et al.,* 2009; Wadham *et al.,* 2013). Glacial meltwaters have been shown to supply significant levels of silicon to surface waters, however may be low in nitrate (González *et al.,* 2013). High nutrient bottom waters may therefore also be crucial in stimulating primary productivity through the supply of nitrate (and other limiting nutrients) to surface waters (Iriarte *et al.,* 2014). This has been exemplified by spring blooms stimulated by vertical mixing of the water column (González *et al.,* 2013). Distinct gradients in light attenuation may be created vertically and horizontally in glacially-fed fjords, due to the outflow of sediment in glacial runoff (Keck *et al.,* 2000). Glacier runoff has been shown to carry high sediment loads from glacial weathering, which are exported to downstream ecosystems (Hawkings *et al.,* 2015). This sediment load may restrict light levels in surface waters close to the glacier outflow, however this effect is likely to reduce with distance from the runoff source (Keck *et al.,* 2000; Hawkings *et al.,* 2015).

### 1.4.3 Microbial communities and fjord systems

The vertical and horizontal environmental gradients in nutrients, salinity, light and sediment created by glacial outflow may influence microbial community structure and biogeochemical function (González *et al.,* 2013; Gutiérrez *et al.,* 2015). A stratified water column may encourage distinct regions of microbial growth depending on salinity tolerance, with organisms associated with freshwater environments more likely to be isolated in surface waters (Gutiérrez *et al.,* 2015). Research by Gutiérrez *et al.,* (2015) using 16s rRNA found distinctive seasonal meltwater influences on fjord microbial community structure from the Jorge Montt glacier (SW Chile). In Autumn, the high availability of meltwater was shown to encourage the presence of freshwater-tolerant organisms, alongside a greater bacterial richness in surface waters (Gutiérrez *et al.,* 2015). Additionally, research by Dethier and Schoch (2005) found that changes in salinity to benthic populations reduced species richness. Nutrient and temperature gradients have also been shown to be key controls on the composition of fjord communities under the influence of glacial meltwater (Renner *et al.,* 2012). The outflow of particulate matter in glacial runoff may have a significant influence on microbial community composition, function and primary productivity (Iriarte *et al.,* 2014). Although particulate matter may supply additional nutrients to surface waters (Hawkings *et al.,* 2015), the sediment may reduce light for photosynthetic organisms (González *et al.,* 2013). The reduction of light attenuation, alongside the increase in turbidity caused by glacial outflow, has been shown to limit primary productivity, reduce microbial abundance and modify community composition in glacially fed fjords (Keck *et al.,* 2000; Iriarte *et al.,* 2014). Research by Aracena *et al.,* (2011) on the primary productivity of Chilean fjords showed that glacially fed waters had the lowest levels due to increased turbidity and reduced light levels. Overall, glacier outflow may have a significant influence on microbial community structure and function, through the development of environmental gradients (Iriarte *et al.,* 2014). The impacts of this may be identified at the scale of the microbial community structure but also on the wider ecosystem function, such as the distribution and growth of fish (Landaeta *et al.,* 2012; Gutiérrez *et al.,* 2015).

The microbial community of fjord benthic sediments may also be influenced through horizontal gradients in salinity, nutrients and temperature, created by glacial outflow (Keck *et al.,* 2000). In particular, organic matter has been shown to increase from oceanic sediments to inner fjords, in relation to discharge from terrestrially fed rivers (Aracena *et al.,* 2011). However, glacially fed fjords are likely to have reduced organic matter contents, largely due to the high inorganic sediment fluxes in rivers created from glacial weathering (Silva *et al.,* 2008; Aracena *et al.,* 2011). The limited organic matter content of glacier fed fjord sediments may also relate to the reduced primary productivity in surface waters from turbidity and limited light levels (Aracena *et al.,* 2011). Consequently, glacier meltwaters may influence the structure, function and productivity of microbial communities in both the water column and the sediments of fjords.

### 1.5 Proglacial environments and climate change

The structure and function of proglacial microbial communities may be substantially modified with global climate change (Yde *et al.,* 2011). Enhanced warming of polar regions will act to accelerate glacier melt rates and extent, promoting deglaciation (IPCC, 2013). Continued ice melt will expose undeveloped soils in proglacial forefields and drive increased meltwater fluxes into fjords (IPCC, 2013).

It is important to understand forefield microbial communities and their succession, alongside how this may change with global climate change (Schulz *et al.,* 2013; Bradley *et al.,* 2014). Climate warming may modify soil successional pathways, the dominant microbial and plant communities, the biogeochemical cycling in these environments and the composition and fertility of developed soils (Davidson and Janssens, 2006; Schulz *et al.,* 2013). Fjords draining glacial meltwater will experience enhanced freshwater fluxes with ice melt (Davila *et al.,* 2002;

IPCC 2013). Changes to the salinity balance may modify the dominant microbial community structure and function, enhance water column stratification and subsequently modify local biogeochemical cycles (Meerhoff *et al*., 2013; Gutiérrez *et al.,* 2015). Additionally, enhanced meltwater fluxes may increase sediment inputs to fjords and modify nutrient content (Iriarte *et al.,* 2010; González *et al.,* 2013). Not only does this have implications on the in-situ microbial communities, but also on nutrient dynamics and wider ecosystem functioning (Gutiérrez *et al.,* 2015). This is particularly significant in regions such as South West Chile, due to the commercial importance of Salmon fisheries, which rely on the supporting ecosystem to maintain productivity (Iriarte *et al.,* 2010). By Investigating the current microbial community structure function and diversity, we can better understand how these may change in future years.

## 1.6 Methodological considerations for investigating microbial communities

### 1.6.1 Methods available

A variety of methods are available to investigate microbial community composition and functional diversity. These techniques range from traditional microbial culture-based studies, to more recent molecular methods, such as metagenomics (Ward *et al.,* 1990; Janssen *et al.,* 2002; Riesenfeld, 2004; Daniel, 2005; Teeling and Glockner, 2012). Whilst single organisms can be investigated through culture-based methods, developments in DNA sequencing have enabled whole environmental samples to be evaluated at once, using 16s amplicon sequencing and more recently, metagenomics (Tringe *et al.,* 2005; Fierer *et al.,* 2012). These methods are particularly interesting as they enable researchers to isolate the taxonomic diversity (16s), or both the taxonomic and functional potential of microbial communities (metagenomics) (Handelsman, 2004; Daniel, 2005; Fierer *et al.,* 2012). This provides a more comprehensive insight into natural microbial communities, in comparison to growing a single organism independently (Torsvik and Øvreås, 2002). The available techniques to investigate microbial communities are discussed in this section, in the context of scientific developments and recent advances.

### 1.6.2 Culture-based studies

Microbial cultures have been extensively used in soil research to investigate single organisms, isolated from environmental samples (Teeling and Glockner, 2012). Cultures are used to identify new organisms, optimum growth conditions, the ability of organisms to grow under extreme stressors (such as darkness and low temperatures), metabolic functions and more recently, used in DNA sequencing (Kirk *et al.,* 2004). For example, Frey *et al.,* (2013) carried out microbial cultures of bacteria isolated from the Damma glacier forefield to understand their role in granite weathering and therefore soil formation. However, less than 1% of organisms can be successfully cultured in laboratory conditions (Riesenfeld, 2004; Teeling and Glockner, 2012). This was first identified as the 'great plate count anomaly' whereby laboratory cultures are unable to grow population sizes equal to those observed in natural samples (Handelsman, 2004). Consequently, the majority of organisms and their metabolisms, remain unidentifiable through culture-based techniques (Reed *et al.,* 2014; Riesenfeld, 2004). This has led to the emergence of molecular techniques, to investigate microbial genetic and functional diversity, without the need to culture (Streit and Schmitz, 2004; Thomas *et al.,* 2012).

### 1.6.3 First generation DNA sequencing

The emergence and subsequent developments of DNA sequencing technologies have revolutionised our understanding of microbial communities (Rondon *et al.,* 2000; Torsvik and Øvreås, 2002). Understanding the DNA sequence of a single organism (genomics), or multiple organisms (metagenomics) enables microbial taxonomy, phylogenetics and functional diversity to be inferred (Riesenfeld *et al.,* 2004; Tringe *et al.,* 2005; Tringe and Rubin, 2005). This is particularly significant for uncultured organisms, as sequencing allows an examination of the genetic structure, metabolic pathways and community diversity, without the need for laboratory growth experiments (Handelsman, 2004; Daniel, 2005; Xu, 2006).

First generation DNA sequencing was based on the chain termination (or Sanger) method (Sanger *et al.,* 1977; Swerdlow *et al.,* 1990; Wooley *et al.,* 2010). The isolated DNA or genes of interest were first amplified prior to sequencing, either by cloning plasmid vectors, or through artificial replication in using polymerase chain reaction (PCR) (Erlich, 1989; Weisburg *et al.,* 1991; Newton *et al.,* 1997). In PCR, the DNA strands are denatured, a primer annealed and subsequently extended by DNA polymerase, to amplify the number of fragments (Shendure and Ji, 2008; Wooley *et al.,* 2010). Di-deoxynucleoside triphosphates (ddNTPs) are incorporated, which act to halt DNA chain extension (Fierer *et al.,* 2005). The DNA sequence is identified through electrophoresis, based on the fluorescent tag of the terminating ddNTPs (Shendure and Ji, 2008). Sequencing is repeated to ensure all areas of the gene (or genome) of interest are covered, however the read lengths returned by Sanger sequencing (up to 1,000 base pairs) may still leave unresolved long repeat sequences (homopolymer regions) (Wooley *et al.,* 2010).

### 1.6.4 Taxonomic marker genes

PCR-based amplification is now fundamental in DNA sequencing and has given rise to improved taxonomic assignment of environmental samples (Riesenfeld *et al.,* 2004; Blazewicz *et al.,* 2013). Universal genes, such as 16s rRNA in bacteria, can be used as taxonomic marker genes, whereby gene specific primers are used in PCR, and the target regions are sequenced, rather than full genomes (Nübel *et al.,* 1997; Handelsman, 2004). Sequences are subsequently compared to databases, and operational taxonomic units (OTUs) defined based on rRNA sequence similarity (Streit and Schmitz, 2004). The genetic structure and function of organisms in environmental samples is therefore inferred from the nearest sequenced neighbour, identified on the database (Riesenfeld *et al.,* 2004). Additionally, non-universal, function specific, target genes can be used, such as nifH for nitrogen fixation, to understand the taxonomic diversity of a microbial population involved in a certain metabolic pathway (Gaby and Buckley, 2012). Molecular markers have become a popular mechanism for understanding taxonomic diversity, however do not isolate overall community microbial function or abundance, as the remaining (un-targeted) functional genes are not sequenced (Thomas *et al.,* 2012).

### 1.6.5 Next generation DNA sequencing

Advances in sequencing technologies have led to the emergence of next generation (NGS) (or massively parallel) sequencing (Mardis, 2008). These technologies vastly improve on the speed, depth and cost of traditional Sanger sequencing methods (Metzker, 2010). This is due to the high throughput of NGS machines and the ability to sequence millions of DNA fragments at one time (Mardis, 2008; Shendure and Ji, 2008). The popular NGS platforms include Illumina, PacBio and Ion Torrent (Metzker, 2010; Liu *et al*., 2012). The use of these platforms has risen with developments in bioinformatics, such as increased data storage capacities and techniques for large dataset analysis (Mardis, 2008; Horner *et al.,* 2009). Each platform provides a unique method for DNA sequencing, however they are all based on a 'cyclic array' approach (Shendure and Ji, 2008). This is where DNA is cycled through repeated steps of enzyme-initiated manipulation and DNA bases are read by imaging (Shendure and Ji, 2008). NGS techniques can be applied to both single genomes and metagenomes, to sequence specific genes, or the whole genomes/communities (Quaiser *et al.,* 2002; Tringe and Rubin, 2005; Delmont *et al.,* 2011). All NGS sequencing platforms can produce mate-pair or paired-end reads, whereby each DNA strand is sequenced in two directions, thereby providing the distance between each paired read (Shendure and Ji, 2008). This helps resolve structural rearrangements, such as INDELS (insertions or deletions), and repeat regions, especially when a fully sequenced reference genome is not available (Hajirasouliha *et al*., 2010; Metzker, 2010; Miller *et al.,* 2010). NGS platforms produce varying read lengths, however the longer the read length provided, the higher the sequencing error rate (Teeling and Glockner, 2012). However, longer reads are beneficial for resolving large repeat regions, enhancing taxonomic and functional annotations and revealing structural rearrangements (Wommack *et al.,* 2008). Illumina sequencing is currently widely used due to reasonable read lengths (150 – 300bp) and minor error rate (Illumina, 2018).

However, NGS sequencing technologies still retain several limitations (Mardis, 2008; Ansorge, 2009; Alkan *et al.,* 2011; Teeling and Glockner, 2012). The cost of purchasing the sequencing platforms may hinder their use in some labs, alongside accounting for the cost of each round of DNA sequencing (Ansorge, 2009; Grada and Weinbrecht, 2013). Furthermore, as with Sanger sequencing, homopolymer regions may not be accurately resolved, due to spanning longer lengths than the short reads returned through sequencing (Alkan *et al.,* 2011; Grada and Weinbrecht, 2013). Additionally, base call errors generally increase towards the 3' end of read fragments, and therefore the error rate is not consistent throughout the sequencing run (Miller *et al.,* 2010). Furthermore, the datasets returned by NGS place high computational demands on downstream analysis, often requiring complex bioinformatic pipelines to assemble, annotate, and investigate the resulting DNA sequences (Horner *et al.,* 2009; Grada and Weinbrecht, 2013).

Further technological advances from NGS platforms have enabled the development of third generation sequencing platforms (Schadt *et al.,* 2010). Platforms have been provided by Pacific Biosciences (SMRT), Oxford Nanopore, Life Technologies (FRET) and Ion Torrent (Branton *et al.,* 2008; Wash and Image, 2008; Ozsolak, 2012; Quail *et al.,* 2012; Roberts *et al.,* 2013). Whilst these technologies are still being established, they aim to improve the time and cost efficiency of DNA sequencing (Wash and Image, 2008; Schadt *et al.,* 2010). SMRT (Single Molecule sequencing in Real Time), produced by PacBio, is based on sequencing individual fragments of DNA, and can provide read lengths of up to 15kb, however error rates are still high (Roberts *et al.,* 2013).

### 1.6.6 Metagenomics

Metagenomics is a more recent approach to DNA analysis and involves sequencing the complete DNA of a microbial community without PCR amplification (Handelsman, 2004). As multiple genes are sequenced, this approach allows both taxonomic and functional annotation of the microbial community (Tringe *et al.,* 2005). Consequently, metagenomics provides an alternative to 16s amplicon sequencing, as the presence of functional genes allows an insight into the community's metabolic potential, and thus, biogeochemical importance (Handelsman, 2004; Daniel, 2005). Additionally, because this approach does not require DNA amplification, it is not susceptible to the bias typically introduced through PCR (Risenfeld *et al.,* 2004). However, much greater quantities of DNA are required prior to sequencing, which restricts the analysis of low biomass samples (Yilmaz *et al*., 2010). As metagenomics supplies substantially more data than amplicon sequencing and requires greater downstream processing, the post-sequencing analysis process is more demanding (Schmieder and Edwards, 2011). The relatively recent introduction of this technique means there is a lack of formalised guidance or software for the analysis of metagenomic data, hindering its uptake by the scientific community.

### Metagenome sequence assembly and annotation

NGS metagenome sequencing supplies short DNA fragments, which depending on the sequencing platform, can range from 100bp (SOLiD) to 1000bp (Sanger) in length (Miller *et al.,* 2010). These raw reads can be directly interpreted using analysis packages such as MG-RAST, for taxonomic and functional annotation (Glass *et al.,* 2010). As these annotations are often made by aligning the DNA fragments to genes in databases, such as NCBI GenBank, the short length of the fragments limits the quality and magnitude of the possible alignments (Howe *et al.,* 2014; Vázquez-Castellanos *et al*., 2014). Consequently, to provide better sequence annotations, the read fragments can be assembled into longer contigs using a sequence assembler (Nagarjan and Pop, 2013). This is particularly important for accurately understanding the metabolic function or phylogenetic diversity of the sequenced organisms (Pignatelli and Moya, 2011). This is because annotating genes with functional and taxonomic identity involves alignment to databases, and sequences which are longer in length will provide more accurate alignments (Van der Walt *et al.,* 2017).

### Sequence assembly

Sequence assembly aims to reconstruct the gene, genome or metagenome that was sequenced (Miller *et al.,* 2010). This involves identifying overlaps in the raw DNA read fragments, using an assembly algorithm, generating longer fragments (contigs) (Narzisi and Mishra, 2011). This is particularly important when using NGS platforms which supply short read fragments that are difficult to directly interpret (Mende *et al.,* 2012). The length of overlaps between raw reads are defined by the user, based on a specified k-mer value (a sequence of k- base calls) (Narzisi and Misha, 2011). The contigs generated can be built into a scaffold, which define the orientation and order that the contigs occurred in the original genome (Myers *et al.,* 2000; Peng *et al.,* 2012). Sequence assemblers assume that similar fragments occur at proximity within a genome, however this may be fundamentally undermined by the presence of repeats that occur throughout the genome, or between genomes (Nagarjan and Pop, 2013). These repeats are less problematic when they fall within the read length, however become harder to resolve when they span longer than the sequencing read length (Narzisi and Misha, 2011). Repeat sequences, short reads and sequencing errors are the key issues that face assembly algorithms (Narzisi and Misha, 2011).

There are two dominant approaches for NGS sequence assembly; reference guided and de novo (Miller *et al.,* 2010; Zhang *et al.,* 2011). Reference guided assembly is used for single genomes, when a 'gold standard' assembled genome is available to assist the assembly of the raw sequence data (Cattonaro *et al.,* 2010). Typically, the raw sequencing reads are aligned against the reference genome, to help guide the assembly (Vezzi *et al.,* 2011). De novo sequence assembly is used in the absence of a reference genome and can be used for both single and metagenome datasets (Zhang *et al.,* 2011). De novo assemblers aim to merge the short sequencing reads based on common overlapping fragments of a specified length (k-mers) (Miller *et al.,* 2010). These assemblers therefore aim to simplify the raw dataset and reproduce the metagenome in the environmental sample (Nagarjan and Pop, 2013). There are three key approaches to de novo assembly, which include; Greedy algorithms, Overlap Layout Consensus algorithms (OLC) and de Bruijn graph algorithms (Miller *et al.,* 2010).

Greedy sequence assembly algorithms work by selecting the highest scoring read overlaps, before merging raw reads into longer contigs (Miller *et al.,* 2010; Zhang *et al.,* 2000). This is carried out until all the available merges have been exhausted (Narzi and Misha, 2011). Heuristic corrections are carried out on each merge, with assembly gaps left in regions where corrections are not possible (Zhang *et al.,* 2000). Greedy algorithms consequently provide a set of assembled contigs, for a single genome or metagenome (Pop, 2009). Examples of assemblers based on this algorithm include PHRAP, CAP3, PCAP, TIGR, SSAKE and Phusion (Sutton *et al.,* 1995; Huang and Madan, 1999; Mullikin and Ning, 2003; Bastide and McCombie, 2007; Warren *et al.,* 2007; Simpson *et al.,* 2009; Pignatelli and Moya, 2011).

OLC algorithms are optimally used on raw reads exceeding 200bp in length (Zhang *et al.*, 2011; Deng *et al.,* 2015). This algorithm works by graphically representing the sequencing reads and the base pair overlaps between them (Miller *et al.,* 2010; Nagarjan and Pop, 2013). The overlaps are generated through pairwise comparisons of the raw reads, by in built aligners (Narzi and Misha, 2011). Each graph node comprises a metagenome read, with the edges connecting nodes representing the overlaps between reads (Miller *et al.,* 2010). Contigs are graphically shown by the paths running through the graph, connecting nodes (Deng *et al.,* 2015). The aim of OLC algorithms is to simplify the raw assembly graph by connecting each node (raw read) to a single 'Hamiltonian' path (Miller *et al*., 2010). However, the presence of inter and intra genome repeats and sequencing errors result in multiple diverging paths (Miller *et al.,* 2010; Deng *et al.,* 2015). Whilst OLC algorithms are computationally expensive, they can incorporate both forward and reverse DNA strands, alongside distinguishing the 5' and 3' read ends (Miller *et al.,* 2010; Narzi and Misha, 2011). Examples of OLC based sequence assemblers include Edena, MIRA and Celera (Denisov *et al.,* 2008; Hernandez *et al.,* 2008).

Finally, de Bruijn graph-based assemblers are optimal for short reads around 100bp, such as those obtained from Illumina sequencing (Miller *et al.,* 2010). Similar to OLC, the algorithm represents the assembly graphically, through a series of edges and nodes (Nagarjan and Pop, 2013). Sequencing reads are split up into fragments of length k (k-mers) and represented by edges on the graph (Narzi and Misha, 2011). The k-mer length is specified by the user and must be below the read length (Narzi and Misha, 2011). The defined k-mer length is significant, as reads significantly longer than length k will be fragmented, prior to being re-assembled (Miller *et al.,* 2010). Nodes on the graph represent the read overlaps, of length $k^{-1}$ (Nagarjan and Pop, 2013). Consequently, nodes connect two edges if they share a common sequence of $k^{-1}$ length (Nagarjan and Pop, 2013). As with OLC algorithms, de Bruijn graphs aim to simplify the assembly by merging reads into contigs and forming a single sequence (an Eulerian path) (Pevzner *et al.,* 2001; Conway *et al.,* 2012). Repeats and errors prevent the assembly of a single large sequence, so contigs are merged and extended before the paths branch, resulting in numerous Eulerian paths (Pevzner *et al.,* 2001; Narzi and Misha, 2011; Pignatelli and Moya, 2011). Miller *et al.,* (2010) identifies three key types of errors that may be represented in a fragmented de Bruijn graph (Miller *et al.,* 2010). 'Bubbles' are created when the Eulerian path splits and subsequently recombines due to errors in the centre of a read (Miller *et al.,* 2010). 'Spurs' are created when a second Eulerian path is generated, due to an error at the 3' end of a sequencing read (Pevzner *et al.,* 2001; Miller *et al.,* 2010). Finally, 'cycles' form when two paths converge due to repeat sequences, present in multiple reads, which ideally need to be separated (Miller *et al.,* 2010). Whilst these graphical representations are computationally expensive, both forward and reverse DNA reads can be represented as k-mers (Narzi and Misha, 2011). Assemblers that utilise the de Bruijn graph algorithm include Velvet, IDBA, SOAPdenovo, CLC, AllPaths, metaSPAdes and AbySS (Butler *et al.,* 2008; Zerbino and Birney, 2008; Simpson *et al.,* 2009; Li *et al.,* 2010; Miller *et al.,* 2010; Peng *et al.,* 2010; Nurk *et al.,* 2017).

### Sequence annotation

Following DNA sequencing and assembly, contigs can be annotated with their taxonomic and functional identity, using a database (Thomas *et al.,* 2012). Annotation software and pipelines exist, such as MG-RAST, which allow short unassembled reads to be annotated (Wooley *et al.,* 2010; Vincent *et al.,* 2013). However, assembled contigs with longer lengths pose fewer challenging alignments to gene databases and can provide more accurate results, given a high-quality assembly (Thomas *et al.,* 2012). Taxonomic and functional annotation of assembled contigs can be carried out using a range of algorithms and software platforms, such as MEGAN and IMG/M (Huson *et al.,* 2007; Chen *et al.,* 2017). These programs provide homology and sequence similarity alignments to databases such as NCBI GenBank, KEGG and eggNOG, to reveal the taxonomic distribution and functional potential of metagenomes (Kaneisha and Goto *et al.,* 2000; Benson *et al.,* 2005; Huerta-Cepas *et al.,* 2015).

### The choice of assembler

When conducting a genome or metagenome assembly, the user is faced with numerous choices. In the absence of a reference genome, a de novo assembly must be carried out, however the assembly algorithm and specific assembler need to be selected (Zhang *et al.,* 2011). There is currently an absence of formalised guidance on assembler selection and evaluation (Arumugam *et al.,* 2010). The decision-making is therefore left to the user, with the dataset size, complexity and availability of computational power all key considerations (Finotello *et al.,* 2011). Consequently, it is important for the user to evaluate and compare assemblers, to make an informed decision (Kurtz *et al.,* 2004; Peng *et al.,* 2012; Howe *et al.,* 2014).

Greedy algorithms are simple, less computationally expensive, and apply well to single genomes, with minimal repeats (Zhang *et al.,* 2011). On the other hand, the graph-based algorithms, OLC and de Bruijn, are more computationally expensive, however function better for complex genomes or metagenome samples, with repeat structures (Zhang *et al.,* 2011). Furthermore, as all NGS platforms vary in terms of the error rate, read lengths, coverage depth and evenness, different assemblers will be better aligned to one platform than another (Earl *et al.,* 2011). Whilst the quality of the assembly does inherit issues from the sequencing platform (base call errors, short read length), the assembler selected will introduce variances through contig length, miss-joined contigs and single nucleotide polymorphisms (SNPs) (Finotello *et al*., 2011; Salzberg *et al.,* 2012). For example, some assemblers, such as SOAPdenvo prioritise extending the contig lengths, at the expense of accuracy (thereby increasing the number of erroneous contigs) (Salzberg *et al.,* 2012). It is therefore important to compare different assembly algorithms and assemblers for the sequenced dataset in question, however this is seldom done (Finotello *et al.,* 2011; Lin *et al.,* 2011).

As metagenomics has been newly introduced into the field of microbial ecology, it is still not widely utilised. This largely relates to the cost, computational intensity and popularity of other community analysis approaches such as 16s amplicon sequencing (Handelsman, 2004). However, the use of this approach is rising, with more papers incorporating metagenomes to study functional ecology (Eloe-Fadrosh *et al.,* 2016a). Consequently, there is a growing need for guidance on assembler use and choice for environmental data analysis. Additional guidance may aid the incorporation of this technique into the research community, highlighting the benefit and need for metagenome assembly and evaluation.

### Selecting an appropriate assembler

To choose an appropriate assembler, several platforms can be compared using the same dataset, and evaluated using selected metrics. However, a range of evaluation metrics exist, none of which provide a complete or standardised evaluation of assembly quality (Arumugam *et al.,* 2010). Traditionally, studies have looked at measures of the assembly size, including the number of contigs produced, the mean contig lengths and the contig N50 length (Earl *et al.,* 2011). However, a long contig length does not fully indicate the assembly quality, as long contigs may be a result of misjoins, and the N50 value cannot be directly compared across assemblies (Finotello *et al.,* 2011). Consequently, an additional measure can look at the accuracy of the assembly, by mapping the assembled contigs to a reference genome if available, identifying any mismatched regions (Miller *et al.,* 2010; Narzisi and Misha, 2011). Quantification of the metagenome coverage may also be utilised, which involves mapping the raw sequence reads to the assembled contigs and identifying the number of which can be correctly aligned (Earl *et al.,* 2011; Huson *et al.,* 2001). The optimal assembly will utilise all raw sequencing reads supplied, and any errors will be random (Kumar and Blaxter, 2010; Bräutigam *et al.,* 2011). Typically, as assemblies tend to trade-off contig length and quality, it is important to combine a series of measures that look at the size, coverage and accuracy of the assembly (Nagarjan and Pop, 2013).

Traditional studies may only choose a single assembler, with selection based on previous findings. This may be due to the complex nature of the raw data and the computational intensity required to carry out an assembler comparison (Najaran and Pop, 2013). Evaluation metrics are often presented to validate the choice of assembler, for example by showing N50 values in comparison to those of other studies (Miller *et al.,* 2010). Some studies have carried out comparisons between assembly algorithms to resolve a specified dataset or have used simulated datasets to test assembler quality (Narzisi and Misha, 2011; Lin *et al.,* 2011). However, these have been restricted to single genome assemblies or using contigs from human samples, or combined human and environmental datasets (Huson *et al.,* 2001; Mavromatias *et al*., 2007; Earl *et al*., 2011; Pignatelli and Moya, 2011; Saltzberg *et al.,* 2012). More recently, a study by Vollmers *et al.,* (2017) compared assemblers using real Illumina sequencing data for forest soil and algal biofilms. This study highlighted the importance of comparing assemblers on multiple metrics, however did not investigate parameter settings, all assembly algorithms or use simulated data to test the accuracy of assemblers (Vollmers *et al.,* 2017). Additionally, Sczyrba *et al., (2017)* compared a range of software for metagenomic data analysis, including assemblers, using artificial metagenomes. However, as this study utilised newly sequenced genomes, the analysis was not specific to bacteria, and in particular, those from heterogeneous environmental samples, such as soils. Thus far, there is no assembler comparison available specifically for studies investigating bacterial communities obtained from sediment or soil.

## 1.7 Summary and research gaps

Diverse and unique microbial life can be found in proglacial regions, including forefield soils and proglacial fjords. These microbes can be subject to distinct environmental gradients and biogeochemical perturbations created by ice melt, which may modify community structure and function. In proglacial soils, ice melt and glacier retreat expose undeveloped soils, providing an opportunity to study microbial colonisation and development. These microbial communities are likely to modify in line with soil development and may have key roles in soil biogeochemical cycling, such as the nitrogen cycle. Understanding community composition and development in terms of structure and function will aid our understanding of microbial colonisation in extreme environments, and the role of these organisms in nutrient cycling. In proglacial fjords, microbial communities are subject to environmental perturbations from glacial freshwater fluxes. These fluxes create distinct changes to salinity, sediment, light attenuation and nutrients in fjords, which again may influence the microbial community taxonomic and functional diversity. As meltwater fluxes continue to increase with climate change, meltwater may pose challenges to the wider biogeochemical functioning of fjords. This is particularly important as fjords in regions such as Chilean Patagonia have been shown to be hotspots of primary productivity, and support commercially significant fisheries. Understanding current community composition and function will help us understand how microbial communities may contribute to biogeochemical cycling and how this may change in the future.

Metagenomics is a technique which can be used to profile both the taxonomic and functional diversity of microbial communities. It can therefore be applied to the understanding of microbial communities in proglacial regions. The assembly of metagenome sequencing data can improve downstream functional and taxonomic annotation, and thus, can raise the accuracy of ecological conclusions. However, the lack of guidance and complex nature of DNA assembly means there is a minimal uptake within the field of microbial ecology. Selecting an appropriate assembler for the data type in question is paramount, to obtain an improved outcome and minimise erroneous results. Consequently, additional guidance for the selection of metagenome assemblers is needed, to improve the outcome of assembly-based studies and encourage the use of DNA assembly. This is particularly needed in the field of soil microbial ecology, as limited focus has been made to this research group.

### 1.8 Aims and objectives

**Objective 1:** To compare the performance of five publicly-available metagenome assemblers for soil bacterial communities

Metagenomics involves the study of all genes present in a microbial community sample. The analysis of the community gene pool provides information on both the taxonomic and functional composition. This provides an insight into the potential role of the microbial community in local biogeochemical cycles. Assembling short metagenome DNA sequencing reads is beneficial to obtain longer fragments (contigs) for alignment to functional gene databases. However, there is currently no formailsed guidance for the assembly of complex bacterial communities isolated from soil samples. In particular, the choice of metagenome assemblers for this research community requires further investigation. This study investigated the performance five publicly available metagenome assemblers, spanning the three dominant assembly algorithms (OLC, de Bruijn graph and Greedy). The assemblers were evaluated based on a series of artificial soil metagenomes of varying complexity, with evaluation based on assembly size, completeness and contiguity. Overall, the choice of assembler was shown to influence assembly quality and thus downstream functional or taxonomic annotation. The more complex de Bruijn graph assemblers, metaSPAdes and CLC provided the highest quality assemblies. This is because these assemblers have more tuneable parameter settings to fit the data type in question and are better suited for short reads returned from Illumina sequencing. It is hoped that the results from this work will help guide assembler selection for the soils community, alongside highlighting the importance of informed choice in the use of metagenome assembly tools.

**Objective 2:** To investigate the similarities and differences in taxonomic composition of diazotrophic bacteria in metagenomes sampled from four Arctic glacier forefields

Published in FEMS Microbiology (Nash *et al.,* 2018).

Microbial nitrogen fixation is crucial for building labile nitrogen stocks and facilitating higher plant colonization in oligotrophic glacier forefield soils. Here, the diazotrophic bacterial community structure across four Arctic glacier forefields was investigated using metagenomic analysis. In total, 70 soil metagenomes were used for taxonomic interpretation based on 185 nitrogenase (nif) sequences, extracted from assembled contigs. The low number of recovered genes highlights the need for deeper sequencing in some diverse samples, to uncover the complete microbial populations. A key group of forefield diazotrophs, found throughout the forefields, was identified using a nifH phylogeny, associated with nifH Cluster I and III. Sequences related most closely to groups including Alphaproteobacteria, Betaproteobacteria, Cyanobacteria and Firmicutes. Using multiple nif genes in a Last Common Ancestor analysis revealed a diverse range of diazotrophs across the forefields. Key organisms identified across the forefields included *Nostoc*, *Geobacter*, *Polaromonas* and *Frankia*. Nitrogen fixers which are symbiotic with plants were also identified, through the presence of root associated diazotrophs, which fix nitrogen in return for reduced carbon. Additional nitrogen fixers identified in forefield soils were metabolically diverse, including fermentative and sulfur cycling bacteria, halophiles and anaerobes.

**Objective 3:** To investigate the bulk microbial community composition along a chronosequence of soil succession in the Midtre Lovénbreen forefield, Svalbard, using metagenomics

Arctic glaciers are currently undergoing retreat with global climate change, revealing undeveloped soils at the glacier terminus. These soils can be used to study succession, utilising a chronosequence based approach along a transect of soil age and development. Understanding microbial communities in the Arctic is important for our understanding of microbial function, diversity and importance in harsh oligotrophic, cold conditions. Additionally, microbial succession in Arctic forefields may provide insights into how extra-terrestrial life may colonise on cold planets such as Mars. Whilst some research has been carried out using 16s rRNA sequencing on forefield communities, metagenomics has yet to be applied to understanding succession. This study used metagenomes spanning the forefield of Midtre Lovénbreen glacier, Svalbard, to investigate microbial community development. A combination of metagenome DNA reads, assemblies and genome binning were used to gain an insight into microbial taxonomy and function during soil development. A diverse range of microbes were recovered, including those with carbon, sulfur and nitrogen cycling metabolisms. Cyanobacteria were detected in early soils, attributed to their ability to fix carbon and nitrogen in oligotrophic conditions, alongside the production of protective EPS. The microbial community was shown to modify along the chronosequence, in line with the establishment of labile carbon and nitrogen pools. Overall, this study provided an insight into the diversity and metabolic potential of microbial communities during forefield succession and highlighted the potential of metagenomic analysis for those studying Arctic microbial ecology.

**Objective 4:** To investigate the composition and potential function of microbial communities sourced from benthic metagenomes in a Chilean fjord.

The fjord systems of Chilean Patagonia contain three UNSECO bio-reserves, support commercially important Salmon fisheries and host high rates of primary productivity. These fjords are particularly of interest due to the interaction of glacial meltwater from Patagonian ice fields with marine waters, harboring a range of physico-chemical conditions for microbial life. However, the microbiology of the fjord sediments, crucial for understanding the wider ecosystem functioning, has received limited research attention.

This study applied metagenomics to understand the taxonomic diversity of uncultured benthic sediment metagenomes from a Patagonian fjord. The uncovered taxonomic diversity was used to drive genome binning and functional analysis, providing insights into the novelty, metabolic potential and ecological diversity of these microbial communities. In particular, extremophiles associated with anoxia and oligotrophy were detected. Despite the harsh nature of the environment, organisms with carbon, nitrogen and sulfur metabolisms were detected, indicating a range of potential biogeochemical cycles. Additionally, the results highlight the novelty of the microbial community, with the potential for new genomes within the samples. The findings provide an insight into this unique environment, whist highlighting areas where targeted single cell genome sequencing and culture-based studies may be beneficial.

# Chapter 2: Comparison of publicly available metagenome assemblers for soil bacterial communities

Maisie V. Nash[1]; Alexandre M. Anesio[2]; Gary Barker[3] and Patricia Sánchez-Baracaldo[1]

[1]School of Geographical Sciences, University of Bristol, BS8 1SS, UK
[2] Department of Environmental Science, Aarhus University, PO box 358, Denmark
[3] School of Life Sciences, University of Bristol, BS8 1TQ, UK

## Contributions and acknowledgements

All work presented in this chapter was carried out by MN. Co-authors AA, GB and PSB provided methodological advice and guidance on editing. This work was presented as part of a conference talk during the Polar and Alpine Microbiology Symposium (2017).

### Funding

This work was carried out as part of the NERC GW4+ Doctoral Training Partnership, awarded to MN. Funding support was provided through NERC grants NE/J02399X/1, awarded to AA, and a Royal Society Research Fellowship awarded to PSB.

## 2.1 Introduction

Natural soils incorporate complex and dynamic bacterial communities, which play important roles in local biogeochemical cycles, such as nutrient cycling and carbon fixation (Handelsman *et al*., 1998; Van Der Heijden *et al.,* 2008). Investigating these communities from a genomic perspective aids understanding of their taxonomic composition and functional potential, and how this changes between locations or over environmental gradients (Cong *et al.,* 2015). However, due to the complex structure of soils, small spatial scale of biogeochemical processes and heterogeneity of bacterial life, it is difficult to extract and culture cells for analysis (Riesenfeld *et al.,* 2004; Teeling and Glockner, 2012). Whilst some studies have used 16s rRNA amplicon sequencing and microarrays to investigate soil bacterial communities, these provide limited information on novel uncultured organisms and the functional genes they contain (Fierer *et al.,* 2012; Cong *et al.,* 2015). A more comprehensive understanding of microbial communities may be obtained through metagenomics, as ideally the complete microbial DNA is sequenced (Daniel, 2005; Handelsman, 2004). This provides the user with information on the community composition and functional genes, therefore highlighting the potential contribution of the community to local biogeochemical processes (Tringe *et al*., 2005; Tringe and Rubin, 2005).

There is currently a limited (but growing) uptake of metagenomic sequencing in the microbial ecology community, largely related to the considerable use and lower cost of 16s amplicon sequencing (Risenfeld *et al.,* 2004). However, metagenomics has been explored more extensively in other fields, for example in studying the human gut microbiome (Guill *et al.,* 2006; Qin *et al.,* 2010; Qin *et al*., 2012). The approach is gaining interest from the microbial ecology community, as more research has been published exemplifying the possibilities with metagenomic data from environmental samples (Piganeau and Moreau, 2007; Makelprang *et al.,* 2011; Pearce *et al.,* 2012; Eloe-Fadrosh *et al.,* 2016a). As there is a lack of guidance on metagenome data analysis, the majority of studies use short unassembled sequencing reads, in programs such as MG-RAST (Meyer *et al.,* 2008; Glass *et al.,* 2010). This may not be optimal, as the short DNA fragments may not provide good alignments to taxonomic and functional databases during metagenome annotation, limiting the potential of the data analysis (Van der Walt *et al.,* 2017). Consequently, assembling the DNA fragments into longer contigs, using a metagenome assembler, can improve the dataset annotation quality (Nagarjan and Pop, 2013). Assembled metagenomes have enabled scientists to explore community functional genes in greater detail, such as through comparing environments or even assembling complete genomes from metagenomic datasets (Sharon and Banfield, 2013; Eloe-Fadrosh *et al.,* 2016b). However, it must be acknowledged that assembly may not always be the optimal approach, for example when investigating highly diverse, novel and fragmented metagenomes, which can be harder to assemble.

A limited number of ecological studies have started assembling metagenome DNA, however these studies often select a single assembler without justification for that choice (Najaran and Pop, 2013). This is widely accepted, as there is limited formalised assembler comparisons or guidance specifically for the analysis of bacterial microbial communities sourced from soils. However, as assemblers are based on multiple algorithms, they will work differently depending on the sequencing platform, community structure and complexity of the input dataset (Zhang *et al.,* 2011). Testing several assemblers for the type of dataset in question, selecting the best performing algorithm, can enhance the quality of the metagenome assembly, and therefore annotation (Kurtz *et al.,* 2004; Peng *et al.,* 2012; Howe *et al.,* 2014). More accurate functional annotations will not only increase the reliability of the ecological conclusions drawn, but also enable more detailed analysis of the microbial community, such as extracting single draft genomes from community datasets (Eloe-Fadrosh *et al.,* 2016b). Several studies have attempted assembler comparison for animal, human and mixed environmental samples (Mavromatis *et al.,* 2007; Lin *et al.,* 2011; Zhang *et al.,* 2011; Peng *et al*., 2012; Deng *et al.,* 2015; Sczyrba *et al.,* 2017; Vollmers *et al.,* 2017). However, many of these comparisons have been for single genome assemblies, not metagenomes, or focused on mixed environmental metagenomes from numerous habitats (Earl *et al.,* 2011; Salzberg *et al.,* 2012; van der Walt *et al.,* 2017; Vollmers *et al.,* 2017). The recently completed CAMI inter-comparison aimed to evaluate software for metagenomic data analysis, including assemblers, based on simulated metagenomes (Sczyrba *et al.,* 2017). However, this study used newly sequenced genomes from a range of sources, spanning Bacteria, Archaea and Fungi (Sczyrba *et al.,* 2017). Thus far, there has been no assembler comparison focused on complex soil bacterial communities, comparing all assembly algorithms.

When conducting an assembler comparison, simulating artificial microbial communities to use as the test dataset(s) is often useful (Mavromatis *et al.,* 2007; Sczyrbra *et al.,* 2017; van der Walt *et al.,* 2017). This allows the user to construct the input dataset, and therefore evaluate how well the assemblers do in recreating it (Mavromatis *et al.,* 2007). These simulated datasets can contain DNA sequences of choice from publicly available databases, alongside artificially introduced sequencing errors, to reflect typical metagenome data preparation (Mavromatis *et al.,* 2007). It is important to conduct assembler comparisons using multiple evaluation measures, including both contiguity and completeness (Arumugam *et al.,* 2010; Vollmers *et al.,* 2017). This will highlight any assemblers that are producing long, but erroneous contigs (Kumar and Blaxter, 2010; Bräutigam *et al.,* 2011).

Here, five publicly available metagenome assemblers, spanning the three main assembly algorithms, were compared for four artificially curated soil metagenomes of different complexity. Assembler evaluation was carried out using both the contiguity and completeness of metagenome assemblies. This study aims provide a more standardized method for the assembly of complex soil metagenome data, aiding assembler choice within the soil bacterial ecology community.

## 2.2 Methodology

For this analysis, five metagenome assemblers were selected for comparison, spanning the three key assembly algorithms (Arumugam *et al.,* 2010; Zhang *et al.,* 2011). A summary of these assemblers can be identified in Table 2.1.

***Table 2.1:*** *The five assemblers used in the evaluation, detailing the assembler version and algorithm, alongside assembly specific details.*

| Assembler | Algorithm | Assembly details |
|---|---|---|
| MetaSPAdes 3.7.0<br>(Nurk *et al.,* 2013) | de Bruijn graph | Assembly run in paired end mode, with error correction switched off (--only-assembler flag) as base quality values are required for this, however are not supplied in the MetaSim datasets. Read coverage cut-off was not set as this cannot be used in paired end mode.<br>Key parameter to modify was the kmer length, which ranged between 21 and 71. |
| SSAKE 3.8.4<br>(Warren *et al.,* 2007) | Greedy | Assembly run in paired end mode.<br>Key parameters to modify were the minimum contig coverage depth, the minimum number of overlapping bases required for a consensus sequence join, and to trim bases when all other extension prospects have been trialled. |
| ABYSS 1.9.0<br>(Simpson *et al*., 2009) | de Bruijn graph | Key parameters to modify were the kmer length, the minimum contig coverage and to pop assembly bubbles below a threshold value. |
| MIRA 4.0<br>(Chavreux, B., 2014) | Overlap layout consensus | Assembly run without base quality values, using the –no_qualities flag, as these are not supplied by Metasim. Assemblies were run using the flag for Illumina input data (SOLEXA_SETTINGS).<br>Key parameters to modify were the minimum read length used in the assembly and the minimum contig length. |
| CLC 4.4.1<br>(Qiagen Bioinformatics, 2016) | de Bruijn graph | Assembly run in paired end mode without scaffolding.<br>The key parameters to modify were the minimum output contig length, the minimum word size for the de Bruijn graph, and the maximum assembly bubble size allowed. |

To test these assemblers, an artificial metagenome was created in MetaSim, incorporating 123 complete soil bacterial genomes downloaded from NCBI GenBank, identified in Table 2.2 (Richter *et al.,* 2008). Simulated Illumina sequencing errors were introduced to the metagenome to replicate typical sequencing error profiles, such as declining sequencing quality with run time (Richter *et al.,* 2008). Four variants of the artificial metagenome were created (A-D), with increasing levels of complexity, each containing 10,000000 reads, 80 base pairs (bp) in length (Table 2.3). The metagenome size was selected as a trade-off between accurately representing soil metagenome sizes, whilst minimising computational requirements. Metagenome complexity modifications were carried out by amending the abundance profile of bacterial species in the MetaSim simulations, in line with the methodology of Mavromatis *et al.,* (2007). The low complexity community (Metagenome A) contained a small number of high abundance organisms, with the remainder of the bacterial population at a lower abundance. This theoretically would allow a good assembly of the high abundance bacteria, due to an increase in read coverage for those organisms (Mavromatis *et al.,* 2007). In contrast, the high complexity community (Metagenome D) contains organisms at a range of abundance levels, with varying levels of coverage. As the population abundance is more variable, the assemblers may create a smaller and more erroneous output, for example by increasing the number of chimeric contigs, whereby reads from different organisms are merged (Mende *et al.,* 2012). This is because there would be reduced sequencing support for each read from low abundance organisms, thereby causing difficulties in assembling regions with long repeats or INDELS. Furthermore, a default dataset with even coverage (or abundance) across all bacteria (Metagenome A) was also included. This is to remove the effect of high abundance organisms, which, due to the associated increase in read coverage, tend to pull the average assembly quality up (Mavromatis *et al*., 2007). In total 100 assemblies were carried out, with the full list available in Appendix 1 Table 1.

***Table 2.2:*** *Complete bacterial genome sequences used for artificial metagenomes created in MetaSim, sourced from NCBI GenBank. GenBank ID numbers are given, alongside genome description.*

| No. | NCBI accession number | NCBI sequence name |
|---|---|---|
| 1 | CP007128.1 | *Gemmatimonadetes* bacterium KBS708, complete genome |
| 2 | NC_000964.3 | *Bacillus subtilis* subsp. subtilis str. 168 chromosome, complete genome |
| 3 | NC_002516.2 | *Pseudomonas aeruginosa* PAO1 chromosome, complete genome |
| 4 | NC_002942.5 | *Legionella pneumophila* subsp. pneumophila str. Philadelphia 1 |
| 5 | NC_002947.3 | *Pseudomonas putida* KT2440 chromosome, complete genome |
| 6 | NC_003030.1 | *Clostridium acetobutylicum* ATCC 824 chromosome, complete genome |
| 7 | NC_003155.4 | *Streptomyces avermitilis* MA-4680 = NBRC 14893, complete genome |
| 8 | NC_003210.1 | *Listeria monocytogenes* EGD-e chromosome, complete genome |
| 9 | NC_003212.1 | *Listeria innocua* Clip11262 complete genome |
| 10 | NC_003366.1 | *Clostridium perfringens* str. 13 DNA, complete genome |
| 11 | NC_003450.3 | *Corynebacterium glutamicum* ATCC 13032 chromosome, complete genome |
| 12 | NC_003888.3 | *Streptomyces coelicolor* A3(2) chromosome, complete genome |
| 13 | NC_003902.1 | *Xanthomonas campestris* pv. campestris str. ATCC 33913 chromosome |
| 14 | NC_004668.1 | *Enterococcus faecalis* V583 chromosome, complete genome |
| 15 | NC_004722.1 | *Bacillus cereus* ATCC 14579 chromosome, complete genome |
| 16 | NC_005085.1 | *Chromobacterium violaceum* ATCC 12472, complete genome |
| 17 | NC_005296.1 | *Rhodopseudomonas palustris* CGA009 complete genome |
| 18 | NC_005957.1 | *Bacillus thuringiensis* serovar konkukian str. 97-27 chromosome |
| 19 | NC_006177.1 | *Symbiobacterium thermophilum* IAM 14863 DNA, complete genome |
| 20 | NC_006270.3 | *Bacillus licheniformis* ATCC 14580, complete genome" |
| 21 | NC_006361.1 | *Nocardia farcinica* IFM 10152 DNA, complete genome |
| 22 | NC_006582.1 | *Bacillus clausii* KSM-K16 DNA, complete genome |
| 23 | NC_006834.1 | *Xanthomonas oryzae* pv. oryzae KACC 10331, complete genome |
| 24 | NC_007005.1 | *Pseudomonas syringae* pv. syringae B728a chromosome, complete genome |
| 25 | NC_007404.1 | *Thiobacillus denitrificans* ATCC 25259, complete genome |
| 26 | NC_007406.1 | *Nitrobacter winogradskyi* Nb-255, complete genome |
| 27 | NC_007761.1 | *Rhizobium etli* CFN 42, complete genome |
| 28 | NC_008009.1 | *Candidatus Koribacter versatilis* Ellin345, complete genome |
| 29 | NC_008255.1 | *Cytophaga hutchinsonii* ATCC 33406, complete genome |
| 30 | NC_008536.1 | *Solibacter usitatus* Ellin6076, complete genome |
| 31 | NC_008555.1 | *Listeria welshimeri* serovar 6b str. SLCC5334 complete genome |
| 32 | NC_008593.1 | *Clostridium novyi* NT, complete genome |
| 33 | NC_008711.1 | *Arthrobacter aurescens* TC1, complete genome |

| | | |
|---|---|---|
| 34 | NC_008726.1 | *Mycobacterium vanbaalenii* PYR-1, complete genome |
| 35 | NC_009328.1 | *Geobacillus thermodenitrificans* NG80-2, complete genome |
| 36 | NC_009434.1 | *Pseudomonas stutzeri* A1501, complete genome |
| 37 | NC_009439.1 | *Pseudomonas mendocina* ymp, complete genome |
| 38 | NC_009441.1 | *Flavobacterium johnsoniae* UW101, complete genome |
| 39 | NC_009515.1 | *Methanobrevibacter smithii* ATCC 35061, complete genome |
| 40 | NC_009636.1 | *Sinorhizobium medicae* WSM419 chromosome, complete genome |
| 41 | NC_009674.1 | *Bacillus cytotoxicus* NVH 391-98, complete genome |
| 42 | NC_009720.1 | *Xanthobacter autotrophicus* Py2, complete genome |
| 43 | NC_009792.1 | *Citrobacter koseri* ATCC BAA-895, complete genome |
| 44 | NC_010001.1 | *Clostridium phytofermentans* ISDg, complete genome |
| 45 | NC_010002.1 | *Delftia acidovorans* SPH-1, complete genome |
| 46 | NC_010337.2 | *Heliobacterium modesticaldum* Ice1, complete genome |
| 47 | NC_010571.1 | *Opitutus terrae* PB90-1, complete genome |
| 48 | NC_010572.1 | *Streptomyces griseus* subsp. griseus NBRC 13350, complete genome |
| 49 | NC_010617.1 | *Kocuria rhizophila* DC2201 DNA, complete genome |
| 50 | NC_010655.1 | *Akkermansia muciniphila* ATCC BAA-835, complete genome |
| 51 | NC_010725.1 | *Methylobacterium populi* BJ001, complete genome |
| 52 | NC_010995.1 | *Cellvibrio japonicus* Ueda107, complete genome |
| 53 | NC_011000.1 | *Burkholderia cenocepacia* J2315 chromosome 1, complete genome |
| 54 | NC_011001.1 | *Burkholderia cenocepacia* J2315 chromosome 2, complete genome |
| 55 | NC_011002.1 | *Burkholderia cenocepacia* J2315 chromosome 3, complete genome |
| 56 | NC_011666.1 | *Methylocella silvestris* BL2, complete genome |
| 57 | NC_011886.1 | *Arthrobacter chlorophenolicus* A6, complete genome |
| 58 | NC_012483.1 | *Acidobacterium capsulatum* ATCC 51196, complete genome |
| 59 | NC_012490.1 | *Rhodococcus erythropolis* PR4 DNA, complete genome |
| 60 | NC_012560.1 | *Azotobacter vinelandii* DJ, complete genome |
| 61 | NC_012660.1 | *Pseudomonas fluorescens* SBW25 complete genome |
| 62 | NC_012669.1 | *Beutenbergia cavernae* DSM 12333, complete genome |
| 63 | NC_012778.1 | *Eubacterium eligens* ATCC 27750, complete genome |
| 65 | NC_012781.1 | *Eubacterium rectale* ATCC 33656, complete genome |
| 65 | NC_012881.1 | *Desulfovibrio salexigens* DSM 2638, complete genome |
| 66 | NC_012969.1 | *Methylovorus glucosetrophus* SIP3-4, complete genome |
| 67 | NC_013061.1 | *Pedobacter heparinus* DSM 2366, complete genome |
| 68 | NC_013131.1 | *Catenulispora acidiphila* DSM 44928, complete genome |
| 69 | NC_013132.1 | *Chitinophaga pinensis* DSM 2588, complete genome |
| 70 | NC_013595.1 | *Streptosporangium roseum* DSM 43021, complete genome |
| 71 | NC_013739.1 | *Conexibacter woesei* DSM 14684, complete genome |
| 72 | NC_013743.1 | *Haloterrigena turkmenica* DSM 5511, complete genome |
| 73 | NC_013757.1 | *Geodermatophilus obscurus* DSM 43160, complete genome |
| 74 | NC_013861.1 | *Legionella longbeachae* NSW150, complete genome |
| 75 | NC_013892.1 | *Xenorhabdus bovienii* SS-2004 chromosome, complete genome |
| 76 | NC_013929.1 | *Streptomyces scabiei* 87.22 complete genome |
| 77 | NC_013947.1 | *Stackebrandtia nassauensis* DSM 44728, complete genome |
| 78 | NC_014103.1 | *Bacillus megaterium* DSM319, complete genome |
| 79 | NC_014158.1 | *Tsukamurella paurometabola* DSM 20162, complete genome |
| 80 | NC_014170.1 | *Xenorhabdus nematophila* ATCC 19061 plasmid XNC1_p, complete genome |
| 81 | NC_014228.1 | *Xenorhabdus nematophila* ATCC 19061 chromosome, complete genome |
| 82 | NC_014259.1 | *Acinetobacter oleivorans* DR1, complete genome |
| 83 | NC_014323.1 | *Herbaspirillum seropedicae* SmR1, complete genome |
| 84 | NC_014391.1 | *Micromonospora aurantiaca* ATCC 27029, complete genome |
| 85 | NC_014551.1 | *Bacillus amyloliquefaciens* DSM7 complete genome |
| 86 | NC_014622.2 | *Paenibacillus polymyxa* SC2, complete genome |
| 87 | NC_014623.1 | *Stigmatella aurantiaca* DW4/3-1, complete genome |
| 88 | NC_014734.1 | *Paludibacter propionicigenes* WB4, complete genome |
| 89 | NC_014814.1 | *Mycobacterium gilvum* Spyr1, complete genome |
| 90 | NC_014958.1 | *Deinococcus maricopensis* DSM 21211, complete genome |
| 91 | NC_015177.1 | *Pedobacter saltans* DSM 12145, complete genome |
| 92 | NC_015514.1 | *Cellulomonas fimi* ATCC 484, complete genome |
| 93 | NC_015677.1 | *Ramlibacter tataouinensis* TTB310, complete genome |
| 94 | NC_016109.1 | *Kitasatospora setae* KM-6054, complete genome |
| 95 | NC_016629.1 | *Desulfovibrio africanus* str. Walvis Bay, complete genome |
| 95 | NC_016803.1 | *Desulfovibrio desulfuricans* ND132, complete genome |
| 97 | NC_016845.1 | *Klebsiella pneumoniae* subsp. pneumoniae HS11286 chromosome |
| 98 | NC_017384.1 | *Ketogulonigenium vulgarum* WSH-001 chromosome, complete genome |
| 99 | NC_017770.1 | *Solitalea canadensis* DSM 3403, complete genome |
| 100 | NC_017960.1 | *Enterococcus faecium* DO chromosome, complete genome |

| | | |
|---|---|---|
| 101 | NC_018750.1 | *Streptomyces venezuelae* ATCC 10712 complete genome |
| 102 | NC_020800.1 | *Xanthomonas axonopodis* Xac29-1, complete genome |
| 103 | NC_020815.1 | *Xanthomonas citri* subsp. citri Aw12879, complete genome |
| 104 | NC_020816.1 | *Xanthomonas citri* subsp. citri Aw12879 plasmid pXcaw19, complete genome |
| 105 | NC_020817.1 | *Xanthomonas citri* subsp. citri Aw12879 plasmid pXcaw58, complete genome |
| 106 | NC_020990.1 | *Streptomyces albus* J1074, complete genome |
| 107 | NC_020995.1 | *Enterococcus casseliflavus* EC20, complete genome |
| 108 | NZ_AP014683.1 | *Burkholderiales bacterium* GJ-E10 DNA, complete genome |
| 109 | NZ_CP002190.1 | *Bdellovibrio bacteriovorus* W, complete genome |
| 110 | NZ_CP007215.1 | *Enterobacter sacchari* SP1, complete genome |
| 111 | NZ_CP007557.1 | *Citrobacter freundii* CFNIH1, complete genome |
| 112 | NZ_CP009124.1 | *Streptomyces lividans* TK24, complete genome |
| 113 | NZ_CP009576.1 | *Listeria ivanovii* subsp. londoniensis strain WSLC 30151, complete genome |
| 114 | NZ_CP009962.1 | *Collimonas arenae* strain Cal35, complete genome |
| 115 | NZ_CP009963.1 | *Collimonas arenae* strain Cal35 plasmid, complete sequence |
| 116 | NZ_CP010028.1 | *Deinococcus swuensis* strain DY59, complete genome" |
| 117 | NZ_CP010946.1 | *Mycobacterium chelonae* genome |
| 118 | NZ_CP011253.2 | *Pandoraea oxalativorans* strain DSM 23570, complete genome |
| 119 | NZ_CP011451.1 | *Nitrosomonas communis* strain Nm2, complete genome |
| 120 | NZ_CP012329.1 | *Bacillus pumilus* strain NJ-M2, complete genome |
| 121 | NZ_CP012382.1 | *Streptomyces ambofaciens* ATCC 23877, complete genome |
| 122 | NZ_CP013106.1 | *Halomonas huangheensis* strain BJGMM-B45, complete genome |
| 123 | NZ_HG916826.1 | *Pseudomonas pseudoalcaligenes* CECT 5344 complete genome |

***Table 2.3:*** *Four artificial bacterial metagenomes created in MetaSim. The community complexity description is given, alongside the number of organisms at each abundance level. The relative abundance values used in MetaSim are given in parentheses.*

| **Metagenome** | **Community complexity** | **Metagenome composition** |
|---|---|---|
| A | Default test community (all organisms at same abundance) | 123 even abundance (100) |
| B | Low complexity | 4 high abundance (200)<br>119 standard abundance (100) |
| C | Medium complexity | 20 high abundance (200)<br>93 standard abundance (100)<br>10 low abundance (50) |
| D | High complexity | 20 high abundance (200)<br>20 medium-high abundance (150)<br>63 standard abundance (100)<br>20 low abundance (50) |

Each of the assemblers were subsequently run on the four metagenome variants (A-D) in turn. However, whilst many studies currently run assemblies using default parameter settings, it must be acknowledged that modifying the parameter values to fit the input metagenome size, structure and complexity can produce a better assembly (Howe *et al.,* 2014). Consequently, whilst a full parameter space optimisation was not the aim of this analysis, for each assembler, five parameter sets were tested on each metagenome, to obtain an improved result. The importance of this is to generate longer and more accurate contigs, which may provide enhancements for subsequent functional and phylogenetic annotations (Nielsen *et al.,* 2014). A schematic for this methodology can be identified in Figure 2.1, and the full list of parameter sets used is available in Appendix 1 Table 2. Key parameters to modify were the k-mer length (or word size) used to build assemblies, alongside the minimum read length and coverage. Parameter values were selected systematically, to test periodic intervals across the parameter space.

The final assemblies were evaluated based on both contiguity and completeness metrics. This allows the comparison to incorporate both the assembly length and coverage of the input dataset, to avoid large erroneous contigs masking the quality of the assembly (Earl *et al.,* 2011; Finotello *et al.,* 2011). The assembly contiguity was evaluated in Quast, using the measures of maximum contig length, contig N50 and number of contigs over 1000bp (Gurevich *et al.,* 2013). Assembly size was also evaluated, based on number of contigs in the assembly, and overall assembly size (bp). To analyse assembly completeness, the input metagenome reads were mapped to the assembled datasets using BWA (Li and Durbin, 2009). This identifies the amount of the input metagenome that has been used in the assembly, shown through the percentage coverage of the input dataset. Assembly chimerism was explored by searching for chimeric contigs using VSEARCH, enabling reference-based chimera detection (Rognes *et al.,* 2016). Using VSEARCH, the output metagenome assemblies were searched against the input artificial metagenomes using a global alignment to detect for chimeric sequences.

Additionally, the accuracy of the best metagenome assemblies in terms of the taxonomic distribution of organisms was evaluated using MetaPhlAn2 (Segata *et al.,* 2012). This method uses clade-specific markers to profile the assembled metagenome taxonomy and provide an output taxonomic distribution of the dataset (Segata *et al.,* 2012). This allows an insight into the presence of chimeric contigs, which may create erroneous organisms which were not present in the input dataset, due to joining sequences from unrelated bacteria (Lai *et al.,* 2012). This may also cause assemblers to eliminate highly abundant organisms, particularly when containing DNA sequences with numerous repeat regions or INDELS, as the assemblers find these difficult to resolve. This can be identified by viewing the top most abundant genus' between assemblies of the same metagenome, in comparison with the input dataset. Overall, the best performing assembly will most closely match the taxonomic profile of the input metagenome. Phylogenetic heat maps were created displaying the 25 most abundant clades in each metagenome in a logarithmic scale. Clustering was performed by average linkage and Euclidean distances for both clades and samples (Segata *et al.,* 2012). Finally, the best performing parameter set for each assembler (for each metagenome) was selected based on the above metrics.

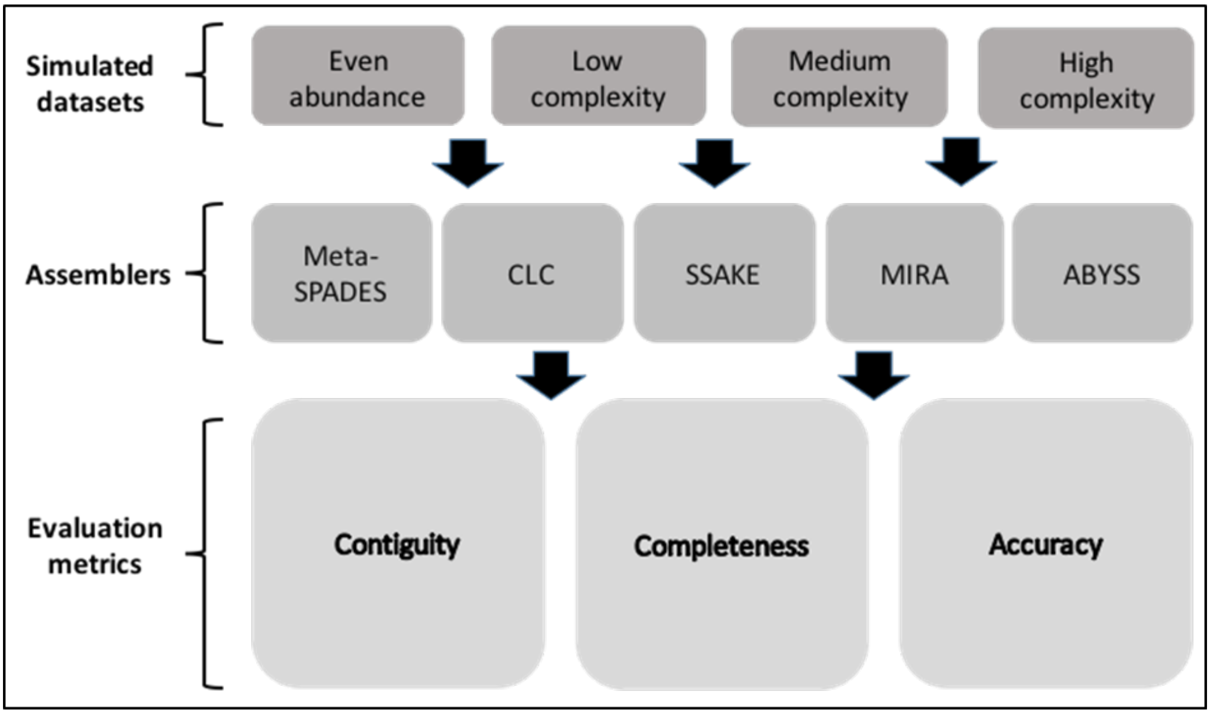


***Figure 2.1:*** *Overview schematic of the methodology used in assembler comparison.*

## 2.3 Results

### 2.3.1 Assembly coverage and contiguity

The best overall assemblies from the five tested assemblers are summarised in Table 2.4, with the full results available in Appendix 1 Table 3. As the study is primarily focused on assembler comparison (rather than parameter optimisation), the parameter set shown in Table 2.4 is the best performing out of the five tested, and not a product of testing the whole parameter space. For each assembly, a range of contiguity and completeness metrics are presented, including assembly N50, commonly used to describe contiguity. However, as the N50 is not a standardised metric, and relies on the dataset size, it should not be used to directly compare assemblies of different magnitudes.

***Table 2.4:*** *Summary table of metagenome assemblies, split by metagenome complexity (A-D). The single overall best performing parameter set for each assembler is shown, with results for all parameter sets available in Appendix 1. The parameter set is given in parentheses. Assembly size (number of contigs) and contiguity (N50, maximum contig length, contigs over 1000bp) statistics are given. The assembly completeness is identified by % coverage, which is the percentage of raw input metagenome reads that could be mapped to the final assembly.*

| Assembler | **CLC (2)** | **ABYSS (2)** | **MIRA (1)** | **MetaSPAdes (1)** | **SSAKE (4)** |
|---|---|---|---|---|---|
| Assembly | A | A | A | A | A |
| Number of contigs | 1334018 | 1227327 | 186625 | 666045 | 351881 |
| Contigs over 1000bp | 1369 | 143 | 373 | 939 | 21 |
| Maximum contig length | 5060 | 4641 | 5387 | 5299 | 1725 |
| N50 | 232 | 109 | 203 | 597 | 122 |
| % coverage of BWA mapping | 59.35 | 22.13 | 10.30 | 42.08 | 8.89 |

| Assembler | **CLC (2)** | **ABYSS (2)** | **MIRA (1)** | **MetaSPAdes (1)** | **SSAKE (4)** |
|---|---|---|---|---|---|
| Assembly | B | B | B | B | B |
| Number of contigs | 1289769 | 1204906 | 205976 | 645872 | 366959 |
| Contigs over 1000bp | 3234 | 146 | 389 | 2285 | 29 |
| Maximum contig length | 4585 | 4730 | 4773 | 4940 | 1643 |
| N50 | 234 | 112 | 210 | 624 | 123 |
| % coverage of BWA mapping | 59.56 | 23.56 | 12.24 | 43.03 | 9.75 |

| Assembler | **CLC (2)** | **ABYSS (2)** | **MIRA (1)** | **MetaSPAdes (1)** | **SSAKE (4)** |
|---|---|---|---|---|---|
| Assembly | C | C | C | C | C |
| Number of contigs | 1179804 | 1189639 | 272144 | 613234 | 417318 |
| Contigs over 1000bp | 6144 | 147 | 418 | 4622 | 28 |
| Maximum contig length | 7265 | 5090 | 5552 | 5386 | 1999 |
| N50 | 243 | 121 | 217 | 662 | 124 |
| % coverage of BWA mapping | 60.80 | 27.47 | 16.81 | 46.69 | 11.82 |

| Assembler | **CLC (2)** | **ABYSS (2)** | **MIRA (1)** | **MetaSPAdes (1)** | **SSAKE (4)** |
|---|---|---|---|---|---|
| Assembly | D | D | D | D | D |
| Number of contigs | 1038353 | 1253133 | 400485 | 613787 | 536263 |
| Contigs over 1000bp | 10876 | 189 | 456 | 8162 | 32 |
| Maximum contig length | 12412 | 4345 | 5931 | 10045 | 1743 |
| N50 | 281 | 138 | 218 | 682 | 126 |
| % coverage of BWA mapping | 67.00 | 36.04 | 24.67 | 57.31 | 16.13 |

As the best performing assembly will have both the longest contigs, and include the greatest portion of the input dataset, it is important to evaluate based on both contiguity and completeness, as opposed to a single metric. Figure 2.2 exemplifies the percentage coverage against a commonly used contiguity metric, maximum contig length. The results are provided for each metagenome (A-D), with increasing complexity (Table 2.4; Figure 2.2). For all metagenomes, the de Bruijn based assemblers, CLC, metaSPAdes and ABYSS output the assemblies with the highest coverage and maximum contig length (Table 2.4; Figure 2.2). For the best performing assembler, CLC, the coverage and contiguity scores range between 59 to 67% and 4585 to 12412 bp, between metagenomes, respectively (Table 2.4; Figure 2.2). In contrast, the two worst scoring assemblers are MIRA and SSAKE, based on OLC and Greedy assembly algorithms (Table 2.4; Figure 2.2). The lowest scoring assembler for all metagenomes, SSAKE, output coverage and contiguity scores ranging between 8 to 16% and 1643 to 1999bp, respectively.

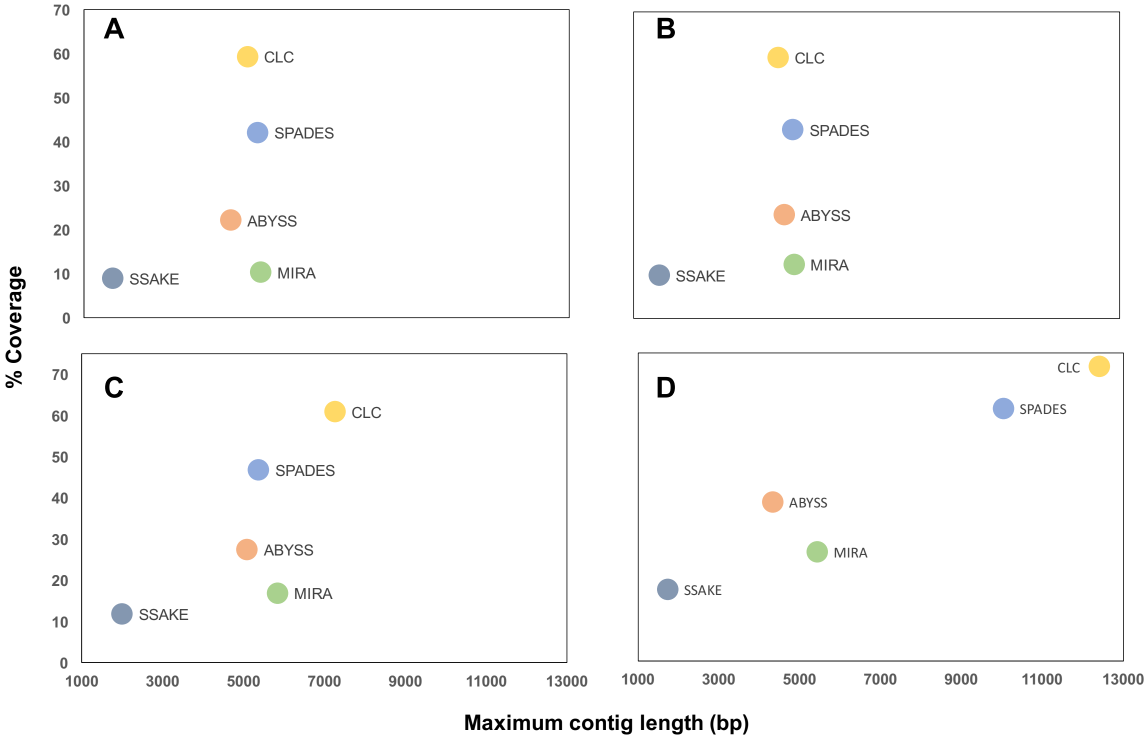


***Figure 2.2:*** *Maximum contig length (bp) and percentage coverage for assembled artificial metagenomes (A-D). For each assembler, the results for the best performing parameter set from the five tested are shown. Metagenome A: even organism abundance; Metagenome B: low organism complexity; Metagenome C: medium organism complexity; Metagenome D: high organism complexity.*

A similar pattern to that observed in contiguity is also identified when analysing assembly size (as number of contigs), with assembly coverage (Figure 2.3). The largest assemblies with the most coverage are produced by de Bruijn based assemblers CLC, metaSPAdes and ABYSS (Figure 2.3). Again, CLC is the best performing assembler across all metagenomes tested, with assembly sizes ranging between 1038353 to 1334018 contigs in metagenomes D and A respectively (Table 2.4). These assemblies are in contrast to those obtained from SSAKE, which range between 351881 and 536263 contigs for metagenomes A and D. Results from a one-way analysis of variance (ANOVA) on the best performing parameter sets, identifies significant differences between assemblers for each maximum contig length, number of contigs and percentage coverage (P= <0.05, Appendix 1 Table 4). This identifies that significant differences in assembly coverage, contiguity and size can be obtained by using different assemblers.

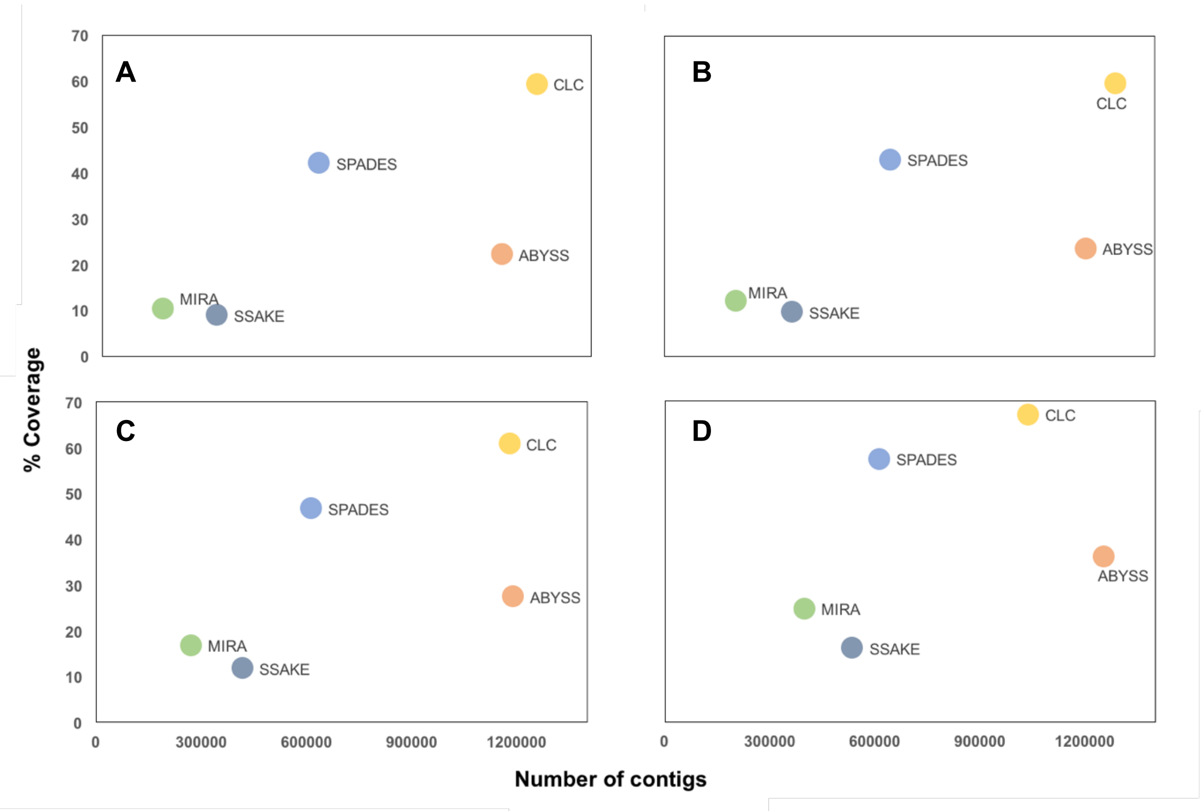


***Figure 2.3:*** *Number of contigs and percentage coverage for assembled artificial metagenomes (A-D). For each assembler, the results for the best performing parameter set from the five tested are shown.*

***Table 2.5:*** *Number of contigs, maximum contig length and percentage BWA coverage for each assembler. Values are reported as averages across all assemblies carried out for each assembler, over the four metagenomes (A-D), and the five parameter sets tested for each metagenome assembly.*

| **Assembler** | **Number of contigs** | **Maximum contig length** | **% coverage** |
|---|---|---|---|
| CLC | 455,954 | 5,728 | 30 |
| metaSPAdes | 188,893 | 3,699 | 16 |
| ABYSS | 246,006 | 2,206 | 8 |
| MIRA | 266,292 | 3,967 | 13 |
| SSAKE | 241,520 | 1,363 | 6 |

On average, de Bruijn based assembler ABYSS produces a maximum contig length of 2,206, with a mean maximum contig number of 246,006 (Table 2.5). This is comparable to both metaSPAdes and CLC, which obtained a maximum number of contigs, on average of 188,893 and 455,954, respectively (Table 2.5). However, ABYSS obtains lower assembly coverage values than the other de Bruijn graph assemblers, CLC and metaSPAdes (Table 2.5).

When looking at the best performing parameter sets (Table 2.4) and overall assembly averages (Table 2.5) SSAKE and MIRA, based on simple Greedy and OLC algorithms, performed worst in terms of assembly coverage (Figures 1-2). SSAKE produced the least complete metagenome assemblies, with coverage values at 6% of the input data on average, and MIRA outputting a slightly higher average of 13% (Table 2.5). SSAKE assembler also produced the lowest number of contigs, and contigs over 1000bp, with maximum values of 536,263 and 32 contigs, for assembly D (Table 2.4). This compares to the best performing CLC assembler, which output at total of 1,038,353 contigs for assembly D, with 10,876 contigs over 1000bp (Table 2.4). This represents a percentage difference of 64% and 199% for total number of contigs and contigs over 1000bp between CLC and SSAKE for assembly D (high complexity) (Table 2.4). This pattern is reflected in the lower complexity datasets, with percentage differences of 111% and 196% for these variables between CLC and SSAKE for assembly B (low complexity) (Table 2.4). This identifies SSAKE as producing less complete assemblies, with smaller more fragmented contigs, than the de Bruijn based assembler CLC.

For OLC based assembler, MIRA, the number of contigs output is comparable to de Bruijn based assemblers ABYSS and metaSPAdes on average (Table 2.5). Average maximum contig numbers are 226,292 for MIRA, compared to 188,893 and 246,006 contigs for metaSPAdes and ABYSS respectively, however, these are still lower than CLC averages at 455,954 contigs (Table 2.5). When looking at contiguity statistics, maximum contig length is again comparable (if not higher than) metaSPAdes and ABYSS for both the best performing parameter sets (Table 2.4, Figures 1-2) and the assemblies on average (Table 2.5). For example, for high complexity metagenome (D) MIRA has a maximum contig length of 5,931bp compared to 4,345bp, 10,045bp and 12,412bp for ABYSS, metaSPAdes and CLC respectively (Table 2.4). This indicates that both assembly size and contig length are comparable to the more complex de Bruijn based assemblers, whilst coverage is significantly lower (Table 2.4). This indicates that metaSPAdes and CLC outperform ABYSS in terms of assembly quality, as they are able to incorporate a greater amount of the input data in the resulting assemblies, thereby increasing assembly coverage.

### 2.3.2 Modifying community complexity

For each assembler, it is useful to compare the results across the different artificial metagenomes tested, as this will highlight how metagenome complexity influences the assembly outcome. Interestingly, in terms of the assembly coverage, N50 and number of contigs over 1000bp, overall assembler performance increases with greater metagenome complexity (Table 2.4; Table 2.5). For example, on average, assembly coverage (for best performing assemblers) rises from 36% to 50% between assembly A and assembly D (Table 2.4). This trend can be seen across all assemblers tested, with number of contigs over 1000bp increasing between 13% (CLC) and 770% (metaSPAdes), with an average increase of 2,473 (56%) in maximum contig length between metagenome A and D (Table 2.4). This indicates that with an increase in metagenome complexity, both the assembly contiguity and coverage improve, creating longer contiguous sequences, compared to shorter, fragmented sequences. However, whilst there are improvements in the assembly coverage and contiguity, the overall assembly size does not show a significant difference between low and high complexity metagenomes. On average, the assembly size only increases by 15,225 contigs (2%) between metagenomes A and D (Table 2.4). For de Bruijn graph assemblers, CLC and metaSPAdes, the number of contigs decreased by 22% and 8%, respectively (Table 2.4). Therefore, whilst contig length and coverage increases with greater complexity, the number of contigs assembled does not show a substantial improvement.

To test the significance of the metagenome complexity on the output assembly quality. A one-way ANOVA was carried out for each assembler independently. The contiguity, complexity and assembly size scores outlined in Table 2.4 were compared for each assembler across the four metagenomes tested. Overall, no significant difference could be observed in the assembly outcomes, when analysing metagenomes of different complexity (Appendix 1 Table 5). Consequently, whilst the metagenome complexity may play a role in modifying contiguity and completeness of assemblies, overall this change was not substantial. Therefore, it can be argued that the optimal assembler selected for the data type (e.g. soil or sediment microbial communities) could be applied to numerous datasets of a similar type, with different community compositions.

### 2.3.3 Assembler parameterisation

Whilst it may be possible to identify a suitable generic assembler for soil microbial communities, each assembler has a range of parameter values that can be modified (Appendix 1 Table 2). Whist assembler optimisation is not the fundamental aim of this study, it is important to acknowledge the influence these have on assembly outcomes. Hence, for each assembler, 5 different parameter sets were tested, with the full results available in Appendix 1 Table 3. This gives an indication of the sensitivity of assembly outcomes to parameter values and helps highlight assemblers where more thorough parameter optimisation may be required. The results of parameter testing for each assembler are given in Figures 4 and 5, showing percentage coverage and maximum contig length, respectively. This gives an indication of the possible spread of assembly quality that can be obtained by modifying parameters, such as the Kmer length. De Bruijn graph assemblers CLC, MetaSPAdes and ABYSS scored the highest values for both contiguity and coverage (Figure 2.4; Figure 2.5). CLC recorded the maximum values in both metrics, across all metagenomes tested, followed by metaSPAdes (Figure 2.4; Figure 2.5). For example, for CLC, maximum contig length reached 14,158 bp and a coverage of 67% for parameter sets 2 and 3 respectively (Metagenome D; Figure 2.4; Figure 2.5). This is compared to SSAKE, the worst performing assembler, which reached maximum values of 2064 bp (contig length) and 16% (coverage) for parameter sets 3 and 4 respectively (Metagenome D) (Figure 2.4). MetaSPAdes and ABYSS recorded maximum coverage values of 57% and 36% for Metagenome D, with the highest contig lengths reaching 10,045 bp and 50,90 bp respectively (Figure 2.4; Figure 2.5).

Despite de Bruijn based assemblers showing the most promising assembly outcomes, the spread of contiguity, completeness and size values across parameter sets far exceeds the simpler OLC and Greedy algorithm-based assemblers (Figure 2.4). This indicates simpler assemblers have a smaller range of outcomes, in comparison to more complex, highly modifiable assemblers. For example, for Metagenome D, CLC and metaSPAdes have a range of maximum contig lengths between 11,964 bp and 6,818 bp, respectively (Figure 2.4). This is compared to SSAKE, which had a range of 278 bp. This pattern is also reflected in the range of coverage scores, with CLC showing a spread of 66%, compared to 15% for SSAKE, in assemblies of Metagenome D (Figure 2.4). Consequently, whilst de Bruijn based assemblers can provide the highest scoring coverage and contiguity values, they can also provide the largest range in scores. This is compared to the assemblers based on simpler OLC and Greedy algorithms, that express a narrower range of outcomes, despite modifying parameter values. This is indicative of pronounced sensitivity to parameterisation in more complex assemblers, and thus a greater need for parameter set optimisation or testing.

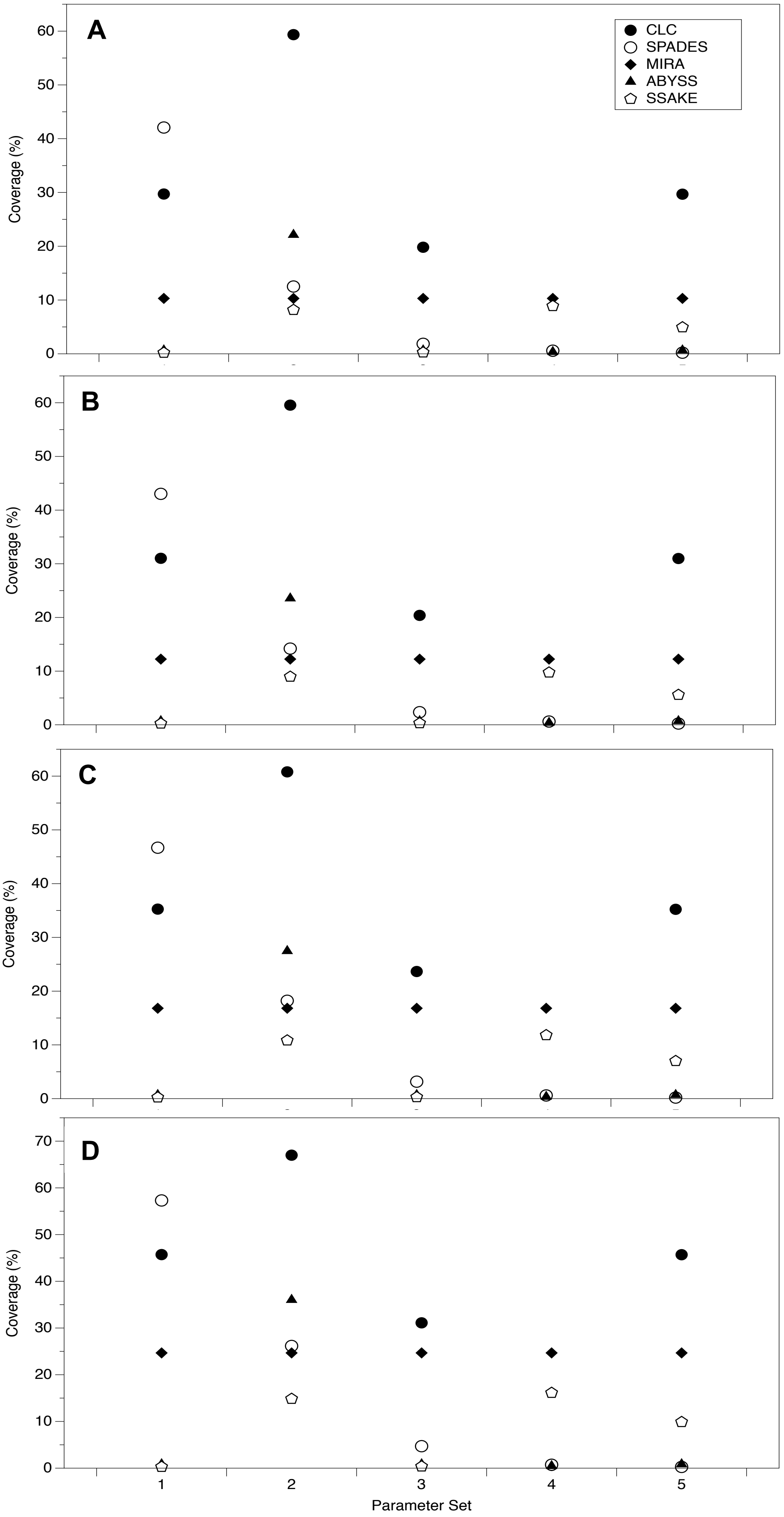


***Figure 2.4:** Coverage (%) of metagenome assemblies against the input dataset, for artificial metagenomes A-D. Results for each parameter set (1-5) are shown, with values for each assembler distinguished. Assemblers tested include: CLC, metaSPAdes MIRA, ABYSS, SSAKE.*

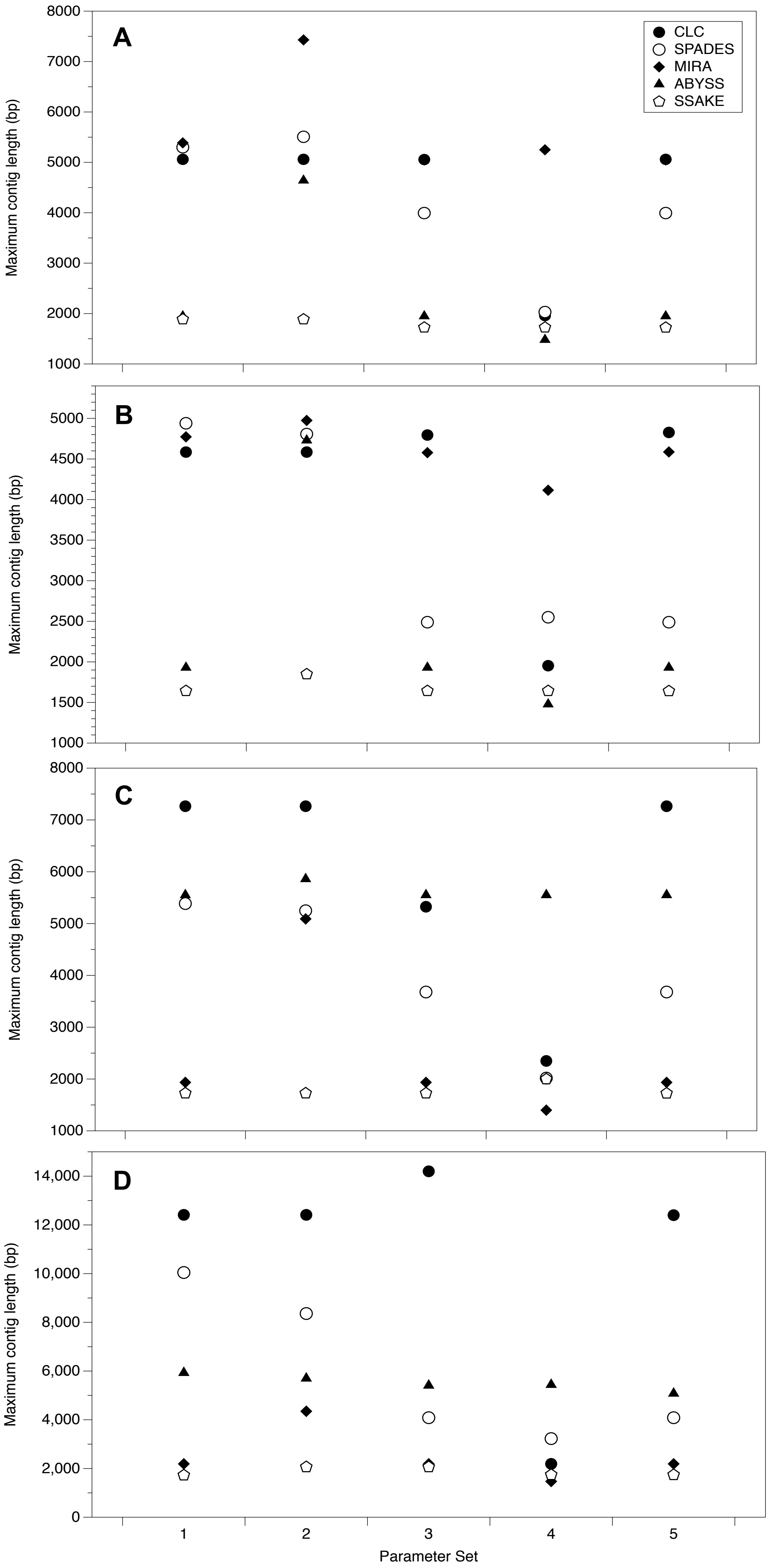


***Figure 2.5:*** *Maximum contig length (bp) of metagenome assemblies against the input dataset, for artificial metagenomes A-D. Results for each parameter set (1-5) are shown, with values for each assembler distinguished. Assemblers tested include: CLC, metaSPAdes MIRA, ABYSS, SSAKE.*

### 2.3.4 Assembly chimeras

For all assemblies, chimeric (incorrectly assembled) contigs were identified using VSEARCH, which utilises an optimal global aligner to identify misassembled contigs in the final assemblies (Rognes *et al.,* 2016). This highlights which assemblers produce more accurate contigs, in comparison to those which produce longer, but incorrect sequences, which would adversely affect downstream functional analysis (Wooley *et al.,* 2010). As identified in Table 2.4, both CLC and metaSPAdes produced the longest contig lengths and metagenome coverage. However, CLC produced the largest percentage of chimeric contigs, in relation to the total number of contigs assembled, with a maximum number of chimeras at 0.1% for metagenome B, compared to 0.006% for metaSPAdes (Figure 2.6). This indicates that whilst CLC produces long contigs, they are slightly more prone to errors. The lowest number chimeric contigs were found in MIRA and SSAKE assemblers, accounting for a maximum of 0.002% of total assembled contigs (Figure 2.6). However, as these assemblies are generally smaller and more fragmented, they may be less useful for downstream functional analysis. For all assemblers tested, assemblies of metagenome D (high complexity) produced the lowest percentage of chimeric contigs, ranging between 0 and 0.06%, between MIRA and CLC assemblers respectively (Figure 2.6). This is in agreement with the coverage and contiguity statistics presented in Table 2.4, identifying metagenome D as having the best overall assembly statistics across all the artificial metagenomes tested (Table 2.4).

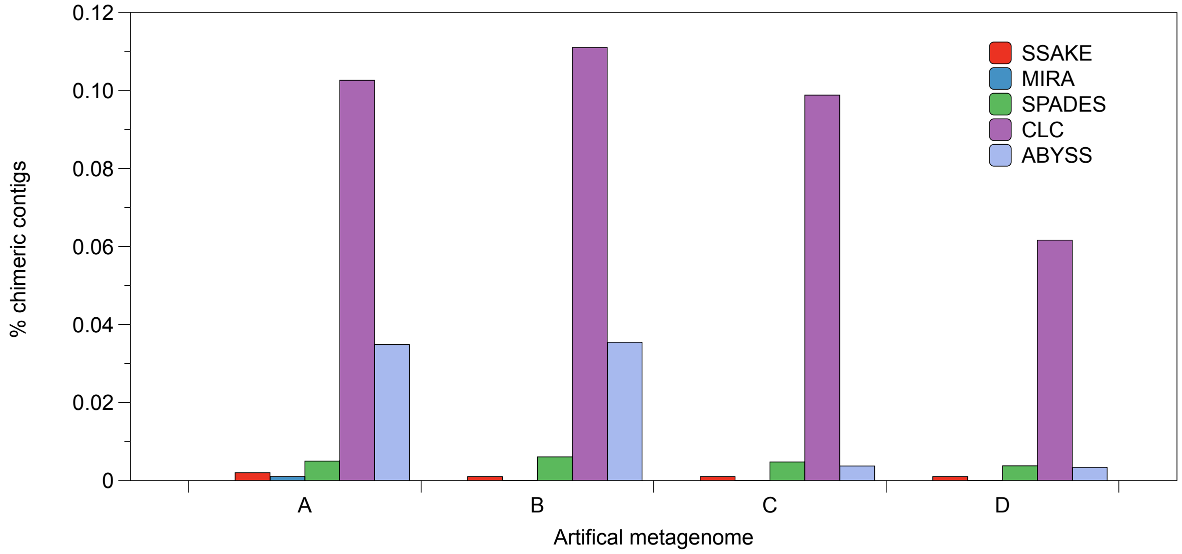


***Figure 2.6:*** *Chimeras as a percentage of total contigs for assemblies of metagenomes A-D, for each assembler. The results for the best performing parameter set for each assembler are shown (identified in Table 2.4).*

### 2.3.5 Taxonomic distribution of assemblies

Alongside using chimera searches to highlight assembly accuracy, identifying the taxonomic distribution of outputs can also provide an insight into assembly quality. MetaPhlAn2 was used to evaluate the taxonomic distribution of assemblies of each metagenome across all five assemblers tested, in comparison to the input unassembled metagenome (Figure 2.7; Segata *et al.,* 2012). This helps to highlight erroneous assemblies by isolating unusual community profiles, or over/under abundant organisms in assembled data (Segata *et al.,* 2012). Greedy based assembler, MIRA, was shown to perform poorly in representing the input metagenome in resulting assemblies. At the genus level, this assembler over represented genus' including *Listera, Pseudomonas and Xanthomonas*, in comparison to the input dataset (Figure 2.7). This was shown to vary significantly between input metagenomes, with substantial over representation of *Gammaproteobacteria* and *Deltaproteobacteria* for high complexity Metagenome C, in comparison to the other assemblers (Figure 2.7). This pattern was not reflected in Metagenomes A, B and D. However, SSAKE, the second greedy-based assembler tested, appeared to more accurately represent the input metagenome in resulting assemblies (Figure 2.7). Additionally, whilst outputs from ABYSS, CLC and metaSPAdes (de Bruijn graph based assemblers), did not completely recreate the input dataset, they did not show a consistent bias in the taxonomic distribution of organisms (Figure 2.7). Consistent bias in assemblies is more problematic for downstream analysis, as inaccurate conclusions about the community composition may be drawn, as opposed to random assembly error, which is unlikely to skew the overall community distribution.

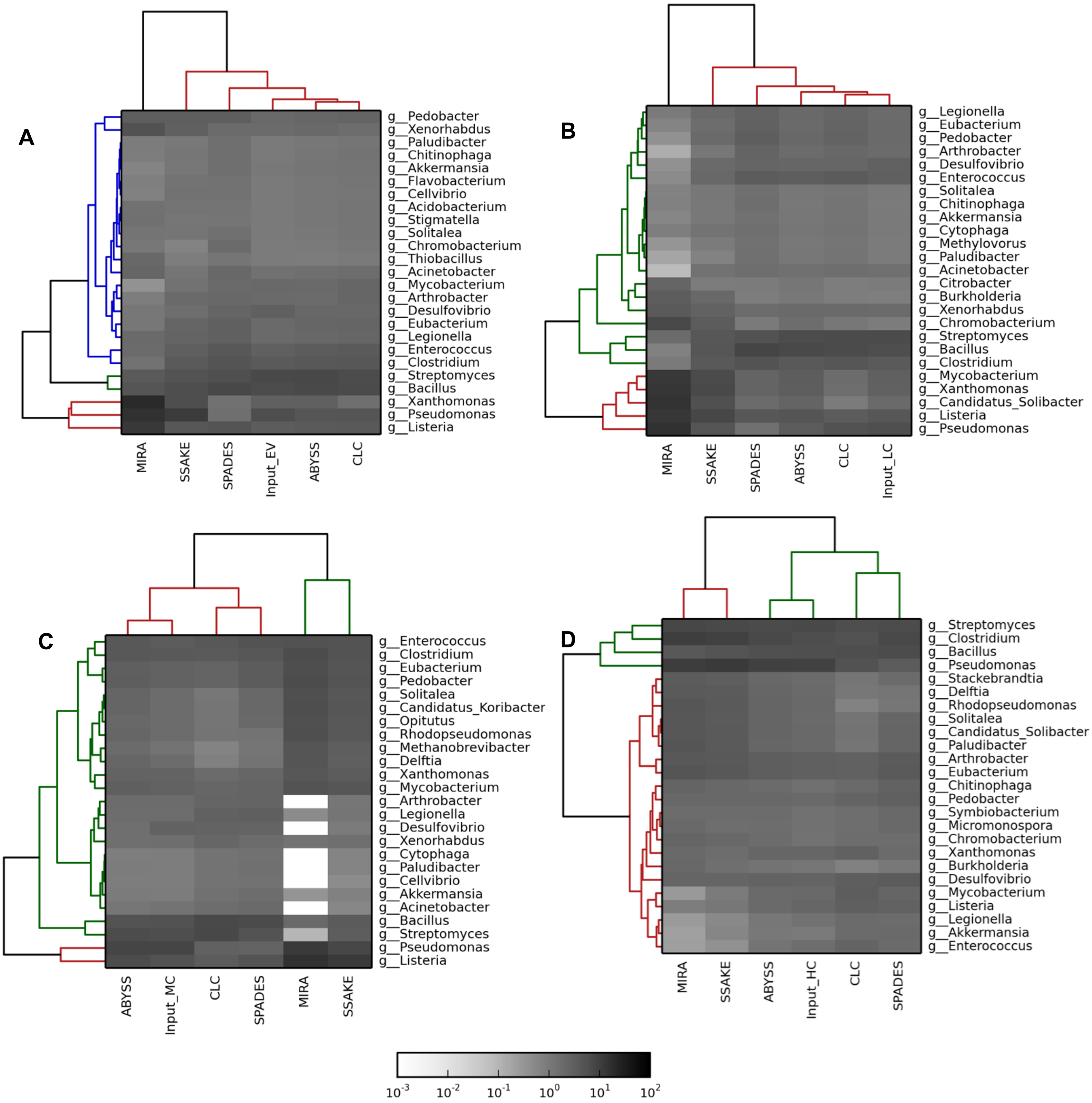


***Figure 2.7:*** *Taxonomic heat maps for each artificial metagenome assembled (A-D), showing the top 25 most abundant genus', clustered based on Euclidean distances. Relative genus abundance for the best parameter set for each assembler are shown, alongside the input microbial community for comparison (flagged by Input_MC).*

## 2.4 Discussion

### 2.4.1 Premise of the study

This study tested the performance of five metagenome assembly tools for the assembly of complex soil bacterial metagenomes. The artificial metagenomes were simulated using an identical community composition, but were modified based on organism abundance profiles, to produce metagenomes at a range of complexities.

Simulated metagenome datasets are generally simplifications of true environmental communities (Vazquez-Castellanos *et al.,* 2014). However, as the aim of this study was to test metagenome assemblers, rather than to study the taxonomy of the dataset, this does not undermine the conclusions drawn (Vazquez-Castellanos *et al.,* 2014). Additionally, using test datasets where the species composition is known *a priori*, allows a direct comparison of the assemblers (Mavromatis *et al.,* 2007). This would not be possible using true bacterial metagenomes, as the species composition is unknown prior to assembly (Mavromatis *et al*., 2007). Consequently, using simulated datasets allows a comparison of how well the resulting assemblies represent the 'true' input bacterial sequence composition. The best performing assemblers will most accurately represent the composition of the test dataset, have minimal chimeric contigs, and have a large contig length to maximise the quality of downstream functional and taxonomic annotation (Howe and Chain, 2015).

### 2.4.2 Assembly evaluation metrics

This study has demonstrated the importance of using more than one evaluation metric for assembly comparison, and that measures incorporating contiguity, completeness and accuracy should be used. This is important as no single metric is comprehensive enough to summarise overall assembly quality (Kurtz *et al.,* 2004; Peng *et al.,* 2012; Vollmers *et al.,* 2017). When looking at assemblies based solely on contiguity, CLC shows the overall highest number of contigs and contig length, in comparison to SSAKE which produces the shortest contigs, and therefore most fragmented assemblies. This is important, as long contigs, such as those produced by CLC (and metaSPAdes) can provide better alignments to gene databases for functional annotation. However, when also considering the number of chimeric contigs, CLC exemplified the highest number of chimeras across all tested datasets (Figure 2.6). This indicates that whilst CLC is producing long contigs, these are sometimes miss-assembled, incorporating read fragments from different organisms. This could subsequently produce some incorrect taxonomic and functional annotations in downstream analysis

(Mavromatis *et al.,* 2007). Whilst this is not a substantial proportion of the overall number of contigs produced during CLC assemblies (0.1%) it is important to select an assembler that can produce both long, and correctly assembled contigs. As metaSPAdes was able assemble contigs with substantial lengths (Table 2.4) and with a very minimal number of chimeras, at 0.006% of total contigs (Figure 2.6), this assembler also provides a reasonable trade-off between these metrics. Interestingly, an inter-comparison of assemblers by Vollmers *et al.,* (2017) based on real Illumina sequencing data (forest soil and algal biofilm) selected metaSPAdes as having the best trade-off between assembly size and coverage of diversity, in comparison to the assemblers tested. Whilst this study was primarily focused on de Bruijn graph-based assemblers, and did not explore parameter values, it does indicate the performance of this assembler is upheld in real sequence data (Vollmers *et al.,* 2017).

### 2.4.3 Simulated bacterial communities

This study also demonstrated the use of simulated datasets in conducting metagenome assembly evaluation. Using artificial metagenome datasets allowed the assemblies to be compared based on percentage coverage and taxonomic composition, in comparison to the 'true' community structure (Sczyrba *et al.,* 2017). This consequently enabled the accuracy of the assemblers to be evaluated, alongside using more standard measures of assembly size and contiguity. Out of the two simpler 'Greedy' based assemblers tested (MIRA and SSAKE), MIRA showed the highest number of contigs, contig length and coverage of the input dataset (Table 2.4). However, when investigating the taxonomic composition of MIRA assemblies in comparison to the input dataset, MIRA is shown to both under and over represent the 25 most abundant organisms (Figure 2.7). This assembler is therefore not completely representing the true species composition and therefore may produce incorrect downstream annotations and ecological conclusions. Consequently, this demonstrated the value of using artificial datasets, and the importance of selecting a good assembler, in order to produce un-fragmented accurately assembled contigs (Risenfeld *et al*., 2004).

Here, metagenome assemblers were also tested on several simulated communities, at a range of complexities, as opposed to a single dataset. This is because each assembler is formatted based on a different algorithm, with varying parameter options and will therefore suit datasets with different sizes and community structures (Miller *et al.,* 2010; Sczyrba *et al.,* 2017). Consequently, assemblers showed some differences in performance across the simulated communities, in relation to the algorithm on which they are based. Overall, high complexity metagenome 'D' was shown to be assembled with the highest coverage, contig lengths and the minimal number of chimeras (Figure 2.2; Figure 2.6). This contrasts with low complexity metagenome ‘B’, and even abundance metagenome ‘A’, which demonstrated highest numbers of chimeric contigs and lowest contig lengths (Table 2.4; Figure 2.6). These differences in outcomes between metagenomes can be attributed to the community abundance structure, as the taxonomic composition is identical. These differences in assembly outcome may be due to raised abundance of some organisms, and subsequent increases in its read coverage in the dataset. Organisms with high abundance are likely to be well assembled, pulling the quality of the assembly up, as there is more data support for each DNA sequence. Metagenome B was classified as the low complexity community as it was comprised of 4 highly abundant organisms, with the remainder as low abundance flanking organisms (Table 2.3). This is in line with Mavromatis *et al*., (2007) whom created test communities in MetaSim, to reflect simpler bacterial populations with a select group of organisms. This is similar to Metagenome A, where all the organisms have the same level of abundance. It is likely that these metagenomes were assembled poorly as there were limited organisms at a high abundance, and therefore most organisms had a low read coverage. This low coverage could have increased the difficulty in resolving artificial sequencing errors, gaps and repeats (Sims *et al.,* 2014). On the other hand, Metagenome D had more organisms at a higher abundance, and therefore the increase in read coverage may have improved the assembly outcome, despite having a more complex species distribution (Table 2.4; Sims *et al.,* 2014). However, the overall differences between assemblies of the different metagenomes were found to be insignificant for all assemblers (Appendix 1 Table 5). In the instance of low coverage datasets, read based analysis, as opposed to assemblies may therefore be beneficial for taxonomic investigations. This is because using reads will be able to profile more of the community, as opposed to assemblies which may not utilise all of the sequencing data.

### 2.4.4 Choice of assembler

In contrast to community complexity, overall assembler selection was found to have a significant influence on the quality of the assembly (Appendix 1 Table 4). This indicates that overall assembler selection for the type of data in question (e.g. soil or sediments) is important, however the complexity of that community does not have a substantial influence on the assembly outcome (Table 2.4; Figure 2.2; Figure 2.3). For metaSPAdes assemblies, the percentage coverage increased by 15% between Metagenomes A to D, up to a maximum of 57% (Table 2.4). This can be compared to the increase of 7% up to a maximum coverage of 16% for SSAKE. Thus, the difference between SSAKE and metaSPAdes coverage scores for Assembly D is 41%, exceeding the 15% and 7% increase in coverage obtained by changing metagenome complexity for metaSPAdes and SSAKE, respectively (Table 2.4).

### 2.4.5 The importance of parameterisation

The analysis highlighted that modifying key parameter values, such as the kmer length, can influence the assembly outcome (Figure 2.4; Figure 2.5). This was an important aspect to test, as many studies will run assemblies using default parameter values, which may not obtain the most optimal results (Howe *et al.,* 2014; Vollmers *et al.,* 2017). This was shown to be particularly evident for more complex de Bruijn Graph assemblers such as CLC and metaSPAdes (Figure 2.4; Figure 2.5). This is likely to relate to the wider range of parameter settings available in these assemblers or their sensitivity to changing key variables such as k-mer values. Whilst simpler assemblers often produce lower quality results, these are generally more consistent between parameterisations. Consequently, in order to obtain an improved assembly outcome, it is beneficial to iterate through several sets of parameter values, before making the final selection. This is in agreement with the CAMI inter-comparison, which identified parameters such as the kmer length, to be highly influential on assembly outcome (Sczyrba *et al.,* 2017). For studies which are highly focused on obtaining the most optimal assembly quality, parameter space optimisation can be used to select the most favourable parameter values (Chikhi & Medvedev, 2013). Some assemblers, such as metaSPAdes already incorporate optimisation of kmer length selection, by iterating through several values during the assembly process, before selecting the most appropriate to the dataset in question (Nurk *et al.,* 2017).

### 2.4.6 Best performing assemblers

Overall, CLC and metaSPAdes, de Bruijn graph-based assemblers, were shown to perform best in terms of contiguity, size and coverage across all datasets (Figure 2.2; Figure 2.3). Greedy algorithm based assembler, SSAKE was shown to perform worst over these metrics. Furthermore, OLC based assembler, MIRA, did not show a substantial improvement in assembly outcome from the worst performing assembler, SSAKE (Figure 2.2; Figure 2.3). However, when considering the number of incorrectly assembled contigs, CLC shows the highest proportion of chimeras relative to its dataset size (Figure 2.6). Whilst this is not a substantial proportion, it is higher than the other assemblers tested. In contrast, simpler assemblers, MIRA and SSAKE show a smaller proportion of incorrectly assembled contigs in their resulting assemblies (Figure 2.6). However, as these assemblers have created much shorter contig lengths, and therefore smaller, more fragmented assemblies, the number of long, incorrectly joined reads is lower overall. However, as the contig lengths are much shorter for these assemblers than the de Bruijn graph-based assemblers, CLC and metaSPAdes, the accuracy of downstream functional annotation could be limited (Narzisi and Misha, 2011).

Consequently, it can be suggested that metaSPAdes and CLC produce suitable assemblies for the investigation of complex bacterial metagenomes formed from environmental samples. This is because the assemblies produced are accurate, large and provide long contigs for downstream annotations (Table 2.4; Figure 2.6; Figure 2.7). Whilst CLC does produce slightly more chimeric contigs, this is arguably a small proportion of contigs and therefore should not impact the inferences from the bulk of the data.

### 2.5 Conclusion

This study has used artificially simulated metagenomic data to test publicly available metagenome assemblers, to provide better guidance on assembler selection for those investigating soil and sediment bacterial communities. The assemblers were selected to cover the range of assembly algorithms available, including de Bruijn graph assemblers (metaSPAdes, CLC, ABYSS), OLC assemblers (MIRA) and simple Greedy assemblers (SSAKE).

The study has shown the importance of using more than one metric when evaluating assemblies. Using one type of metric, such as contiguity, may provide misleading results, as long contigs may not necessarily be accurate. Incorrectly assembled contigs may provide inaccurate functional and taxonomic annotation in downstream analysis and may subsequently influence ecological conclusions. Evaluating based on assembly size may also be misleading, as large assemblies of short contigs will not provide optimal alignments to databases for functional annotation. Fragmented contigs may therefore inaccurately represent key groups of the dataset, as shown by MIRA assemblies, which missed out several highly abundant genera's. Consequently, testing the assemblers on artificial communities proved useful, as the assemblies could be evaluated based on their accuracy and coverage of the 'true' dataset. A combination of contiguity, assembly size, coverage and accuracy metrics are therefore needed to adequately summarise the quality of the metagenome assembly, or to compare between assemblies.

Assembler selection was shown to be important for the output quality, with de Bruijn based assemblers, CLC and metaSPAdes producing assemblies with the highest coverage, contiguity and assembly size. Assembly outcomes were shown to be significantly different across the assemblers tested, and therefore identifying a suitable assembler for the data type is important for obtaining high quality results. Whilst the assemblers were tested across communities at a range of complexities, no significant difference could be found between these, when looking at a single assembler. It is arguably more important to select an appropriate assembler for the data type, which can then be applied to similar data without significant disadvantages. Some differences were obtained when modifying the community complexity, mainly in terms of assembly coverage, N50 and number of contigs over 1000bp. Therefore, if computational power and time permits, it is recommended to test the assemblers for the dataset in question, rather than just the data type. However, we would not expect substantial reductions in assembly quality when using an assembler tested on a similar data type.

Modifying parameter values was also shown to influence output quality. Whilst the assembly itself can improve the dataset quality for functional and taxonomic annotation, by testing several parameter sets, the assembly outcome can be improved. Whilst providing a full parameter optimisation was not the aim of this study, we have shown that the more complex de Bruijn graph assemblers are more sensitive to parameter settings, such as the kmer length. It would therefore be beneficial to test a few parameter sets, or carry out a parameter optimisation, if assembly quality was a key concern.

Overall, we have identified the metaSPAdes and CLC assemblers as providing the highest quality assemblies for soil and sediment bacterial datasets. These assemblers are recommended for use on datasets of similar types. However, the more the user attempts to improve the assembly outcome (i.e. testing for the dataset in question and modifying parameter settings), the greater the outcome quality can be. Improving the quality of the assembly will not only increase the reliability of the functional and taxonomic annotations, but also allow more detailed investigation of the microbial community. For example, improved annotation may aid extraction of draft genomes from metagenome datasets.

### 2.6 Observations

The results of this chapter have highlighted a series of observations, when considering metagenome assembly within an analysis pipeline.

- Assembling metagenome reads can increase their length, which may help downstream annotation.
- Assemblers do not all perform the same, the most appropriate choice will depend on the data type and complexity.
- If possible, testing a selection of assemblers may help determine the most appropriate choice.
- Testing of assemblers can be carried out on real environmental data, or a simulated community (which will allow the evaluation of accuracy).

- Assemblers should be evaluated on multiple metrics: contiguity, completeness, size and accuracy.
- Parameter values can influence assembly outcomes, especially for de Bruijn graph assemblers. Testing several parameter sets may help improve assembly quality.
- Once testing has been carried out, the same assembler can be used on multiple samples from a similar environment, without significant cost to assembly quality.
- Assembly may not always be the most appropriate choice, for example, given high fragmented datasets with low sequencing coverage.

### 2.7 Recommendations for best practice

The results of this analysis have highlighted some recommendations for best practice when using metagenome assembly as part of an analysis pipeline. Here it has been shown that de Bruijn graph assemblers, such as CLC and metaSPAdes are most appropriate for complex soil metagenomes. This is likely related to the use of a graph-based algorithm which identifies read overlaps and read layout, alongside multiple tuneable parameters.  However, we recommend that if the resources are available, several de Bruijn graph-based assemblers are tested on the dataset under analysis, to obtain an improved outcome. Comparing and selecting an assembler can increase read coverage, assembly contiguity and completeness, which will improve downstream annotation opportunities. This analysis has shown that the simpler Greedy assemblers and those designed for single genomes should not be applied to metagenomic assembly. This is because these assemblers produce small localised assemblies through read overlaps, often creating fragmented datasets which hinder annotation.

Once an assembler has been selected, multiple parameter sets should be tested to obtain an improved result. This is especially the case for de Bruijn graph assemblers, which are sensitive to key parameter values such as the kmer length. A consideration should also be made to the expected composition and complexity of the dataset. This analysis has shown substantial differences in assembly outcome based on the complexity of the community, and that assembler choice should take this into account. For example, a simple community monopolised by a few microbes may assemble reasonably well with a Greedy based assembler, given the increased read coverage for these organisms. However, with a complex community structure, such as found in soils, a variability in microbial abundance and therefore read coverage requires the use of a more tuneable assembler. Given limited computational resources and time, it is recommended that those assembling soil metagenomes use a de Bruijn graph assembler, and test several values of kmer length as a minimum. Some assemblers, such as metaSPAdes incorporate kmer length optimisation as part of the assembly process.

### 2.8 Limitations and additional work

As the aim of this study was to test overall assembler function on simulated bacterial communities, parameter space optimisation was not carried out. Whilst the study could have been conducted by running each assembler on default values, we also wanted to show how modifying some parameter settings could enhance the results available. In this study we tested a total of five parameter sets for each assembler, including the default values. Whilst this was sufficient to highlight the sensitivity of more complex de Bruijn graph assemblers to parameters such as the kmer length, greater analysis of the parameter space could be carried out. This would include running optimisation of key parameters such as the kmer length, minimum contig length and bubble size (Chikhi and Mendelev, 2013). Identifying the optimal parameter set is likely to improve the quality of the assembly outcome. This would be particularly important for studies that are interested in more detailed aspects of the microbial community structure, such as gene arrangements or assembling single draft genomes (Eloe-Fadrosh *et al.,* 2016b).

A subset of publicly available assemblers were selected for this study, in order to cover the three main assembly algorithms available. To obtain the most optimal assembler for soil microbial data, all publicly available assemblers could be included. This would require collaboration between several research centres, as carried out by Assemblethon 2 for the assembly of single vertebrate genomes (Bradman *et al.,* 2013). However, we have shown here that by selecting a de Bruijn graph assembler, such as CLC or metaSPAdes, assemblies of improved quality can be obtained compared to using simpler OLC or Greedy assemblers, which are better fitted to longer sequencing reads. In future, the soils community may wish to conduct a wider inter-comparison for metagenome data, in order to highlight the best performing assembler. This in itself has limitations, as the assemblers may only be tested on one data type, and as the field is constantly evolving with the release of new assemblers, this will also only be a snapshot of the assembly platforms available.

This study tested the selected assemblers on a series of artificial bacterial metagenomes. This was carried out so that the ‘true’ bacterial community was known, and therefore the assembly accuracies could be evaluated. The assemblers could also be tested on real bacterial data, however, this would limit the evaluation metrics to contiguity, coverage and size criteria. The assemblies could also be repeated in order to test the reliability of each assembly, and how variable the assembly outcomes can be. As far as the authors are aware, there is no current knowledge on the reliability of assemblers when repeating assemblies of the same dataset.

# Chapter 3: Metagenomic insights into diazotrophic communities across Arctic glacier forefields

Maisie V. Nash[1]; Alexandre M. Anesio[2]; Gary Barker[3], Martyn Tranter[1], Gilda Varliero[3], Emiley A. Eloe-Fadrosh[4], Torben Nielsen[4], Thomas Turpin-Jelfs[1], Liane G. Benning[567] and Patricia Sánchez-Baracaldo[1]

[1]School of Geographical Sciences, University of Bristol, BS8 1SS, UK
[2] Department of Environmental Science, Aarhus University, PO box 358, Denmark
[3] School of Life Sciences, University of Bristol, BS8 1TQ, UK
[4]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, US
[5]GFZ German Research Centre for Geosciences, Telegrafenenberg, 14473 Potsdam, Germany
[6]School of Earth and Environment, University of Leeds, LS2 9JT, Leeds, UK
[7]Department of Earth Sciences, Free University of Berlin, Malteserstr, 74-100, Building A, 12249, Berlin, Germany

### Contributions and acknowledgements

This chapter has been published in FEMS Microbiology Ecology, volume 94, issue 9 (12 June 2018). The full reference is detailed below:

Nash, M. V., Anesio, A. M., Barker, G., Tranter, M., Varliero, G., Eloe-Fadrosh, E. A., ... & Sánchez-Baracaldo, P. (2018). Metagenomic insights into diazotrophic communities across Arctic glacier forefields. *FEMS microbiology ecology*, *94*(9), fiy114

MN carried out the data analysis, interpretation and manuscript formulation. AA, GB and MT carried out field sampling. TTJ supplied the raw carbon and nitrogen data for the Midtre Lovénbreen forefield. GV carried out the read mapping. All co-authors were involved in the refinement and corrections of the manuscript. Additional support came from Katherine Wright, Joshua Blacker and James Bradley during field sampling. Jane Coghill and Christy Waterfall at the Bristol Genomics Facility carried out the DNA sequencing and Nathan Chrismas provided data analysis support.

### Funding

This work was carried out as part of the NERC GW4+ Doctoral Training Partnership, awarded to MN. Funding support was provided through NERC grants NE/J02399X/1, awarded to AA and NE/J022365/1 awarded to LGB. Funding support for PS-B came from a Royal Society University Research Fellowship and LGB also acknowledges the financial support through the Helmholtz Recruiting Initiative. Funding support for GV came from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675546 – MicroArctic. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

### 3.1 Introduction

Arctic glaciers are undergoing fast retreat, exposing soils that have been locked under ice for thousands of years (Bradley, Singarayer and Anesio, 2014). Microbial communities have been identified as the primary colonisers of these newly exposed soils (Schmidt *et al.,* 2008) and are important for building up initial carbon and nitrogen pools, enhancing soil stability through the release of exopolymeric substances, and mediating forefield soil pH (Sattin *et al.,* 2009; Schulz *et al.,* 2013). However, there is a lack of coherent understanding on the diversity and biogeochemical importance of these bacterial communities in relation to nitrogen fixation (Brankatschk *et al.,*2011). Bacterial nitrogen fixation uses the enzyme nitrogenase to convert atmospheric nitrogen ($N_2$) into fixed ammonia ($NH_3$) for biological uptake by non-diazotrophic organisms (Brill 1975). As nitrogen is a key nutrient for microbe and plant growth, nitrogen limited forefield soils may place restrictions on heterotroph colonisation, productivity and succession (Duc *et al.,* 2009). Subsequently, diazotrohic organisms have been proposed as crucial facilitators of succession in newly exposed forefield soils (Knelman *et al.,* 2012). Nitrogen-fixing cyanobacteria have been identified as key in building these initial nitrogen stocks, and therefore expediting the establishment of heterotrophic organisms (Kaštovská *et al.,* 2005; Schmidt *et al.,* 2008; Duc *et al.,* 2009).

Whilst the importance of early diazotrophs is evident, similarities and variations in the nitrogen-fixing communities across forefields, in terms of both diversity and phylogeny, have received limited attention. The majority of research to date has focused on understanding changes in nitrogen fixation within individual forefields, along transects or chronosequences of soil development (Duc *et al.,* 2009; Brankatschk *et al.,* 2011). Thus far, the taxonomic diversity and abundance of the nifH gene, encoding nitrogenase for nitrogen fixation, has been shown to decrease with soil age and distance from the glacier terminus, in line with increasing fixed nitrogen in soils, and a reduced need for diazotrophy (Duc *et al.,* 2009; Brankatschk *et al.,* 2011). The dominant diazotrohic community composition in forefields is likely to be influenced by factors such as soil physicochemical status, climate, topography, the establishment of plants and any disturbances, such as water flow pathways, which may elicit both similarities and differences in diazotrophy between sites (Hodkinson, et al., 2002; Nicol *et al.,* 2005; Schütte *et al.,* 2010; Liu *et al.,* 2012). Furthermore, the current body of evidence surrounding microbial succession in forefields has a limited geographical range, with most studies conducted in the Damma Glacier forefield in Switzerland (Duc *et al.,* 2009; Frey *et al.,* 2010;

Bernasconi *et al.,* 2011; Brankatschk *et al.,* 2011; Brunner *et al.,* 2011; Zumsteg *et al.,* 2013, 2013; Bradley *et al*., 2015). Investigation across multiple glacier forefields is needed to fully explore similarities and differences between forefields in terms of diazotrophic community composition and their phylogenetic relations (Schütte *et al.,* 2010). This will help highlight the microbial community diversity involved in nitrogen fixation among glacier forefields.

Bacterial nitrogen fixation is encoded by clustered nitrogenase (nif) genes, typically through an enzyme containing an iron (Fe) cofactor and a molybdenum-iron (Mo-Fe) cofactor (Dixon and Kahn, 2004). Overall, the abundance of bioavailable nitrogen controls the transcription of nitrogenase genes, whilst the variant of nitrogenase transcribed is regulated by the presence of molybdenum (Oda *et al.,* 2005; Teixeira *et al.,* 2008). In the absence of Mo, nitrogenase is transcribed with vanadium (Fe-V co-factor), or exclusively with iron (Fe-Fe cofactor) in the absence of both Mo and V (Raymond *et al.,* 2004; Teixeira *et al.,* 2008). These nitrogenases are in turn encoded by the nifHDK, vnfH-vnfDGK and anfHDGK operons (Dixon and Kahn 2004; Teixeira *et al.,* 2008). The phylogenetically conserved nifH gene can be used to classify bacterial diazotrophs into Clusters I-IV based on the nitrogenase (Chien and Zinder 1996). Cluster I covers the typical Mo nifH, whilst Cluster II covers the alternative vnfH and Cluster III generally includes a diverse range of anaerobic bacteria (Zehr *et al.,* 2003). Furthermore, Cluster IV contains organisms with ‘nif-like’ sequences, as opposed to conventional nif genes (Zehr *et al.,* 2003).

Previous research conducted on microbial succession in glacial forefields, including those on functional genes, has mostly focused on marker gene data, such as the universal bacterial marker 16s rRNA and amplified nifH (Schmidt *et al.,* 2008; Brankatschk *et al.,* 2011; Rime *et al.,* 2015). However, studies are now applying alternative methods, such as metagenomics, to study microbial communities (Wooley, Godzik and Friedberg, 2010). This is because metagenomics can provide gene sequences for the entire microbial community gene pool, rather than target sequences (Handelsman 2004; Daniel 2005). Thus, both microbial diversity and functional potential can be inferred using one approach (Wooley, Godzik and Friedberg 2010; Thomas *et al.,* 2012). To maximise the quality of the output metagenome, the short DNA fragments from next generation sequencing can be assembled (Vázquez-Castellanos *et al.*, 2014). This generates longer continuous DNA reads (contigs), which can provide enhanced functional and taxonomic annotations (Howe *et al.,* 2014; Vázquez-Castellanos *et al.,* 2014).

In this study, we investigated 70 soil metagenomes spanning transects and chronosequences across four Arctic forefields in N-Sweden, Greenland and Svalbard. The datasets have been assembled separately and subsequently annotated for use in a comparative metagenomics analysis. This study leads on from Chapter 2, by utilising metagenome assembly with metaSPAdes in an environmental setting. Here, we use metagenomics to present an investigation into the taxonomy and phylogenetic relationships of the functional genes recovered relating to bacterial nitrogen fixation in the four forefields. This analysis aims to contribute to the existing knowledge on pioneer microbial communities, helping to identify key genera of diazotrophic bacteria, which may have a key role building labile nitrogen stocks and soil development in oligotrophic forefield soils.

## 3.2 Materials and methods

### 3.2.1 Field sampling

Four Arctic glacier forefields were selected for sampling and analysis, in front of Rabots glacier (Rb), N-Sweden (67° 54′ 25.6284″ N, 18° 26′ 51.0792″ E); Storglaciären (St), N-Sweden (67° 52′ 21.1116″ N, 18° 34′ 2.676″ E); Midtre Lovénbreen (Ml), Svalbard (79° 6′ 1.8″ N, 12° 9′ 21.996″ E) and Russell Glacier (Rl), Greenland (67° 9′ 23.4324″ N, 50° 3′ 50.342″ W). Samples were obtained in July 2013 (Midtre Lovénbreen) and July 2014 (Russell, Rabots and Storglaciären). Surface soil from each site was sampled using a chronosequence/transect-based approach, constructing three parallel transects along the forefield moving away from the terminus (Bradley, Singarayer and Anesio 2014). Chronosequence-based sampling was used to capture the diversity in nutrient concentration and microbial taxonomy of each forefield, to make more holistic comparisons between glacial forefields. Bulk surface samples were collected into sterile Whirlpak bags, and frozen at −20°C. Observationally, the sites comprised soils at very different development stages. A ‘typical’ smooth successional chronosequence from bare ground, to more developed, plant colonised soil was observed in the Ml forefield. However, the other sites sampled had a more heterogeneous chronosequence, with earlier and often more patchy plant colonisation.

### 3.2.2 Soil organic carbon and total nitrogen content

Soil total nitrogen (TN) and total organic carbon (TOC) were determined using mass spectrometry on a FlastEA 1112 nitrogen and carbon elemental analyser. The protocol described in Hedges and Stern (1984) was used for sample preparation. In brief, for TN analysis soil samples were weighed and dried at 50°C overnight, before subsamples were transferred into tin capsules. For TOC analysis, 2 ml of 1 M HCL was incrementally added to 0.1 g of sample (Wo) until effervescence stopped. Subsequently samples were again dried overnight at 50°C, left to equilibrate with hydroscopic salts, and re-weighed (Wf). Finally, subsamples were transferred into tin vials for analysis. The percentage of TOC in each sample was calculated using a correction for acidification induced weight change (Equation 2.1). Where possible, three environmental replicates were analysed for each TN and TOC per sampling site.

***Equation 2.1:*** *Correction for weight change during acidification of samples for organic carbon elemental analysis.*

$$\%OC = \left[\frac{100 \text{ x mgOC}}{\text{mg sample}}\right] x \left[\frac{Wf}{\text{Wo}}\right]$$

Where Wo is sediment dry weight prior to acidification, and Wf is dry weight after acidification.

3.2.3 DNA extraction, library preparation and sequencing

As this study was focused on the microbial diversity in bulk surface soil, DNA was extracted using a Mo-Bio DNAEasy PowerSoil DNA extraction kit (QIAGEN, UK), with DNA yield quantified using a Qubit 2.0 fluorometer. Samples that yielded less than 50 ng of DNA during extractions were pooled with their field replicates prior to sequencing. This method has been previously shown to obtain high DNA yields from soils and has been used for soil microbial diversity analysis in a number of studies, including root microbiomes (Fierer *et al.,* 2007; Allison *et al.,* 2008; İnceoğlu *et al.,* 2010; Carvalhais *et al.,* 2013; Vishnivetskaya *et al.,* 2014). However, as this approach is not directly targeting the soil rhizosphere communities, there may be limitations to DNA extraction from this subset of the microbial community. Metagenomes were sequenced using an Illumina Next-Seq 500 (Rb, St and Rl) and an Illumina-Mi Seq (Ml), with a TruSeq library prep kit at the University of Bristol Genomics facility. A total of 70 metagenomes were sequenced across the four sites using 2x 150bp (Rb, ST, Rl) and 2x 100bp (Ml) paired-end reads (Appendix 2 Table 1). Sequencing read output for each site can be identified in Appendix 2 Table 2, ranging between 3 817 852 and 10 510 0186 reads per metagenome.

### 3.2.4 Metagenome assembly and annotation

The 70 sequenced datasets were quality trimmed and subsequently assembled individually using the metaSPAdes 3.10.0-dev assembler, a development release of metaSPAdes, tested in Chapter 2 (Bankevich *et al.,* 2012). These assemblies were carried out in collaboration with the DOE Joint Genome Institute (Walnut Creek, CA), using the BFC algorithm for read error correction (Li 2015), and the --meta and --only-assembler flags. Furthermore, incremental Kmer lengths were used (22, 33, 55 and 77) to identify the most appropriate parameter value for assembly. Assembly size for each metagenome ranged between 241,660 and 429,543,524 bases (Appendix 2 Table 2). Functional annotation of the 70 metagenomes was subsequently carried out using the Integrated Microbial Genomes with Microbiome Samples (IMG/M) system (Chen *et al.,* 2017). Rarefaction curves were created in MG-RAST 4.0.3 for each metagenome (Appendix 2 Figures 1–4; Meyer *et al.,* 2008). Each metagenome was evaluated based on the number of contigs assembled and species obtained, to highlight metagenomes that may be under sampled through sequencing. Under sampling can occur in highly diverse metagenomes, where the sequencing is not adequate to reveal all taxa present in the sample (Torsvik and Øvreås 2002). Consequently, in under sampled datasets, some organisms, particularly those which were less abundant, may not be included in the output metagenome (Rodriguez and Konstantinidis 2014).

For each metagenome, the nifH gene for nitrogen fixation was searched using the Basic Local Alignment Search Tool for Proteins (BLAST-p) with an e-value of 1e$^{-5}$and extracted. As nifH genes are generally found in a phylogenetically conserved nitrogenase cluster (with nif D, K, N and E), these genes were also searched for and extracted (Howard and Rees 1996). Nif genes were dereplicated, removing duplicate copies, using VSEARCH 2.6.0, leaving a total of 185 assembled nif genes for subsequent analysis (Rognes *et al.,* 2016). The nif genes used for the analysis have been deposited in GenBank, under accession numbers MH551286 - MH551470. Gene abundance was calculated as a combined value of nifHDKNE, normalised in relation to the abundance of the bacterial single copy housekeeping gene, rpoB, for each site (Vos *et al.,* 2012; Ishii *et al.,* 2015). As this method relies on sequencing unamplified genes, the nif gene counts are limited and may not be exhaustive for individual samples. This is particularly the case for unamplified sequencing of complex microbiome datasets, such as soil samples (Rodriguez and Konstantinidis 2014). Additionally, diazotrophs can contain multiple different nif genes, and several copies of a single variant, so should not be used as a measure to enumerate the explicit number of diazotrophs in each sample (Zehr *et al.,* 2003).

Finally, the raw sequencing reads were mapped to the extracted nif contigs for each metagenome using the BWA-MEM algorithm (Li and Durbin 2009). The alignment score (AS) of each read/contig is reported, which numerically indicates the quality of the alignments.

### 3.2.5 Nif taxonomy

The taxonomic distribution of all nif sequences (HDKNE) was carried out using a Last Common Ancestor (LCA) analysis in MEGAN 6.9.0 (Huson *et al.,* 2016). For each forefield, nifHDKNE sequences were nucleotide BLAST (BLASTn) searched against an NCBI GenBank database of complete bacterial genomes. The sequences were subsequently binned based on the NCBI taxonomy, using an LCA algorithm, and visualised at the genus level for each forefield (Huson *et al.,* 2016).

### 3.2.6 Gene phylogeny

A phylogeny for nifH, based on clusters identified in Zehr *et al*., (2003), was carried out, as this gene is supported by the largest body of research. Sample nifH sequences were aligned to sequences of cultured isolates, largely derived from the phylogeny by Deslippe and Egger (2006). GenBank and UniProtKB accession numbers for cultured isolates are available in Table 3.1. DNA sequence alignments were generated in SATé 2.2.7, using MAFT, MUSCLE and FASTTREE (Liu *et al.,* 2011). The GTR+CAT model was implemented, with the decomposition set to longest (to minimise long branch attraction) and a maximum number of iterations set to 8. Alignments were manually edited in Mesquite, alongside generating Nexus and Phylip format files (Maddison and Maddison 2017). Maximum likelihood phylogenies were carried out using the CIPRES implementation of RAXML-HPC2 8.2.10 on XSEDE[1], with 1000 bootstrap iterations (Stamatakis, 2014). The GTR+G model of nucleotide substitutions was implemented, as identified with j model test (Guindon and Gascuel 2003; Darriba *et al.,* 2012). Trees were evaluated using Figtree 1.4.3[2], before annotation with EvolView v2[3] (He *et al.,* 2016). Graphical enhancements were made using Inkscape 0.92.2[4]. Comparisons between nifH sample sequences and cultured isolates were made using NCBI BLASTn[5], to identify nearest cultured relatives.

***Table 3.1:*** *GenBank and UniProtKB accession numbers for nifH sequences derived from Deslippe and Egger (2006), for use in the nifH phylogeny.*

| **Database** | **Accession number** | **Species** | **Gene** |
|---|---|---|---|
| GenBank | X13519.1 | *Azotobacter vinelandii* | vnfH |
| GenBank | AY367395.1 | *Kiebsiella variicola* | nifH |
| GenBank | AF484674.1 | *Methylomonas rubra* | nifH |
| GenBank | AF216883.1 | *Azomonas agilis* | nifH |
| GenBank | V01215.1 | *Rhizobium meliloti* | nifH |
| GenBank | U97122.1 | *Azoarcus tolulyticus* | nifH |
| UniProtKB | P26251 | *Azorhizobium caulinodans* | nifH |
| GenBank | AJ515294.1 | *Paenibacillus azotofixans* | nifH |
| GenBank | Z31716.1 | *Nostoc* sp. | nifH |
| UniProtKB | P33178 | *Anabaena* sp. | nifH |
| UniProtKB | P08925 | *Frankia alni* | nifH |
| GenBank | X57006.1 | *Frankia* sp. | nifH |
| GenBank | ABQ25379.1 | *Geobacter uraniireducens* | nifH |
| GenBank | M23528.1 | *Azotobacter vinelandii* | anfH |
| UniProtKB | P16269 | *Azotobacter vinelandii* | anfH |
| UniProtKB | Q07942 | *Azotobacter capsulatus* | anfH |
| GenBank | AF065617.1 | *Chlorobium tepidum* | anfH |
| GenBank | AY221832.1 | *Pelodictyon lutolum* CC11OA0 | anfH |
| GenBank | AF227926.1 | *Desulfovibrio salexigens* | nifH |
| GenBank | AY040513.1 | *Desulfomicrobium baculatum* | anfH |
| UniProtKB | P25767 | *Methanococcus thermolithotrophicus* | nifH |
| GenBank | AY029234.1 | *Methanosarcina mazei* | nifH |
| UniProtKB | P00456 | *Clostridium pasteurianum* | nifH |
| GenBank | AF065618.1 | *Desulfonema limicola* | anfH |
| GenBank | AF216881.1 | *Acetobacterium woodii* | anfH |

## 3.3 Results and Discussion

### 3.3.1 Soil carbon and nitrogen

***Table 3.2:*** *Summary statistics for total nitrogen (TN) and total organic carbon (TOC) across the four forefields (Midtre Lovénbreen Ml, Russell Rl, Storglaciären St and Rabots Rb). The average, minimum, maximum and standard deviation (SD) across each forefield is given. The detection limit for both TN and TOC was 1 mg g$^{-1}$. Sites recording values below detection (b.d) are shown.*

| **TN (mg g$^{-1}$)** | **Average** | **Minimum** | **Maximum** | **SD** |
|---|---|---|---|---|
| **Ml** | b.d. | b.d | 4.90 | 1.56 |
| **Rl** | 1.95 | b.d | 6.94 | 2.15 |
| **St** | b.d. | b.d | 4.19 | 0.93 |
| **Rb** | 1.04 | b.d | 3.35 | 1.33 |

| **TOC (mg g$^{-1}$)** | **Average** | **Minimum** | **Maximum** | **SD** |
|---|---|---|---|---|
| **Ml** | 10.56 | b.d. | 72.36 | 21.14 |
| **Rl** | 26.36 | b.d. | 82.70 | 26.35 |
| **St** | 2.78 | b.d. | 27.89 | 6.25 |
| **Rb** | 6.81 | b.d. | 22.90 | 9.66 |

The range of values obtained within and between forefields for TOC and TN for samples from each forefield is listed in Table 3.2. These values include TOC and TN from both microbial and plant sources. Looking at average nutrient contents, comparing across the forefields, TN content ranges from averages below detection to 1.95 mg g$^{-1}$, between St and Rl, respectively (Table 3.2). TOC content follows the same trend, increasing from the two Swedish glaciers (St and Rb), to Ml and Rl. Results from a one-way ANOVA analysis for each nutrient did not show any statistically significant differences in the TN measured between forefields ($P > 0.05$). However, concentrations of TOC were found to vary significantly ($P$ = 0.002) (Table 3.3). Additional analysis of the TOC variance between forefields using a post-hoc Tukey analysis revealed the significant difference was between the St and Rl forefields, with Rl containing almost 10 times the TOC content of St on average ($P < 0.01$, Table 3.2; Table 3.4).

***Table 3.3*** *ANOVA comparing differences between the four forefields (Midtre Lovénbreen, Russell, Rabots and Storglaciären) based on total nitrogen (TN) and total organic carbon (TOC). Significant differences observed between the forefields are noted at the 0.01 or 0.05 level.*

| | TN | TOC |
|---|---|---|
| **f-ratio value** | 2.46 | 5.375 |
| **p-value** | 0.071 | 0.002357 |
| **Significance level** | Not significant | 0.05 |

***Table 3.4:*** *Results of a post-Hoc Tukey analysis, comparing differences between forefields (Midtre Lovénbreen Ml, Russell Rl, Storglaciären St and Rabots Rb), based on total nitrogen (TN) and total organic carbon (TOC). Significant differences between the forefields are noted at the 0.01 significance level.*

| TN | Treatments | Q statistic | p-value | Inference |
|---|---|---|---|---|
| | **Ml vs Rl** | 2.65 | 0.25 | insignificant |
| | **Ml vs Rb** | 0.14 | 0.89 | insignificant |
| | **Ml vs St** | 0.89 | 0.89 | insignificant |
| | **Rl vs Rb** | 1.54 | 0.68 | insignificant |
| | **Rl vs St** | 3.64 | 0.05 | insignificant |
| | **Rb vs St** | 0.72 | 0.89 | insignificant |
| **TOC** | **Treatments** | **Q statistic** | **p-value** | **Inference** |
| | **Ml vs Rl** | 3.56 | 0.07 | insignificant |
| | **Ml vs Rb** | 0.52 | 0.89 | insignificant |
| | **Ml vs St** | 1.66 | 0.62 | insignificant |
| | **Rl vs Rb** | 2.79 | 0.21 | insignificant |
| | **Rl vs St** | 5.39 | 0.00 | ** p<0.01 |
| | **Rb vs St** | 0.56 | 0.89 | insignificant |

Samples from the Rl forefield revealed the widest range in both TOC (below detection—82.70mg $g^{-1}$) and TN (below detection—6.94mg $g^{-1}$), respectively (Table 3.2). This contrasts with the Rb forefield, where TOC and TN values expressed a smaller range, from below detection to 22.90 mg $g^{-1}$ and below detection up to 3.35 mg $g^{-1}$, respectively. A range of values is expected across sites within each forefield, due to soil development which takes place over successional chronosequences and given variations in sources of autochthonous and allochthonous material (Bradley, Singarayer and Anesio 2014), for example, in the deposition of aeolian material (such as soot), or the presence of ancient *in situ* organic pools, exposed by glacier retreat (Schulz *et al.,* 2013; Bradley *et al.,* 2015). For example, across the MI chronosequence TN and TOC increase from below detection and 2.85 mg g$^{-1}$, to 4.4 mg g$^{-1}$ and 14.5 mg g$^{-1}$, in line with expected soil development (Table 3.5; Bradley *et al.,* 2016). However, whilst differences in soil nutrient content do occur between sites, the values fall into the general range observed from other forefields (1–2 mg g$^{-1}$ nitrogen, and 0.1–40 mg g$^{-1}$ carbon) (Bradley, Singarayer and Anesio 2014) and are indicative of a generally oligotrophic environment.

***Table 3.5****: Initial and final concentrations of total nitrogen (TN) and total organic carbon (TOC) in forefield soils. Values are shown at the start of the transect/ chronosequence (by the glacier terminus) and at the end of the transect.*

| | **Start TN (mg g$^{-1}$)** | **End TN (mg g-1)** | **Start TOC (mg g$^{-1}$)** | **End TN (mg g$^{-1}$)** |
|---|---|---|---|---|
| **MI** | b.d | 4.40 | 2.85 | 14.47 |
| **St** | b.d | b.d | b.d | 1.24 |
| **Rb** | b.d | 1.74 | b.d | b.d |
| **RI** | b.d | 1.33 | b.d | 17.46 |

### 3.3.2 Rarefaction analysis

Rarefaction analysis was used to investigate the coverage of diversity in each metagenome, identifying any datasets where species content may be under sampled (Appendix 2 Figures 1-4). For each forefield, an assortment of both adequately sequenced and under sampled metagenomes were obtained (Figure 1-4, Appendix 2). Metagenomes that show rarefaction curves to reach saturation are likely to adequately profile the microbial diversity in the samples, for example metagenomes MI 7, RI 15, St 16 and St 17 (Figures 1, 2 and 3, Appendix 2). However, those metagenomes in which species number does not reach saturation are most likely to exclude taxa, for example ML1, ML 20, RI 14 and RI 20 (Figure 1 and 2, Appendix 2). In these metagenomes, the least abundant taxa are most probably excluded from the dataset, due to the reduced abundance of DNA for sequencing from these organisms (Rodriguez and Konstantinidis 2014). Whilst this does not detract from conclusions drawn on the organisms present in the samples, the full depth of diversity in under-sampled metagenomes cannot be highlighted. This issue is often prevalent in highly complex datasets such as soil and can only be resolved through continued deeper sequencing of those metagenomes (Rodriguez and Konstantinidis 2014).

### 3.3.3 Nif genes recovered

The total abundance of dereplicated *rpoB* normalised contigs containing nif genes (nifHDKNE), in relation to the variation of TN and TOC, spanning all sampling sites is shown in Figure 3.1. A total of 185 nif genes contained on assembled contigs were recovered from the datasets. In 75% of samples where nif genes were detected, the TN and TOC concentrations fell below 1 and 5 mg $g^{-1}$, respectively (Figure 3.1). Conversely, in samples where nif genes were not detected, 61% and 49% measured below 1 and 5 mg $g^{-1}$, of TN and TOC, respectively (Figure 3.1). As sequencing output varied substantially between metagenomes, further sequencing may reveal additional genes due to the complex nature of soil microbiome samples (Table 2, Appendix 2; Rodriguez and Konstantinos 2014). However, this may indicate that samples with limited TN/TOC could have a larger relative abundance of genes for diazotrophy, as these were recovered through the sequencing effort undertaken. Interestingly, a similar trend between nitrogen fixation and TN has been reflected by the assays carried out by Telling *et al*., (2011), whereby fixation rates on Arctic glaciers were negatively correlated with total inorganic nitrogen content. Additionally, a link between nif gene abundance and activity is supported theoretically, as fixation becomes less metabolically beneficial when labile nitrogen stocks increase (Gutschink *et al,.*1978). When applied to forefield soils, both TN and TOC have been shown to increase over successional chronosequences, indicating nitrogen fixation may become less profitable with soil development (Duc *et al.,* 2009; Brankatschk *et al.,* 2011; Bradley, Singarayer and Anesio 2014). Furthermore, research by Brankatschk *et al*., (2011) identified a link between nif gene abundance and enzyme activity, indicating sites with high numbers of nif genes, such as Storglaciären, would have enhanced nitrogen fixation activities. However, the relationship between gene abundance and nitrogen fixation activity is not always fully defined, as areas with low nitrogenase activity have previously been linked to high gene abundance in the Damma Glacier (Swiss Alps) (Duc *et al.,* 2009).

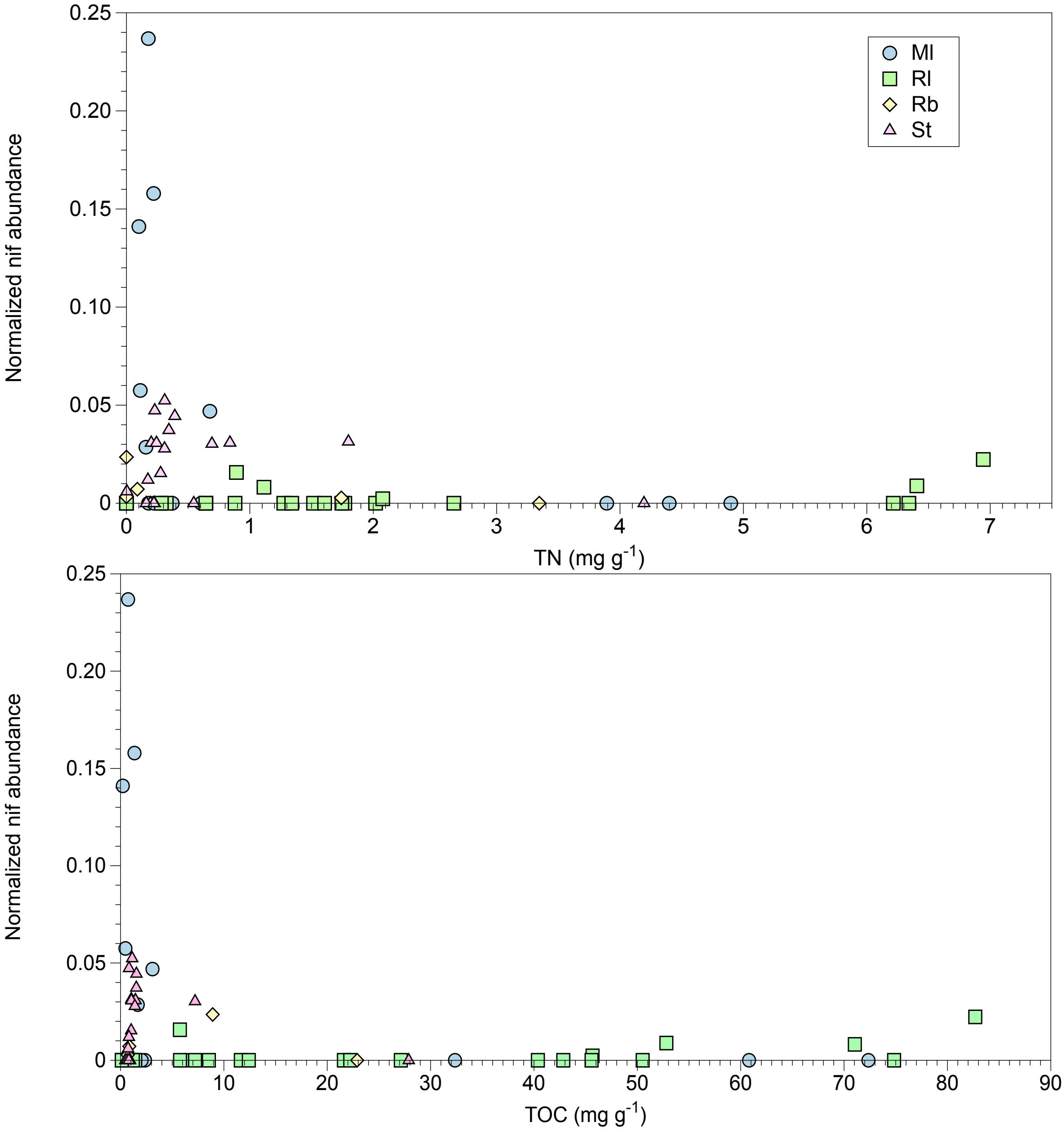


***Figure 3.1:*** *Relationship between normalized nif gene abundance (nifHDKNE) and concentration of total organic carbon (TOC) and total nitrogen (TN) per gram of soil, across all sampling sites. Nif gene abundance values are normalized against the bacterial single copy housekeeping gene, rpoB, for each metagenome. Values across the different forefields are noted, including: Midtre Lovénbreen (Ml), Russell (Rl), Rabots (Rb) and Storglaciären (St).*

The results of mapping sequencing reads to the nif genes is provided in Table 3 (Appendix 2). This highlights the Alignment Score (AS), which indicates the alignment quality between reads and contigs (Table 3, Appendix 2). The number of nif genes for each score threshold is provided, alongside the percentage of reads with AS over 60. The Alignment Score ranges between 0 and the maximum length of the reads (0–100 for Ml dataset and 0–150 for Rb, St and Rl datasets). For each forefield, the percentage of alignments with an AS greater than 60 was $1.06x10^{-3}$ (Ml), $4.23x10^{-5}$ (Rl), $2.38x10^{-4}$ (Rb) and $9.56x10^{-4}$ (St). Plots of the normalised nif genes recovered and the number of reads aligning to genes with an AS over 60, for each metagenome, are available in Figure 5–8 (Appendix 2).

### 3.3.4 Nitrogenase clusters

Our newly sampled bacteria were analysed and grouped with previously published relatives, as shown in Zehr *et al*., (2003). Forefield sequences were distributed across Cluster I (23 sample sequences) and III (3 sample sequences), with no representatives in Cluster II or IV (Figure 3.2). Thus, 88.5% of sample sequences were attributed to Cluster I, which contains the typical Mo nifH, indicating the presence of plentiful molybdenum in soils for the nitrogenase cofactor (Zehr *et al.,* 2003).

Environmental samples in Cluster I included the groups Alphaproteobacteria, Betaproteobacteria, Cyanobacteria and Firmicutes (Figure 3.2). The first group, associated with Alphaproteobacteria and Betaproteobacteria, incorporated five environmental samples that clustered most closely with *Azorhizobium caulinodans* and *Azoarcus tolulyticus*. These are plant-associated diazotrophs, important for establishing stocks of fixed nitrogen for legume uptake, supporting plant growth (Hurek and Hurek 1995; Dreyfus, Garcia and Gillis 1988). The second group was comprised of six sample sequences, clustering with the Cyanobacteria, *Nostoc* and *Anabaena*, which are free living nitrogen fixers (Zehr *et al.,* 2003). Cyanobacteria have been proposed as crucial for building labile nitrogen pools in newly exposed soils, important for facilitating heterotroph colonisation, and have been identified in other forefields using 16s rRNA amplicon sequencing (Schmidt *et al.,* 2008; Duc *et al.,* 2009; Frey *et al.,* 2013). Group 3 contained 11 highly related sample nifH sequences, grouping closely to *Frankia*. This genus is composed of nitrogen-fixing bacteria that are symbionts of actinorhizal plant roots, and again provides evidence for bacterial support of plant growth and establishment, through supplies of fixed nitrogen (Benson and Silvester 1993). Whilst the forefields may have a low diversity of root symbiotic diazotrophs, this may also relate to sub-optimal cell lysis and separation of root-associated cells during the DNA extraction process, or that these organisms were at a low abundance and thus not captured through sequencing.

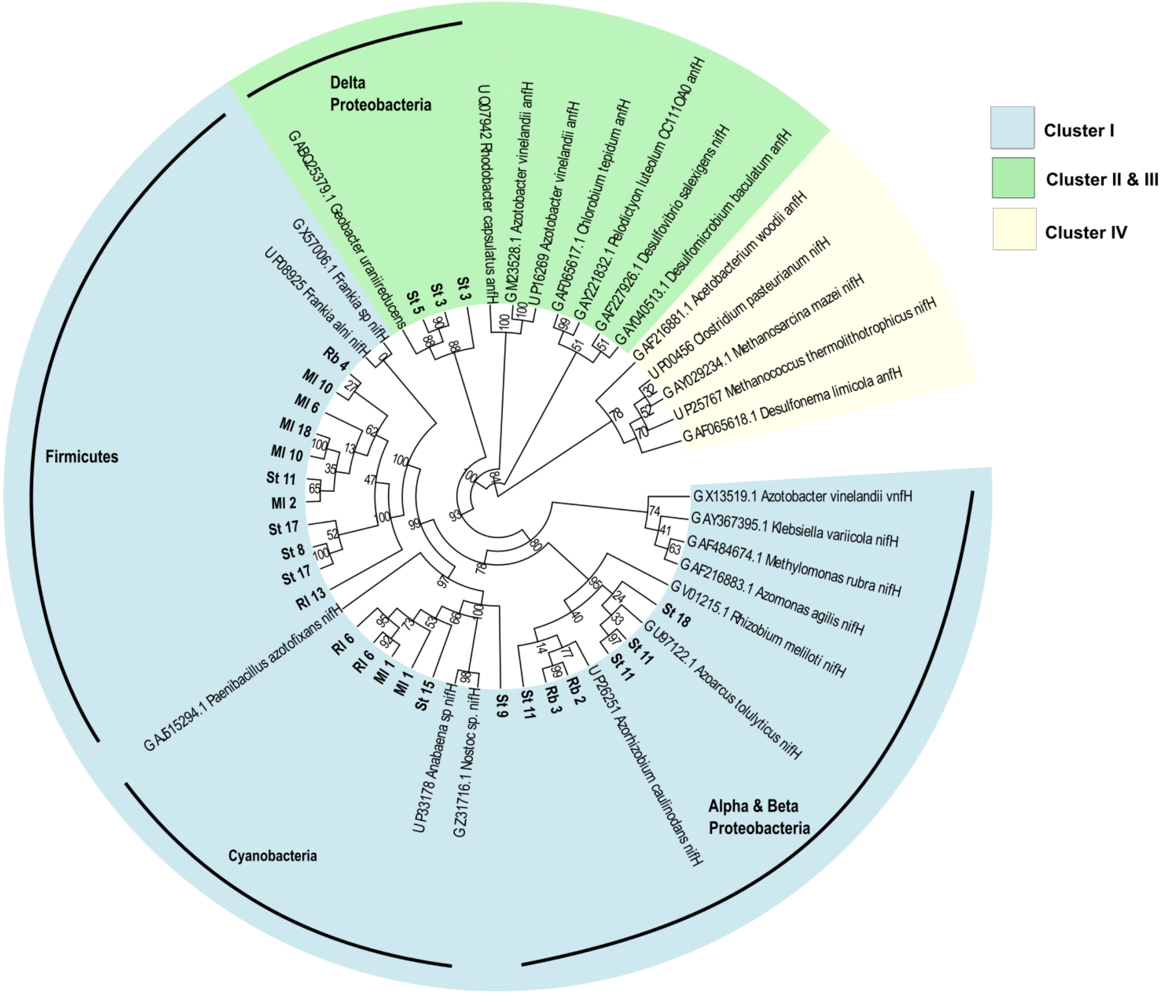

Cluster I
Cluster II & III
Cluster IV
Delta
Proteobacteria
Firmicutes
Cyanobacteria
Alpha & Beta
Proteobacteria
G ABQ25379.1 Geobacter uraniireducens
St 5
St 3
St 3
U Q07942 Rhodobacter capsulatus anfH
G M23528.1 Azotobacter vinelandii anfH
U P16269 Azotobacter vinelandii anfH
G AF065617.1 Chlorobium tepidum anfH
G AY221832.1 Pelodictyon luteolum CC111OA0 anfH
G AF227926.1 Desulfovibrio salexigens nifH
G AY040513.1 Desulfomicrobium baculatum anfH
G AF216881.1 Acetobacterium woodii anfH
U P00456 Clostridium pasteurianum nifH
G AY029234.1 Methanosarcina mazei nifH
U P25767 Methanococcus thermolithotrophicus nifH
G AF065618.1 Desulfonema limicola anfH
G X13519.1 Azotobacter vinelandii vnfH
G AY367395.1 Klebsiella variicola nifH
G AF484674.1 Methylomonas rubra nifH
G AF216883.1 Azomonas agilis nifH
G V01215.1 Rhizobium meliloti nifH
St 18
G U97122.1 Azoarcus tolulyticus nifH
St 11
St 11
U P26251 Azorhizobium caulinodans nifH
Rb 2
Rb 3
St 11
St 9
G Z31716.1 Nostoc sp. nifH
U P33178 Anabaena sp nifH
St 15
Ml 1
Ml 1
Rl 6
Rl 6
G AJ515294.1 Paenibacillus azotofixans nifH
Rl 13
St 17
St 8
St 17
Ml 2
St 11
Ml 10
Ml 18
Ml 6
Ml 10
Rb 4
U P08925 Frankia alni nifH
G X57006.1 Frankia sp nifH
0
88
90
88
100
100
99
51
51
78
32
52
70
84
100
93
74
41
63
80
95
24
33
97
40
77
99
14
78
100
98
66
53
73
95
92
97
99
100
52
100
47
65
35
100
13
62
27
100

***Figure 3.2:** nifH maximum likelihood phylogeny of sample sequences (bold) and sequenced samples derived from NCBI GenBank and UniProtKB. Most sample sequences were obtained from the nifH phylogeny of Deslippe and Egger, (2006). For study samples, the Sample ID is given, corresponding to Appendix 2 Table 1. For sequenced samples, the database, organism name and gene are given. Bootstrap support values are given, based on 1000 tree iterations. The nifH clusters (derived from Zehr et al., 2003) are denoted by leaf colours (Cluster I-IV). The tree is rooted on Cluster IV, as this group contains divergent 'nif-like' sequences (Zehr et al., 2003). Key groups containing sample sequences are noted, including Firmicutes, Cyanobacteria, Alphaproteobacteria, Betaproteobacteria and Deltaproteobacteria.*

Environmental samples were also present in Cluster III, which is attributed to a group of anaerobic bacteria (Zehr *et al.,* 2003). The three sample sequences clustered most closely to *Geobacter uraniireducens*, an anaerobe common in sediments under metal reducing conditions, capable of dissimilatory Fe(III) reduction (Shelobolina *et al.,* 2008). However, no sample sequences were linked to Cluster II, which is associated with organisms containing the alternative anfH, containing an Fe–Fe cofactor, used in the absence of molybdenum and Vanadium (Zehr *et al.,* 2003).

These results reflect those of Duc *et al*., (2009), who used clone libraries to evaluate the phylogeny of diazotrophs across the Damma Glacier, Switzerland. Interestingly, nifH sequences from their analysis also grouped with nitrogenase Clusters I and III (Duc *et al.,* 2009). Additionally, genera identified by Duc *et al*., (2009) included the key genera identified in this analysis, such as *Geobacter, Nostoc* and *Anabaena*, suggesting that these organisms are common across forefields (Duc *et al.,* 2009). The prevalence of these organisms may be due to adaptations or attributes to cold environments, such as cold or UV tolerance, and the release of protective exudates (Tamaru *et al.,* 2005; Chattopadhyay 2006; Pattanaik *et al.,* 2007). Cyanobacteria such as *Nostoc* have been shown to produce extracellular polysaccharides (EPS) which are important for desiccation and freeze-thaw tolerance in Arctic environments (Tamaru *et al, .*2005). *Geobacter* are commonly found in anaerobic environments, and therefore may tolerate any anoxia in forefield soils created by frequent meltwater flooding and the formation of melt pools (Duc *et al.,* 2009). The consistent identification of *Geobacter*, *Nostoc* and *Frankia* in forefield soils using nifH analysis indicates that a core group of diazotrophs may be present across Arctic forefields. These diazotrophs may be important for facilitating plant colonisation and establishment, either by building labile pools in newly exposed soils (Cyanobacteria) or through symbiosis (*Frankia, Azorhizobium*).

Results from BLASTn searching each nifH sequence against cultured isolates revealed forefield sample sequences were divergent, with sequence identity ranging between 80%–95% (Table 3.6). This indicates that the diazotrophs present in the samples are novel compared to those which have been previously identified and may be unique or contain adaptations to cold oligotrophic forefield conditions. However, as less abundant organisms will contribute to a minor proportion of the unamplified sequenced DNA and nifH gene pool, using additional nif genes may help highlight the presence of rare organisms in samples (Cowan *et al.,* 2005). This may be especially helpful for metagenomes where sequencing coverage was not sufficient to profile the complete community structure, and thereby some low abundance organisms may not have been represented in the final dataset (Figure 1–4, Appendix 2).

***Table 3.6:*** *NCBI blastn matches for sample sequences, against cultured isolates. The best match accession number and % identity is given. Sequences with no significant matches have been left blank.*

| Sample | Cluster | Blast match | Accession number | % similarity |
|---|---|---|---|---|
| Rb10 | I | *Frankia* sp. | X57006.1 | 83 |
| Ml 10 | I | *Frankia* sp. | CP000820.1 | 84 |
| Ml 6 | I | *Frankia alni* str. | CT573213.2 | 82 |
| Ml 18 | I | - | - | - |
| Ml 10 | I | *Frankia casuarinae* strain | CP000249.1 | 82 |
| St 11 | I | *Frankia* sp. | AY115490.2 | 82 |
| Ml 2 | I | *Frankia alni* str. | CT573213.2 | 82 |
| St 17 | I | *Frankia* HRN18a | X17522.1 | 81 |
| St 8 | I | *Frankia* sp. | X73983.1 | 81 |
| St 17 | I | *Frankia* sp. | HM026362.1 | 81 |
| Rl13 | I | *Leptosprillum ferriphilum* | JN390678.1 | 85 |
| | | | | |
| St3 | III | *Geobacter uraniireducens* | CP000698.1 | 88 |
| St3 | III | *Geobacter lovleyi* | CP001089.1 | 90 |
| St5 | III | *Geobacter uraniireducens* | CP000698.1 | 89 |
| | | | | |
| Rl 6 | I | *Nostoc flagelliforme* | AP018269.1 | 95 |
| Rl 6 | I | *Scytonema* sp. | AP018268.1 | 91 |
| Ml1 | I | *Nostoc punctiforme* | CP001037.1 | 94 |
| Ml1 | I | *Scytonema* sp. | AP018268.1 | 90 |
| St15 | I | *Scytonema* sp. | AP018268.1 | 84 |
| St9 | I | *Anabaena variabilis* | AP018216.1 | 84 |
| | | | | |
| Rb2 | I | *Bradyrhizobium oligotrophicum* | AP012603.1 | 94 |
| Rb3 | I | *Bradyrhizobium oligotrophicum* | AP012603.1 | 91 |
| St 11 | I | *Polaromonas napthalenivorans* CJ2 | CP000529.1 | 87 |
| | | | | |
| St18 | I | *Bradyrhizobium oligotrophicum* S58 | AP012603.1 | 80 |
| St11 | I | *Polaramonas napthalenivorans* CJ2 | CP000529.1 | 94 |
| St11 | I | *Polaramonas napthalenivorans* CJ2 | CP000529.1 | 90 |

### 3.3.5 Diazotroph community structure

LCA analysis with multiple nif genes (HDKNE) identified the key organisms consistent between two or more forefields, including *Geobacter*, *Frankia* and *Nostoc*, which were also highlighted in the nifH analysis. Additional genera, for example *Polaromonas, Pelobacter and Microcoleus* were also identified here through the inclusion of additional nif genes (nifDKNE)

(Figure 3.3). This suggests including multiple nitrogenase genes provides a more holistic view of the diazotroph community structure in each forefield, due to the low copy number of these genes in unamplified samples. This is a particular issue of highly diverse metagenome samples, such as those from soils, as sequencing depth may not profile the complete community structure (Rodriguez and Konstantinidis 2014).

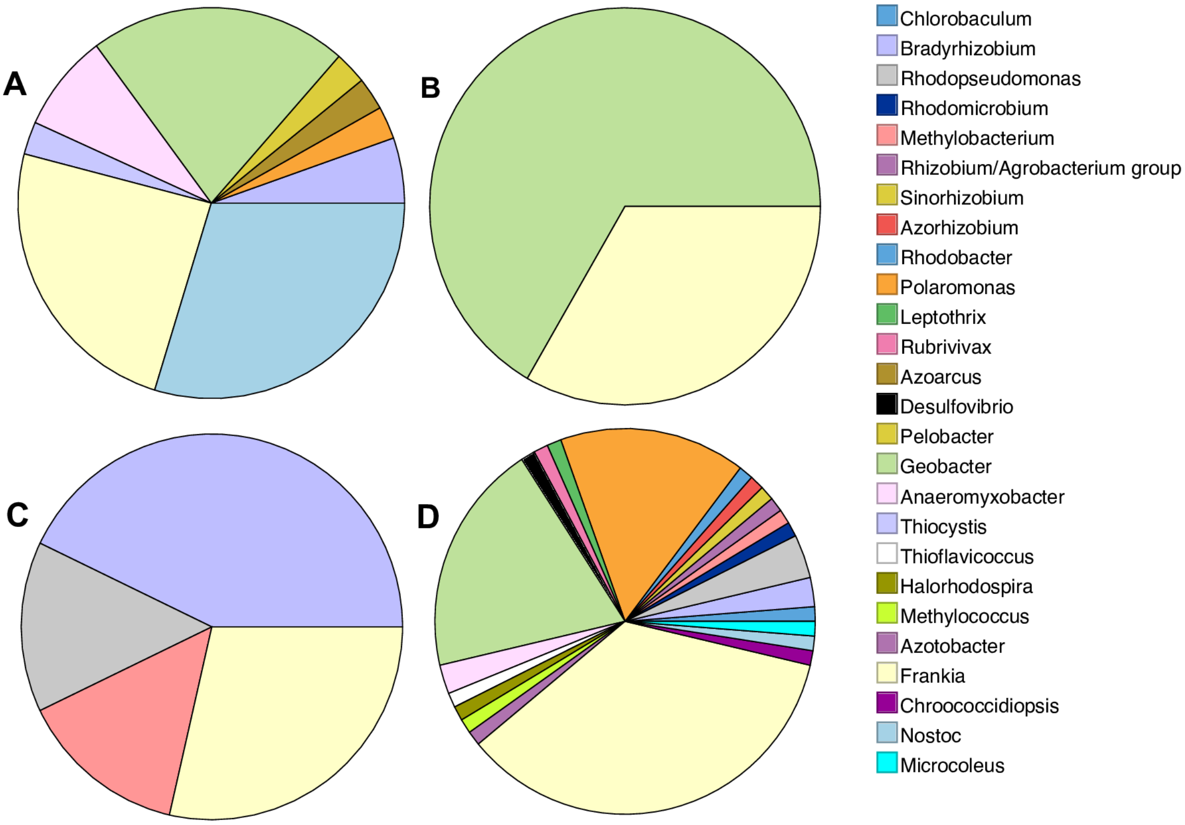


***Figure 3.3:*** *Taxonomic distribution of nif (HDKNE) genes for each forefield at the genus level: Midtre Lovénbreen Ml (A), Russell Rl (B), Rabots Rb (C), Storglaciären St (D). The total nif gene sequence count for each site was 42, 15, 13 and 91, respectively.*

The assignment of nif genes in the Rl forefield covers two key genera, *Geobacter* and *Frankia*. Limited research has been conducted into the presence of *Frankia* in Greenland; however, these organisms are typically associated with common actinorhizal plants (Benson and Silvester 1993; Chaia *et al.,* 2010). This group forms nitrogen-fixing root nodules with *Frankia* in exchange for reduced carbon and therefore are commonly found as early colonisers of undeveloped, oligotrophic soils (Wall 2000; Schwinter 2012). This is in agreement with the limited nitrogen content detected in this forefield, at 2.04 TN $g^{-1}$ (Figure

3.1 and Table 3.2). Additionally, the presence of plants has been identified as a key control on microbial community structure over the Damma Glacier forefield, Switzerland (Miniaci *et al.,* 2007). Furthermore, the identification of the anaerobic *Geobacter* indicates the presence of periodically saturated and anoxic conditions along the forefield, possibly attributed to meltwater flooding (Duc *et al.,* 2009). *Geobacter* are dissimilatory metal and sulfur reducing bacteria and have been proposed as key players in sediment nutrient cycles, oxidation of organic matter, bioremediation and soil gleying (Lovley *et al.,* 1993; Childers *et al.,* 2002; Methe *et al.,* 2003). *Geobacter* have been consistently identified across glacier forefield soils, which may relate to their metabolic diversity, thereby making these organisms well suited to fluctuating environmental conditions in forefield soils (Duc *et al.,* 2009; Edwards and Cook 2015; Rime *et al.,* 2015). This group has been shown to use chemotaxis to access Fe(III) oxides as an electron acceptor, which may explain their prevalence over other non-motile Fe(III) reducers (Hartmann and Brunner 2015). Whilst deeper sequencing in some metagenomes may highlight additional rare diazotrophic bacteria in Rl samples, it is likely that *Geobacter* and *Frankia* were the most dominant nitrogen fixers present, as these were identified through direct sequencing of unamplified DNA (Cowan *et al.,* 2005; Figure 2, Appendix 2).

Similarly, to Rl, the taxonomic diversity detected in the N-Swedish Rb forefield was largely comprised of root associated diazotrophs, including the genera *Bradyrhizobium*, *Frankia, Methylobacterium* and *Rhodopseudomonas* (Figure 3.3). This may relate to the lack of bare soil observed at this forefield, and therefore limited requirement for free living diazotrophs (Miniaci *et al.,* 2007). This site also had a low average soil nitrogen content, at 1.04 mg g$^{-1}$ (Figure 3.1 and Table 3.2), which, alongside the detection of *Rhizobia*, Fabaceae root-nodule symbionts, indicates that nitrogen limitation for plant growth may have been occurring in soils (Mylona *et al.,*1995). Actinorhizal and legume plants, which directly benefit from biological nitrogen fixation through symbiosis, such as Clover, are likely to prevail in developing forefield soils (Fagerli and Svenning 2005; Chaia *et al.,* 2010). This is because they maintain a competitive advantage over other plants in nitrogen limited conditions, typical of newly exposed soils (Menge and Hedin 2009; Bradley, Singarayer and Anesio 2014). Additionally, Rb had a lower average soil TOC content than other forefields, at 6.8 mg g$^{-1}$ (Figure 3.1 and Table 3.2). Thus, *Rhizobia* are likely to benefit from symbiosis with plants through the supply of reduced carbon (Denison and Kiers 2004). Plants may therefore be acting as a control on the forefield microbial community structure, endorsing the presence of root-associated diazotrophs (Miniaci *et al.,* 2007). Rarefaction curves for Rb sites were shown to be nearing saturation, indicating much of the microbial community structure was profiled (Figure 4, Appendix 2). Additional sequencing for these samples may reveal further low abundance taxa; however, it is likely that the most dominant fraction of diazotrophs have been identified adequately through our analysis.

The nif genes recovered from the MI forefield showed a wider taxonomic diversity of diazotrophs and contained sequences linked to the genera *Nostoc, Polaromonas*, *Bradyrhizobium*, *Pelobacter*, *Azoarcus* and *Anaeromyxobacter*. The presence of the Cyanobacteria, *Nostoc*, was expected due to the greater extent of bare soil observed in this forefield, enhancing the need for early colonisers (Frey *et al.,* 2013). Additionally, EPS production enables this group to resist harsh freeze-thaw cycles, common in Arctic environments (Tamaru *et al.,* 2005). Given the high latitude of this forefield, it is also not surprising to find *Polaromonas*, which are known psychrophiles (Irgens *et al.,*1996). The presence of *Bradyrhizobium* and *Frankia* indicate plants may require additional fixed nitrogen through symbiosis, corresponding with the low nitrogen stocks detected (Benson and Silvester 1993; Mylona *et al.,* 1995; Chaia et al., 2010; Figure 3.1 and Table 3.2). Additionally, the presence of legume symbiotic diazotrophs is interesting, as Fabaceae are non-native to Svalbard, having been introduced over the 20th Century (Fagerli and Svenning 2005). The absence of early plant colonisation in the forefield may also have been a control on overall microbial community structure, endorsing a range of non-symbiotic diazotrophs (Knelman *et al.,* 2012). Alongside *Geobacter*, the identification of *Pelobacter, Thiocystis* and *Anaeromyxobacter*, again indicates permanent or periodic anaerobic conditions in the glacier forefield, similarly to RI (Schink and Stieb, 1983; Sanford *et al.,* 2002). *Pelobacter* are anaerobic organisms containing diverse fermentative metabolisms, which may make this group well suited to the rapidly changing conditions in forefield soils (Schink, 2006). For example, *Pelobacter* have been shown to ferment acetylene using acetylene hydratase to acetate for cell growth or using nitrogenase to ethylene through nitrogen fixation (Akob *et al.,* 2017). The genomic results for the MI forefield falls in line with 16s amplicon data presented by Bradley *et al*., (2016). This study also found *Frankia, Rhizobium, Nostoc* and *Geobacter* in the MI forefield (Bradley *et al.,* 2016). The identification of additional organisms such as *Devosia*, *Sphingomonas* and *Rhodoplanes* may relate to the use of amplification in their methodology, thereby aiding the discovery of low abundance organisms (Bradley *et al.,* 2016). Additionally, some metagenomes from this forefield would have benefitted from greater sequencing depth in order to completely profile the microbial community composition (Figure 1, Appendix 2). Therefore, deep sequencing of these samples may reveal additional low abundance diazotrophs, unidentified in this analysis.

Finally, the St forefield contained sequences relating to *Nostoc, Geobacter, Rhizobium, Polaromonas* and *Frankia*, in line with the other forefields sampled (Figure 3.3). This supports the identification of a core group of diazotrophs present across Arctic glacier forefields. However, several diazotrophs detected at this site may also have importance in sulfur cycling, alongside nitrogen fixation (Figure 3.3). The detection of the anaerobic diazotrophs *Geobacter* and *Desulfovibrio* indicates the potential for sulfur reduction, whereby energy is gained through reducing sulfur (S) or sulfate ($SO_4^{2-}$) to hydrogen sulfide ($H_2S$), with the oxidation of organic carbon (Boopathy and Kulpa 1993; Caccavo *et al.,*1994). However, inorganic S and $SO_4$ have been found to be limiting for both plants and microbes in newly exposed glacier forefield soils (Allison *et al.,* 2007; Prietzel *et al.,* 2013). Nevertheless, desulfonating bacteria, whom metabolise organically bound sulfur to labile sulfates, have been found in forefield soils, and may therefore help overcome S limitation (Schmalenberger and Noll, 2009; Prietzel *et al.,* 2013). Additionally, suitable anaerobic growth conditions for sulfur reducing bacteria may occur frequently in stagnated proglacial meltwater pools and during periods of meltwater flushing (Duc *et al.,* 2009). Furthermore, the detection of organisms such as *Chlorobaculum, Thioflavicoccus, Halorhodospira and Thiocystis* indicates the potential for St forefield bacteria to carry out both nitrogen fixation and sulfur oxidation (Figure 3.3). These organisms have the potential to oxidise $H_2S$ to S and $SO_4$, alongside gaining fixed nitrogen through diazotrophy (Imhoff and Pfenning 2001; Chan, Morgan-Kiss and Hanson 2008; Peduzzi *et al.,* 2011; Challacombe *et al.,* 2013). The ability of these organisms to overcome nitrogen limitation through fixation, and to respire anaerobically in anoxic soils, may make this group well suited to harsh forefield environments. Additionally, as *Halorhodospira* is also halophilic, this may indicate resistance to high salinity environments, such as ice brine channels, or evaporation ponds in the St forefield (DasSarma and DasSarma 2006).

The diazotroph community composition observed using LCA nifHDKNE analysis was again largely consistent with those found at the Damma Glacier, Switzerland (Duc *et al.,* 2009; Frey *et al.,* 2013). This includes genera such as *Methylobacterium*, *Bradyrhizobium*, *Azotobacter*, *Anabaena*, *Nostoc* and *Geobacter* (Duc *et al.,* 2009). This supports the results from the nifH phylogeny, indicating the presence of consistent genera across forefields, which may be well adapted to the cold, oligotrophic and high UV conditions. Plant colonisation has also been identified as an influence on the diazotrophic community composition, in agreement with studies on the Damma Glacier, Switzerland (Miniaci *et al.,* 2007; Duc *et al.,* 2009; Zumsteg *et al.,* 2013). However, it is important to acknowledge that additional factors, such as latitude, bedrock minerology, organic matter and aeolian nitrogen deposition, may also have an influence on diazotroph community structure and abundance (Duc *et al.,* 2009; Zumsteg *et al.,* 2013). Some genera found by Duc *et al*., (2009), such as *Oscillatoria*, *Ideonella* and *Paenibacillus* were not identified in this study (Figure 3.3). This may relate to the absence of these organisms in the four forefields in this analysis, but also may relate to the alternate approach used. As this analysis uses unamplified nifH sequences, some low abundance organisms may not be sequenced due to incomplete sequencing depth in highly complex samples (Rodriguez and Konstantinos 2014; Figures 1–4, Appendix 2). Thus, it cannot be ruled out that these organisms were also not present in the forefields, but at a lower abundance than those captured by the sequencing effort (Prakash and Taylor, 2012). In order to profile the complete community of some metagenomes, including low abundance organisms, deeper sequencing would be required, due to the diverse nature of soil samples (Rodriguez and Konstantinos 2014). Despite this, this analysis has been able to capture a diverse group of diazotrophs that appear to be common across glacier forefields and are likely the most abundant fraction of the nitrogen-fixing community, as these were captured by unamplified DNA sequencing (Rodriguez and Konstantinos 2014).

### 3.4 Conclusions

Overall, this study has used metagenomics to understand the diversity of diazotrophs across four Arctic glacier forefields. The results of Chapter 2 were applied to this analysis, to assemble metagenome contigs for interpretation. The subsequent analysis used a nifH phylogeny to identify a key group of diazotrophs across four Arctic forefields, associated with both Cluster I and III nitrogenase, linked to aerobic and anaerobic organisms containing the typical Mo nifH (Zehr *et al.,* 2003). Incorporating multiple nif genes (HDKNE) revealed additional organisms from unamplified metagenome samples, compared to using the nifH gene exclusively. This may relate to the complex nature of soil metagenome samples, whereby sequencing depth is not always adequate to profile the complete microbial community diversity. Thus, to reveal all low abundance diazotrophs, some metagenomes would require additional deep sequencing. Key diazotrophs were found to be metabolically diverse, including genera such as *Geobacter*, *Frankia*, *Nostoc, Polaromonas* and *Bradyrhizobium*. A range of diazotrohic organisms outside the key group were also highlighted, including halophiles, psychrophiles and bacteria associated with fermentative metabolisms and sulfur cycling. Therefore, this analysis has shown a diverse group of diazotrohic bacteria present in Arctic forefield soils, including a consistent core subset. These diazotrophs have the potential to build labile nitrogen stocks in forefield soils, which may support further colonisation and soil development.

### 3.5 Limitations and Future work

Metagenomics was applied in this study to investigate the diversity of diazotrophic bacteria across four Arctic glacier forefields. However, metagenomics is unable to provide evidence for the activity of microbial nitrogen fixation, as it focuses on providing genomic verification for the potential of the pathway (Wooley *et al.,* 2010). To clarify if the diazotrophs recovered were active, transcriptomics could be used. Transcriptomics involves the sequencing of microbial community RNA, highlighting the function of actively expressed genes (Wang *et al.,* 2009). However, this technique only provides a snapshot of the expression at the time of sampling, so cannot be extrapolated to year-round activity (Lowe *et al*., 2017). It would also be interesting to investigate the rate of nitrogen fixation in each forefield and if there are any differences between sites. To do this, gasometric incubations of $N_2$ uptake could be implemented using forefield soil samples, for example utilising acetylene reduction assays, or alternatively utilizing a radio tracer for $N_2$ incorporation, such as $^{15}N_2$ (Burris *et al*., 1972; Hardy *et al.,*1973). This would highlight how active community nitrogen fixation was and how this varies over the chronosequences or between forefields.

Furthermore, as shown by the rarefaction analysis (Figures 1-4, Appendix 2) the sequencing coverage was not sufficient to profile the complete microbial community in all forefield samples. The recovered diazotrophs are likely to be those which were most abundant in the samples, as they would account for a greater portion of the sequenced DNA (Rodriguez and Konstantinidis 2014). To reveal less abundant diazotrophic organisms, additional deep sequencing could be carried out. Not only would this provide a more complete picture of the microbial community, but the enhanced sequencing coverage may enable more detailed analyses, for example extracting complete genomes from the metagenomes (Sharon and Banfield, 2013; Eloe-Fadrosh *et al.,* 2016b). This would allow a more refined investigation of the microbial genomes, for example, by providing genomic evidence for cold, desiccation or UV tolerance adaptations which enable these microbes to survive in Arctic environments (Chrismas *et al.,* 2016).

## Footnotes

[1] https://www.phylo.org/
[2] http://tree.bio.ed.ac.uk/software/figtree/
[3] http://www.evolgenius.info/evolview/
[4] https://inkscape.org/en/
[5] https://blast.ncbi.nlm.nih.gov/

# Chapter 4: Microbial community development over a gradient of soil succession along the Midtre Lovénbreen glacier forefield, Svalbard

Maisie V. Nash[1]; Alexandre M. Anesio[2]; Gary Barker[3], Martyn Tranter[1], Thomas Turpin-Jelfs[1], Liane G. Benning[456] and Patricia Sánchez-Baracaldo[1]

[1]School of Geographical Sciences, University of Bristol, BS8 1SS, UK
[2] Department of Environmental Science, Aarhus University, 4000-Roskilde, Denmark
[3] School of Life Sciences, University of Bristol, BS8 1TQ, UK
[4]GFZ German Research Centre for Geosciences, Telegrafenenberg, 14473 Potsdam, Germany
[5]School of Earth and Environment, University of Leeds, LS2 9JT, Leeds, UK
[6]Department of Earth Sciences, Free University of Berlin, Malteserstr, 74-100, Building A, 12249, Berlin, Germany

### Contributions and acknowledgements

MN carried out the analysis, interpretation and formulation of the manuscript. AA, MT and LB assisted with fieldwork logistics and sampling. TTJ supplied the raw TN and TOC data. PSB and AA helped refine drafts of the manuscript. Additional support came from James Bradley and Katherine Wright during fieldwork sampling.

### Funding

This work was carried out as part of the NERC GW4+ Doctoral Training Partnership, awarded to MN. Funding support was provided through NERC grants NE/J02399X/1, awarded to AA and NE/J022365/1 awarded to LGB. Funding support for PS-B came from a Royal Society University Research Fellowship and LGB also acknowledges the financial support through the Helmholtz Recruiting Initiative.

### 4.1 Introduction

Arctic glaciers face continued ice retreat with warming global temperatures (Zemp *et al.,* 2015). As the terminus of a glacier retreats, undeveloped soils are exposed which were previously shielded by ice cover (Bradley *et al.,* 2014). A transect of soil succession can often be identified in glacier forefields, with newly exposed bare soil close to the glacier terminus and more developed plant colonized soils moving outwards (as reviewed by Bradley *et al.,* 2014). These soils pose an interesting opportunity to understand how land is first colonized, the changes to microbe and plant communities during soil succession and the impact these communities have on soil structure and physicochemical characteristics during development (Edwards and Cook, 2015). Investigating microbial communities during forefield soil succession is important to understand Arctic microbial diversity and the role of microbial communities in soil biogeochemical cycles (Bradley *et al.,* 2014). Additionally, this information will help improve our understanding of how Arctic microbial ecology and nutrient cycles may modify in the future, given the continued retreat of glaciers with global warming (Edwards and Cook, 2015).

A chronosequence based approach can be applied to study soil succession in glacier forefields, employing a space for time substitution, with older soils at a greater distance from the glacier terminus (Tscherko *et al.,* 2003; Bradley *et al.,* 2014). Thus far, the majority of forefield chronosequence studies have focused on plant communities and soil structure during succession (Ohtonen *et al.,* 1999; Strauss *et al.,* 2009; Knelman *et al.,* 2012). A general increase in organic carbon and nitrogen has been shown during succession, alongside a reduction in pH, as soils become colonised by plants (Ohtonen *et al.,* 1999; Strauss *et al.,* 2009; Knelman *et al.,* 2012; Turpin-Jelfs *et al.,* 2019). Microbial community colonisation and development in forefields is important for influencing soil nutrient cycles, physicochemical status and structure (Tscherko *et al.,* 2003; Hahn and Quideau, 2013). Microbial communities are typically the initial colonisers of newly exposed soils and thus facilitate soil development (Chapin *et al.,* 1994; Hahn and Quideau, 2013). In particular, phototrophic and diazotrophic microbes are important for building labile carbon and nitrogen stocks, facilitating the establishment of heterotrophic microbial populations and plants (Chapin *et al.,* 1994; Hahn and Quideau, 2013). Microbial carbon and nitrogen fixation is particularly important following glacier retreat, as newly exposed soils are often oligotrophic (Chapin *et al.,* 1994). Additionally, microbes help to stabilise soils for plant colonisation through the secretion of extracellular polymeric substances (EPS) during growth (Magner and Thomas, 2011).

The initial coloniser microbial communities in glacial forefields have attracted substantial research, as these microbes are key facilitators for further soil development. However, the composition of these initial communities remains unclear (Kastovska *et al.,* 2005). On one hand, diazotrophic and phototrophic bacteria may constitute early colonisers, as these microbes can fix carbon and nitrogen in nutrient deplete conditions, subsequently facilitating labile nutrient pools for heterotroph colonisation (Hodkinson *et al.,* 2002). However, heterotrophic microbes may also be present in initial soils given the availability of organic carbon, which may be derived from ancient organic matter in overridden soils or from aeolian deposition (Kastovska *et al.,* 2005; Bardgett *et al.,* 2007; Bradley *et al.,* 2014). The pioneer microbial community may be derived from ice surface microbes, aeolian deposition, or legacy subglacial microbial communities (Anesio *et al.,* 2009; Boyd *et al.,* 2010). Furthermore, the composition of forefield microbial communities has been shown to modify with soil succession, related to changes in environmental factors such as nutrients, organic matter, water flow pathways and pH (Tscherko *et al.,* 2003; Sakata Bekku *et al.,* 2004; Frey *et al.,* 2013). Additionally, research by Knelman *et al.,* (2012) identified that plant colonisation (alongside the type of plant), as having a key influence on microbial community structure during succession, as opposed to soil pH and nutrient status.

Nutrient and organic matter inputs to forefields may be from both allochthonous (glacial runoff, animal droppings and aeolian deposition) and autochthonous sources (*in situ* microbial activity) (Bradley *et al.,* 2014). Bioavailable nitrogen is needed by both plants and microbes for protein synthesis (Tresder, 2008). Nitrogen is supplied to forefields by diazotrophic nitrogen fixation, aeolian deposition, degradation of organic matter and in washing from snowmelt (Bradley *et al.,* 2014; Bradley *et al.,* 2015; Duc *et al.,* 2009). Nitrogen fixation is important in newly exposed soils, especially when the total nitrogen content is initially low (Duc *et al.,* 2009; Turpin-Jelfs *et al.,* 2019). Denitrification has been shown to increase in line with soil succession, attributed to the increase in water logged soils following plant colonization (Kandeler *et al.,* 2006; Schulz *et al.,* 2011). Additionally, nitrification has been shown to increase with soil age, due to a greater availability of bioavailable nitrogen in older soils (Brankatschk *et al.,* 2011). Nitrogen mineralization from organic matter is an important mechanism for recycling bioavailable nitrogen in forefield soils (Brankatschk *et al.,* 2011). This process converts organic nitrogen from cell death or excrement to labile ammonium (Figure 4.1). A schematic identifying the key processes involved in the soil nitrogen cycle is available in Figure 4.1

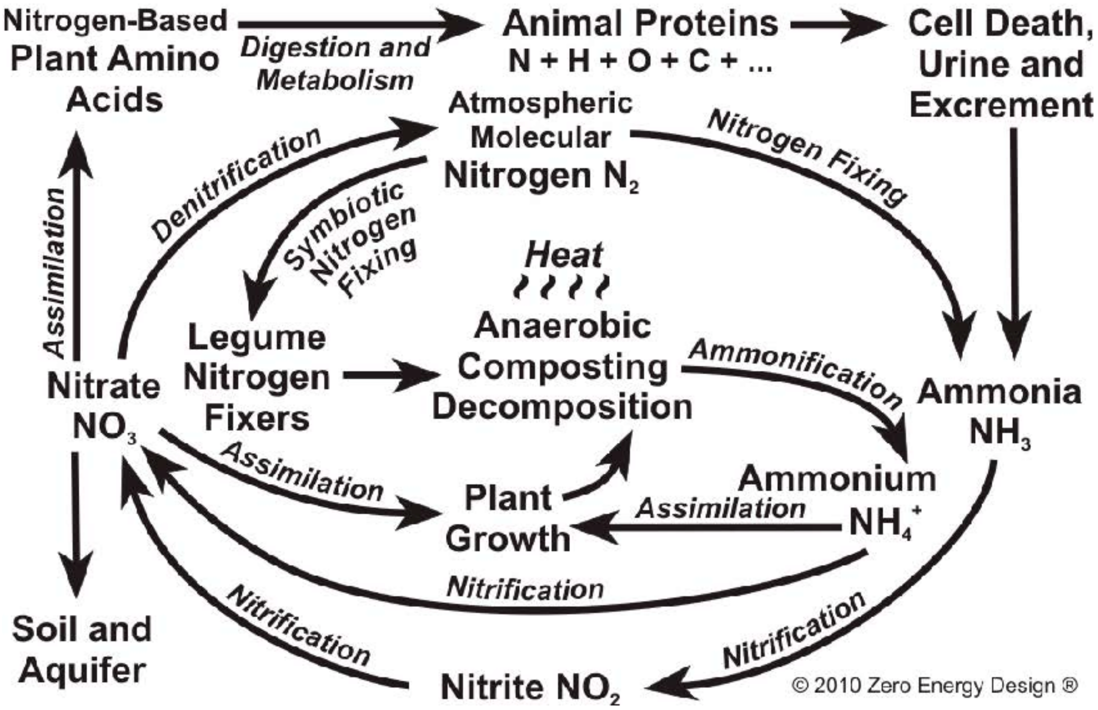


***Figure 4.1:*** *The soil nitrogen cycle. Sourced from Abatenh et al., (2018)*

Furthermore, organic carbon is a crucial nutrient for microbial growth and forms the backbone of molecules. Organic carbon can be supplied through aeolian deposition, carbon fixation and may also be present in the form of ancient organic matter in overridden soils (Hodkinson *et al.,* 2002; Guelland *et al.,* 2013). During the course of succession, organic carbon stocks increase in soils, particularly following plant colonization, whereby plant litter can be recycled by microbial communities back into the soil carbon pool (Knelman *et al.,* 2012; Zumsteg *et al.,* 2013). In early soils, phototrophic bacteria such as cyanobacteria can be crucial, as these microbes fix carbon dioxide into labile organic carbon (Strauss *et al.,* 2012). These microbes may therefore be key facilitators of heterotrophic colonization, as heterotrophs rely on the availability of labile fixed carbon (Freeman *et al.,* 2009; Zumsteg *et al.,* 2013). Finally, sulfur may be cycled in forefield soils through microbial redox reactions. Energy can be produced by chemolithotrophic bacteria through the oxidation of sulphur, coupled to the reduction of carbon or nitrate (Koltz *et al.,* 2011, Wainwright 1978). Sulfate can also be reduced when used as a terminal electron acceptor for the oxidation of organic matter, in the absence of oxygen, forming hydrogen sulfide (Widdel and Hansen 1992). Isotopic evidence for microbial sulfate reduction has been found in Arctic subglacial systems (Wadham *et al.,* 2004) and therefore may continue in soils following ice retreat.

The Damma Glacier, Switzerland, has often been used as a study site to investigate soil, microbial and plant succession (Bernasconi *et al.,* 2008; Duc *et al.,* 2009; Frey *et al.,* 2010; Bernasconi *et al.,* 2011; Brunner *et al.,* 2011). Comparatively, less research is available for the Midtre Lovénbreen forefield, Svalbard (Kastovska *et al.,* 2005; Schutte *et al.,* 2009; Bradley *et al.,* 2016; Nash *et al.,* 2018). Bradley *et al.,* (2016) combined field and laboratory data from the Midtre Lovénbreen glacier with the forefield microbial model SHIMMER 1.0, to help inform parameter values (Bradley *et al.,* 2016). The model results show that microbial biomass developed over the chronosequence and microbial activity was important for nutrient fixation and recycling in soils (Bradley *et al.,* 2016). Consequently, microbial communities may be important facilitators for plant colonization in these oligotrophic forefield soils (Bradley *et al.,* 2016). Furthermore, the microbial community of five Svalbard Glaciers, including Midtre Lovénbreen was investigated by Kastovska *et al.,* (2005). Both cryoconite and barren soils were found to contain the highest abundance of cyanobacteria, however cyanobacteria were also found in vegetated sites. Schutte *et al.,* (2009) found different microbial community compositions in surface and mineral depth soils, alongside a difference between newly exposed and more developed soils in the Midtre Lovénbreen forefield. However, this study also exemplified how forefield disturbances, such as water flow pathways may disrupt the pattern of succession (Schutte *et al.,* 2009). Finally, the Midtre Lovénbreen forefield was used in Chapter 3 of this thesis to investigate the diversity of diazotrophs between four Arctic glacier forefields using functional genes for nitrogen fixation (nif genes) (Nash *et al.,* 2018). Diazotrophs were identified in the Midtre Lovénbreen forefield and included cyanobacteria (*Nostoc*), root associated bacteria (*Rhizobia, Frankia*) and anaerobes (*Geobacter*) (Nash *et al.,* 2018).

Thus far, the majority of research on microbial communities in glacial forefields has used 16s rRNA sequencing or targeted functional gene amplification to investigate microbial community taxonomy (Sigler and Zeyer, 2002; Nemergut *et al.,* 2007; Knelman *et al.,* 2012; Zumsteg *et al.,* 2013). Metagenomics involves unamplified sequencing of the microbial community gene pool, providing information on both the taxonomy and functional genes present (Wooley *et al.,* 2010). Metagenomics has yet to be applied to understanding microbial community development during soil succession. Metagenomic sequencing would be a useful contribution to the current body of literature, as it would provide more information on microbial community functional potential and their role in soil biogeochemical cycles during succession. Furthermore, genome binning is a method which can be used to extract single draft genomes from community metagenome data (Albertsen *et al.,* 2013; Nielsen *et al.,* 2014). This technique is beneficial as it allows draft genomes to be extracted without the need for culturing (Kunin *et al.,* 2008; Albertsen *et al.,* 2013). The completeness of the draft genome can vary depending on the depth of sequencing and the complexity of the microbial community under investigation (Kunin *et al.,* 2008; Albertsen *et al.,* 2013). This means that the gene content, functional pathways and taxonomy of genomes of interest can be investigated further. This technique has yet to be applied to forefield samples and would be beneficial to explore the novelty of genomes in this harsh environment, which is not possible with 16s rRNA analysis.

This study aims to use metagenomics to explore the microbial community composition along a chronosequence of soil succession in the Midtre Lovénbreen forefield, Svalbard. In particular, we hypothesize that the microbial community composition, function and metabolic pathways modify during succession, in line with soil development, the establishment of labile nutrient pools and plants. We hypothesize that autotrophic carbon and nitrogen-fixing bacteria will be prevalent in low nutrient early soils, with heterotrophic microbes dominant in older soils with established nutrient pools. We expect carbon and nitrogen fixation to be detected in recently deglaciated soils, with pathways such as denitrification, carbon remineralization and sulfate reduction to occur in line with nutrient pool development. Additionally, we hypothesis that extremophilic, well adapted or unique genomes may be found in the metagenomes, related to the harsh environmental conditions found in forefield soils. Whilst Chapter 3 looked at the broad diversity of diazotrophs between forefields, this analysis investigates whole community changes within a forefield during succession. In addition, this work follows on from Chapter 2, as the metagenome assembler, metaSPAdes, is applied for functional analysis. The results of this analysis aim to contribute to the current body of literature on forefield microbial communities, providing additional evidence for community functional potential, which has not previously been explored in the context of succession.

## 4.2 Methodology

### 4.2.1 Field Sampling

Surface sediment was sampled along the forefield of Midtre Lovénbreen glacier, Svalbard, in July 2013. Additional sampling of cryoconite sediment and dark ice from the glacier surface was carried out for comparison. Chronosequence based sampling of the forefield was implemented, whereby sediment was obtained from three parallel transects, moving out from the glacier terminus (Bradley *et al.,* 2014; Nash *et al*., 2018). This chronosequence sampling technique was applied to capture the changes in microbial community composition and nutrient content with soil development along the forefield. Additionally, this sampling approach enabled the sites to be dated from their year of exposure following glacier retreat (Brankatschk *et al*., 2011; Bradley *et al.,* 2014; Bradley *et al.,* 2015). Bulk sediment was sampled in triplicate from each site into Whirlpak$^{TM}$ bags, and frozen at -20° C prior to analysis. The sampling site locations, transect numbers (T1-T3) and associated soil age are provided in Table 4.1, with sites ranging from 0 – 2000 years since glacier retreat (Brankatschk *et al.*, 2011; Bradley *et al.,* 2015).

***Table 4.1:*** *Sampling site locations and associated soil exposure age (measured in years since glacier retreat), in line with Brankatschk et al., (2011) and Bradley et al., (2015). Samples obtained from the glacier surface are marked with '*'.*

| Sample Number | Latitude / Longitude | Transect number (T1 - T3) | Estimated soil age (Years) |
|---|---|---|---|
| 1 | 79.101 / 12.156 | T2 | 0 |
| 2 | 79.112 / 12.175 | T2 | 3 |
| 3 | 79.112 / 12.258 | T3 | 3 |
| 4 | 79.118 / 12.094 | T1 | 5 |
| 5 | 79.114 / 12.196 | T2 | 5 |
| 6 | 79.104 / 12.279 | T3 | 5 |
| 7 | 79.152 / 12.216 | T1 | 29 |
| 8 | 79.151 / 12.254 | T2 | 29 |
| 9 | 79.141 / 12.090 | T3 | 29 |
| 10 | 78.928 / 12.254 | T1 | 50 |
| 11 | 78.927 / 12.077 | T2 | 50 |
| 12 | 78.908 / 12.164 | T3 | 50 |
| 13 | 78.901 / 12.076 | T2 | 50 – 113 |
| 14 | 78.901 / 12.076 | T2 | 50 – 113 |
| 15 | 78.901 / 12.076 | T2 | 50 - 113 |
| 16 | 78.990 / 12.083 | T2 | 113 |
| 17 | 78.992 / 12.230 | T2 | 113 |
| 18 | 78.979 / 12.332 | T2 | 113 |
| 19 | 79.768 / 12.144 | T2 | 2000 |
| 20 | 79.768 / 12.144 | T2 | 2000 |
| 21 | 79.768 / 12.144 | T2 | 2000 |
| 22 | 79.484 / 12.092 | N/A | *Cryoconite hole |
| 23 | 79.484 / 12.092 | N/A | *Dark Ice |

### 4.2.2 DNA extraction, library preparation and sequencing

In total, 23 samples were selected for DNA extraction and metagenome sequencing (Table 4.1), spanning the range of soil ages. DNA from bulk soil samples was extracted in line with the methodology detailed in Chapter 3, section 3.2.3 (Nash *et al.,* 2018). Metagenomes were sequenced using an Illumina-Mi Seq, with a TruSeq library prep kit at the University of Bristol Genomics facility. The 23 selected sites were sequenced for metagenomics, providing 2x 100bp paired-end reads. Sequencing read output for each site ranged between 4013376 – 66567072 reads per metagenome (Table 4.2).

***Table 4.2:*** *Statistics for unassembled and assembled metagenomes, including contiguity and completeness metrics. The % Coverage is calculated using the percentage of raw reads mapped back to the assembly.*

| Sample | Unassembled Reads | Assembled Contigs | Contigs > 1000bp | Largest Contig | % Coverage |
|---|---|---|---|---|---|
| 1 | 18800440 | 58428 | 445 | 52064 | 23.2 |
| 2 | 23227446 | 15534 | 18 | 43675 | 3.7 |
| 3 | 17611153 | 34535 | 154 | 314188 | 16.0 |
| 4 | 20076741 | 14007 | 7 | 46511 | 2.9 |
| 5 | 22256997 | 26820 | 31 | 32120 | 6.4 |
| 6 | 21156969 | 33757 | 91 | 51529 | 7.6 |
| 7 | 22729904 | 19399 | 199 | 106118 | 4.6 |
| 8 | 20076740 | 14007 | 7 | 46511 | 2.9 |
| 9 | 22977962 | 15533 | 19 | 43675 | 2.7 |
| 10 | 57959962 | 60667 | 221 | 150624 | 5.1 |
| 11 | 38013530 | 22180 | 3 | 67008 | 1.9 |
| 12 | 38014196 | 22214 | 3 | 67008 | 1.9 |
| 13 | 4013376 | 851 | 0 | 8102 | 1.5 |
| 14 | 22329785 | 13861 | 38 | 1072806 | 3.7 |
| 15 | 22371213 | 11373 | 16 | 19032 | 2.4 |
| 16 | 22371175 | 11374 | 16 | 19032 | 2.4 |
| 17 | 38409045 | 19556 | 296 | 256358 | 3.5 |
| 18 | 66567072 | 75376 | 423 | 768884 | 6.9 |
| 19 | 23966913 | 5931 | 0 | 4646 | 0.9 |
| 20 | 20702648 | 4905 | 0 | 5800 | 0.9 |
| 21 | 29107763 | 14288 | 0 | 8822 | 1.6 |
| 22 | 19259216 | 89619 | 878 | 371579 | 26.6 |
| 23 | 22255419 | 26820 | 31 | 32120 | 9.5 |

### 4.2.3 Rarefaction curves

Rarefaction curves for quality trimmed sequencing reads were created in MG RAST 4.0.3 (Glass *et al.,* 2010; Figure 4.2). Rarefaction curves display the number of species obtained with increasing sequencing effort (number of reads). Curves which reach saturation (i.e. more sequencing effort does not increase the species count) highlight metagenomes which profile the microbial diversity sufficiently (Rodriguez and Konstantinidis, 2014). Metagenomes which are under saturated profile the more abundant factions of the microbial community, however may not sample those which are less abundant. Under saturated rarefaction curves are common when investigating complex microbial communities, such as those in soil, as the sequencing depth may not be enough to recover the entire microbial community composition (Rodriguez and Konstantinidis, 2014). In these cases, additional deep re-sequencing would be useful if the less abundant microbial fractions were of particular interest.

### 4.2.4 Metagenome assembly and mapping

The 23 metagenomes were quality trimmed and subsequently assembled with metaSPAdes v3.11.1 (Nurk *et al.,* 2017). The size and contiguity of the assembled metagenomes were evaluated using QUAST on the KBASE platform (Arkin *et al.,* 2016). Raw reads were mapped back to the assemblies using Bowtie 2 v2.3.2, to identify the proportion of the input reads which were incorporated into the final assemblies (Langmead and Salzberg, 2012). Assembly size for each metagenome ranged between 851 – 89619 contigs and read coverage ranged between 0.9 – 23.2 % (Table 4.2). These read coverage scores mean that over 75% of reads in each sample were not incorporated in to the assembled metagenome. The low coverage scores may be attributed to the complex community structure, combined with a sequencing depth which was insufficient to fully profile all the organisms. Combined, these can result in short fragmented assemblies, as the assembler cannot resolve repeat regions or insertion/ deletions (INDELs).

### 4.2.5 Taxonomic annotation

Read-based taxonomic annotation for each metagenome was carried out in Kaiju v1.5.0 (Menzel *et al.,* 2016). Read based annotation was used, as opposed to assembly annotation, due to the low read recruitment in assemblies (Table 4.2). Consequently, using trimmed reads enabled more of the dataset to be annotated, which is particularly important when aiming to gain an overview of the complete community composition. Kaiju is a kmer based taxonomic classifier, based on protein sequences (Menzel *et al.,* 2016). This classifier is suited to novel metagenome samples and those with sequencing errors, as sequence conservation is generally greater in protein sequences than the corresponding DNA (Menzel *et al.,* 2016).

### 4.2.6 Metagenome binning

Genome binning was applied to recover draft microbial genomes from the metagenomes, using MaxBin2 v2.2.4 (Bankevich *et al.,* 2012; Wu *et al.,* 2015). This binning algorithm uses an expectation- maximum approach, based on contig tetra nucleotide frequencies, to recover draft genomes (Wu *et al.,* 2015). MaxBin2 employs the assembled metagenomes and coverage information from raw reads to group contigs into discrete genome bins. In this instance, using assembled metagenomes was beneficial to provide longer sequences for tetra nucleotide frequencies to be assessed during binning and subsequent gene annotation. BinUtil v1.0.1 was used to extract the genomes from the metagenomes using the KBASE platform (Arkin *et al.,* 2016). Each genome bin was functionally annotated using RAST v0.0.12 on KBASE (Aziz *et al.,* 2008; Overbeek *et al.,* 2013; Arkin *et al.,* 2016). Features were predicted using glimmer3 and prodigal in RAST, before matching to the SEED ontology (Aziz *et al.,* 2008; Overbeek *et al.,* 2013). For each sampling site, extracted genomes were pooled and a function profile was created in KBASE, based on the SEED Ontology (Overbeek *et al.,* 2013; Arkin *et al.,* 2016). The SEED categories relating to nitrogen cycling (nitrogen fixation, nitrification and denitrification) were selected for analysis, given the importance of the nutrient for microbial growth and the limited nitrogen stocks recorded (Table 4.3).

Each genome bin was assigned a taxonomy using BLASTn against NCBI RefSeq (e $10^{-5}$). For each sampling site, the genome bins were aligned to closely related genomes on the KBASE platform, using Species Tree Builder, based on a subset of COG groups (Appendix 3 Table 1). The alignments were used to generate a phylogenetic tree for each site, using Fast Tree 2.2.10 (Aziz *et al.,* 2008). Given the incomplete sequencing depth and low assembly read recruitment, we do not aim to gain any near complete draft genomes or sample all genomes present. However, genome binning can provide an insight into the functional potential of some of the most abundant genomes in the forefield. Additionally, this exemplified the potential of metagenomics to reveal genome specific functions, given the availability of deep sequencing.

### 4.2.7 Soil metadata

Soil pH and temperature for each of the sampling sites was obtained using a Hanna pH meter. Soil total nitrogen (TN) and total organic carbon (TOC) for each of the 23 sediment samples was evaluated using the protocol detailed in Chapter 3, section 3.2.2 (Nash *et al.,* 2018*).*

## 4.3 Results

### 4.3.1 Rarefaction curves

Rarefaction curves for each of the 23 metagenomes are displayed in Figure 4.2. Mid transect sites (50-113 years) display more under saturation than earlier sites, which could be related to an increase in species count and diversity but no increase in sequencing effort (Figure 4.2). Additional sequencing of these samples would be beneficial to recover the less abundant microorganisms. However, no samples were shown to be severely under sequenced and therefore we can be confident in profiling the most abundant constituents of the microbial communities for each site.

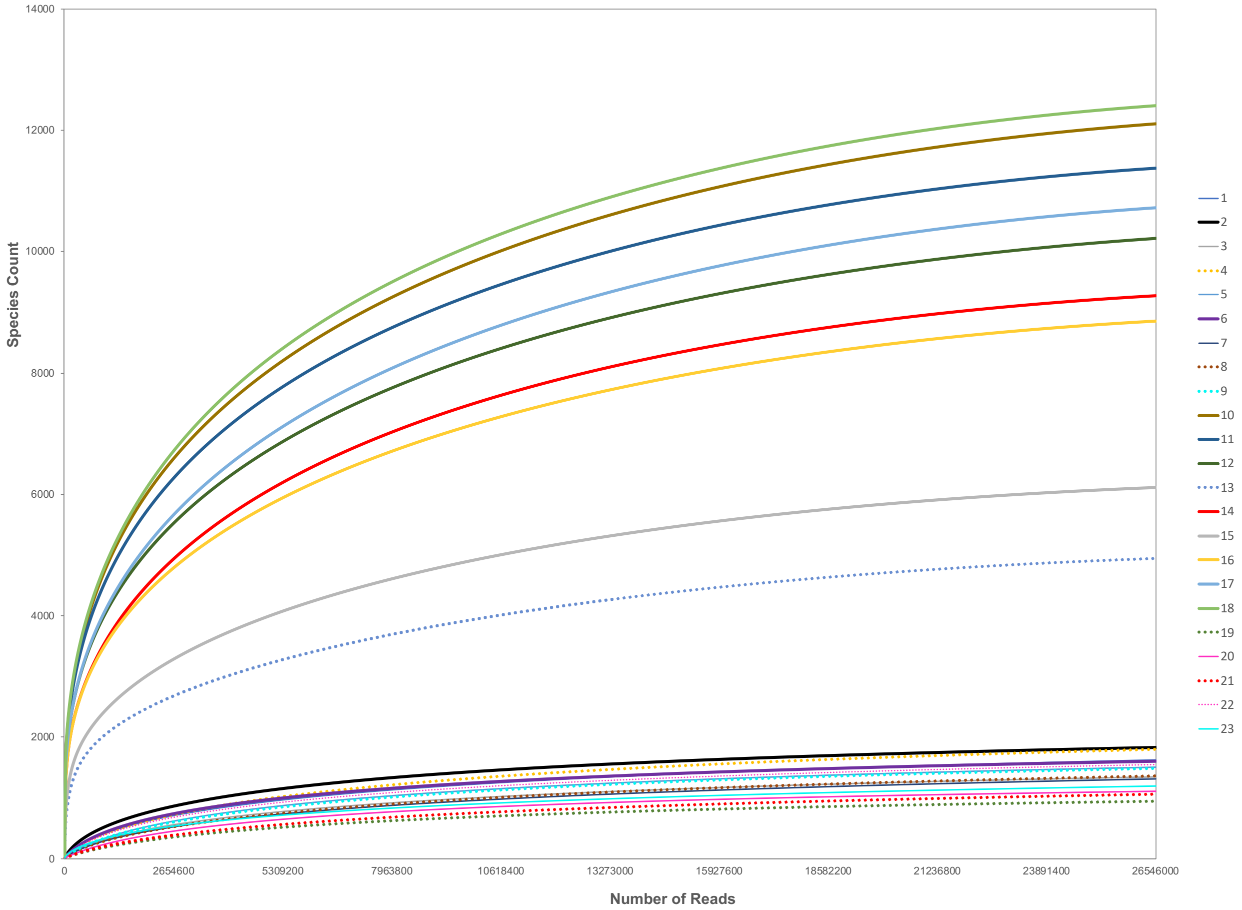


***Figure 4.2:*** *Rarefaction curves for the 23 metagenomes sampled from Midtre Lovénbreen, Svalbard. The curves display the number of species with increasing sampling (or number of reads). Curves that reach saturation display an adequate sequencing depth to profile the microbial community.*

### 4.3.2 Soil metadata

Soil temperature, pH, TN and TOC measurements were obtained from each metagenome sampling site and are displayed in Table 4.3. Soil temperature ranged between 4.8 – 8.5 °C, between sites 1- 15 respectively. Soil pH ranged around neutral, between 6.2 – 8.5, and does not show a distinct variance with soil age (Table 4.3). The TOC values fell in to the general range observed in Arctic soils, with most soils ranging between 0 – 40 mg $g^{-1}$ TOC (Chu *et al.,* 2010). The results for TOC and TN are displayed in Table 4.3 and shown graphically in Figure 4.3. TOC ranged between 2.47 mg $g^{-1}$ – 86.97 mg $g^{-1}$, between sites 8 – 21 (ages 29 – 2000) (Figure 4.3). The highest TOC values were found in the oldest soils (2000 years), the furthest sampled from the glacier terminus (Figure 4.3). Furthermore, in total, 86% of forefield sites had a TN content which fell below the instrumental limit of detection of 1 mg g $^{-1}$ (Table 4.3). However, older sites 19-21 (2000 years) had detectable TN, ranging between 3.89 - 4.89 mg $g^{-1}$ (Table 4.3). Consequently, both TN and TOC values showed an increase with rising soil age (Figure 4.3). Results of a one way ANOVA between all sites showed a significant difference in both TN and TOC, at the 0.01 significance level (Table 4.4). A post-hoc Tukey identified this significant difference to occur between earlier soils (ages 0-113 years) and the 2000 year soils (Table 4.4) This identifies a statistically significant increase in TN and TOC along the forefield, as expected with soil development.

***Table 4.3:*** *Total organic carbon (TOC), total nitrogen (TN), temperature and pH for metagenome sampling sites 1 – 21, soil ages 0- 2000 years. For TN and TOC analysis the instrumental limit of detection was 1 mg g$^{-1}$, with samples falling below this limit marked as b.d. TN and TOC vales are not available for ice sampling sites 22 and 23.*

| Sample Number | Soil age (Years) | TOC (mg g$^{-1}$) | TN (mg g$^{-1}$) | Temperature (°C) | pH |
|---|---|---|---|---|---|
| 1 | 0 | 6.68 | b.d. | 4.8 | 7.5 |
| 2 | 3 | 4.30 | b.d. | 7.6 | 7.6 |
| 3 | 3 | 3.89 | b.d. | 5.9 | 7.0 |
| 4 | 5 | 3.72 | b.d. | 6.6 | 7.5 |
| 5 | 5 | 3.68 | b.d. | 7.5 | 6.2 |
| 6 | 5 | 5.07 | b.d. | 7.6 | 7.3 |
| 7 | 29 | 4.07 | b.d. | 7.0 | 7.5 |
| 8 | 29 | 2.47 | b.d. | 8.3 | 7.5 |
| 9 | 29 | 2.96 | b.d. | 8.0 | 7.5 |
| 10 | 50 | 3.48 | b.d. | 7.4 | 7.5 |
| 11 | 50 | 5.47 | b.d. | 8.4 | 7.5 |
| 12 | 50 | 7.17 | b.d. | 8.5 | 7.3 |
| 13 | 50-113 | 3.09 | b.d. | 8.5 | 7.4 |
| 14 | 50-113 | 4.95 | b.d. | 8.4 | 7.5 |
| 15 | 50-113 | 2.72 | b.d. | 8.5 | 7.3 |
| 16 | 113 | 4.67 | b.d. | 7.8 | 7.5 |
| 17 | 113 | 11.53 | b.d. | 7.8 | 8.5 |
| 18 | 113 | 9.97 | b.d. | 7.7 | 7.5 |
| 19 | 2000 | 54.88 | 3.89 | 7.4 | 7.5 |
| 20 | 2000 | 67.08 | 4.39 | 7.4 | 7.5 |
| 21 | 2000 | 86.97 | 4.89 | 8.1 | 7.5 |

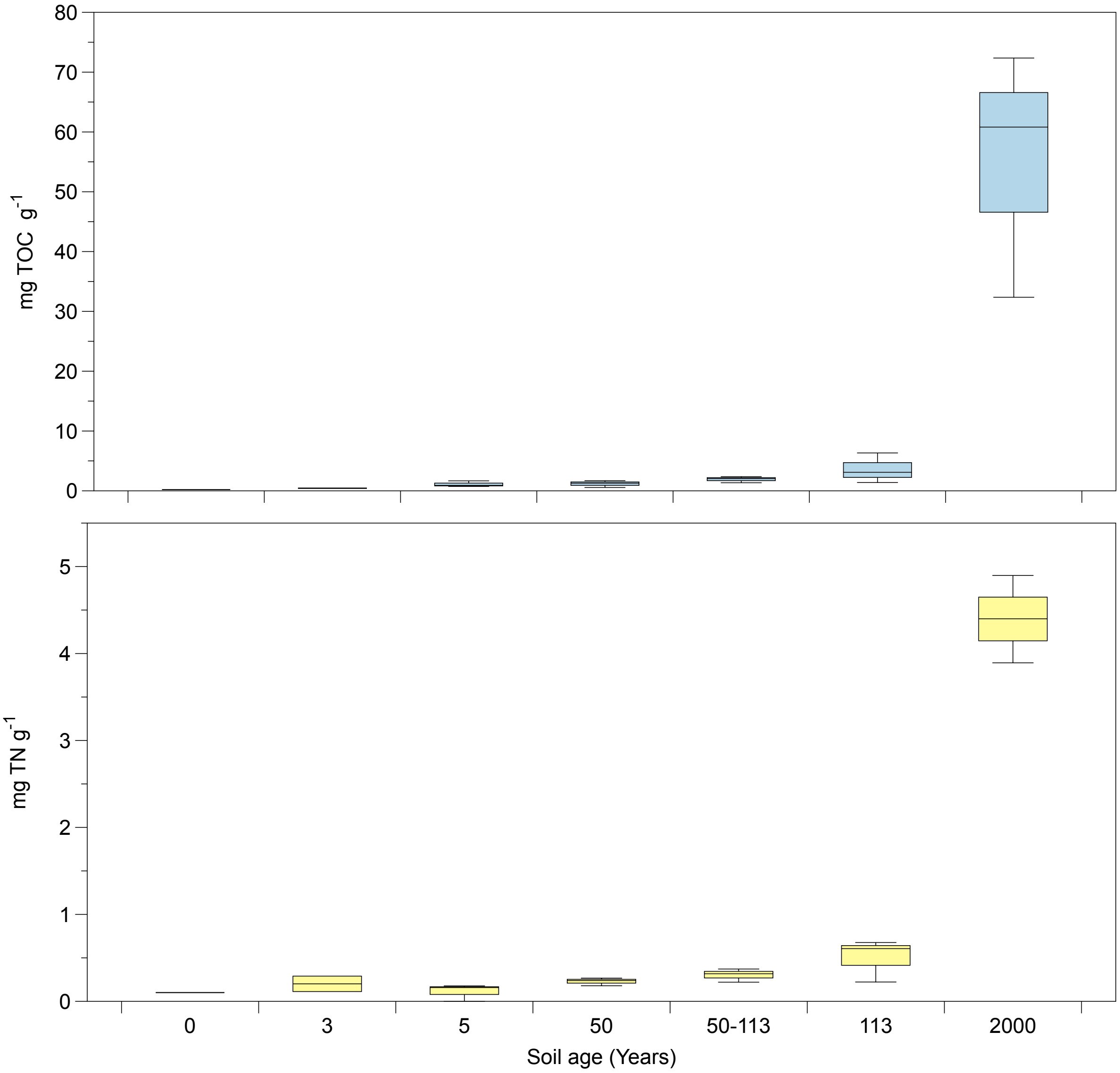


***Figure 4.3:*** *Box plots for total organic carbon (TOC) and total nitrogen (TN) for Midtre Lovénbreen forefield sampling sites, spanning soil ages 0 – 2000 years. The boxplots show the mean, first quartile, third quartile and range of the replicate field samples (where available).*

***Table 4.4:*** One-way ANOVA *comparing TN and TOC differences between all soils. A post-hoc Tukey identifies which sites display the significant difference. Significant differences observed between the sites are noted at the 0.01 or 0.05 level.*

| | **Total organic carbon (TOC)** | **Total Nitrogen (TN)** |
|---|---|---|
| **p value** | 1.5e-08 | 7.3e-15 |
| **significance level** | 0.01 | 0.01 |
| **Pot hoc Tukey**<br>0-113 vs. 2000 years | All (P<0.01) | All (P<0.01) |

### 4.3.3 Microbial community composition

The read-based microbial community composition of Midtre Lovénbreen metagenomes is displayed in Figure 4.4 and in Appendix 3 Table 2 (Genus Level) and in Table 4.5 (Phylum level). Metagenomes spanning the forefield chronosequence (0 years – 2000 years) are displayed alongside cryoconite and dark ice samples for comparison. The most abundant phyla recovered included Proteobacteria and Actinobacteria, accounting for 35% and 21% of annotated protein coding genes, respectively (Table 4.5). Other phyla assigned included Cyanobacteria, Chloroflexi, Verrucomicrobia, Firmicutes and Bacteroidetes, amongst others (Table 4.5). The percentage of protein coding genes attributed to Acidobacteria, Verrucomicrobia and Nitrospirae increased along the forefield, from 1.0%, 0.6% and 0% at 0 years to 11.7%, 4.9% and 0.9% at 2000 years, respectively (Table 4.5). This does not show an absolute change in organism abundance, due to differing depths of sequencing (Figure 4.2). However, as the sequencing profiles the most abundant fraction of the microbial community, it shows these phyla become a greater fraction of the microbial community recovered at older soil ages. Furthermore, in site 0 and dark ice, cyanobacteria account for 13.8% of the protein coding genes (Table 4.5). However, this reduces to 1.59% in soils aged 2000 years (Table 4.5). Whilst the cryoconite sample contains relatively few protein coding genes associated with cyanobacteria (0.7%), 13.8% of genes are attributed to the phototroph Chloroflexi, compared to the range of 1.4 - 3.9% in forefield samples (Table 4.5). Interestingly, no clear pattern can be seen in the distribution of Archaea and Fungi (Table 4.5). The Fungi Ascomyota is present in soils aged 0,3 and 50-113 years, with the Archaea, Thaumarcheaota and Eurayarchaeota present in 113 years and 3 years,113 years and basal ice, respectively (Table 4.5).

Between 34.6 – 57.9 % of metagenome reads were annotated using Kaiju at the genus level (Table 4.6). The reads were assigned to a range of taxa, associated with diverse metabolisms including nitrogen, carbon and sulfur cycling (Figure 4.4). In particular, genera associated with nitrogen fixation (*Rhizobium, Nostoc, Oscillatoria*), denitrification (*Conexibacter*), nitrification (*Nitrospira*), sulfur oxidation (*Chromatiales, Thiobacillus*), sulfur reduction (*Shewanella*) and carbon fixation (*Phormedesmis, Nostoc, Oscillatoria, Synechococcales, Rhodoferax, Rhodoplanes, Crococcales*) were identified (Figure 4.4; Appendix 3 Table 2). Genera associated with both aerobic (*Niabella, Variovorax, Rhodococcus*) and anaerobic metabolisms (*Rhodoferax, Geobacter, Bacillus*) were also identified (Figure 4.4). Whilst we cannot determine the change in abundance of microbes along the chronosequence, due to differing metagenome sequencing depths, we can identify changes to the most abundant fraction of the microbial community. Soils aged 0 years contained the diazotrophs *Rhizobium* and *Nostoc,* alongside the sulfur oxidizer *Thiobacillus* (Figure 4.4). In soils aged 3-5 years, the recovered microbial community included nitrifying, denitrifying and sulfur reducing microbes (Figure 4.4). Carbon fixing (and often diazotrophic) cyanobacteria detected in the metagenomes included *Synechococcales, Phormidesmis, Oscillatoriales* and *Nostocales* (Figure 4.3). Cyanobacteria were only recovered from soils aged 0 years and cryoconite (8.4 and 11% of annotated reads) and not found in more developed soil sites (Figure 4.4). Root associated microbes, including symbiotic diazotrophs (*Frankia, Rhizobium, Massilia, Burkholderia*) were detected throughout the forefield and accounted for 1.2% - 3.4% of annotated reads, between soils ages 0 years and 2000 years, respectively (Figure 4.4).

***Table 4.5:*** *Read based microbial community composition of Midtre Lovénbreen sample sites, presented as the percentage of annotated sequences at the phylum level. Replicate metagenomes for each soil age/ sampling site have been merged.*

| | **0** | **3** | **5** | **29** | **50** | **50 - 113** | **113** | **2000** | **Cryoconite** | **Dark ice** |
|---|---|---|---|---|---|---|---|---|---|---|
| *Acidobacteria* | 1.01 | 1.62 | 3.31 | 4.65 | 6.20 | 4.75 | 4.96 | 11.73 | 1.79 | 1.79 |
| *Actinobacteria* | 8.77 | 22.67 | 24.39 | 29.50 | 24.28 | 17.61 | 25.29 | 15.53 | 20.64 | 20.64 |
| *Ascomycota* | 1.33 | 4.01 | 0.00 | 0.00 | 0.00 | 0.66 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Bacteroidetes* | 13.08 | 4.14 | 7.26 | 7.55 | 7.59 | 8.00 | 7.44 | 6.24 | 1.21 | 10.19 |
| *Candidatus Rokubacteria* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.90 | 0.00 | 0.00 |
| *Chloroflexi* | 1.44 | 2.30 | 2.92 | 3.22 | 3.90 | 2.98 | 2.74 | 2.74 | 13.81 | 1.21 |
| *Cyanobacteria* | 13.76 | 1.96 | 1.50 | 1.83 | 1.48 | 3.04 | 2.37 | 1.59 | 0.71 | 13.81 |
| *Deinococcus-Thermus* | 0.00 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.69 | 0.71 |
| *Euryarchaeota* | 0.00 | 0.27 | 0.00 | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.69 |
| *Firmicutes* | 5.07 | 4.96 | 3.74 | 2.49 | 2.37 | 3.85 | 3.15 | 3.32 | 4.30 | 4.30 |
| *Gemmatimonadetes* | 0.00 | 0.76 | 1.08 | 1.13 | 1.61 | 1.05 | 1.16 | 1.35 | 0.77 | 0.77 |
| *Nitrospirae* | 0.00 | 0.34 | 0.53 | 0.73 | 0.83 | 0.79 | 0.86 | 0.89 | 0.00 | 0.00 |
| *Planctomycetes* | 1.16 | 2.75 | 2.93 | 4.10 | 4.47 | 6.10 | 6.21 | 3.59 | 1.71 | 4.30 |
| *Proteobacteria* | 42.04 | 45.09 | 40.84 | 35.44 | 36.96 | 38.55 | 35.12 | 36.21 | 38.98 | 0.77 |
| *Thaumarchaeota* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.24 | 0.00 | 0.00 | 0.00 |
| *Verrucomicrobia* | 0.58 | 1.21 | 1.74 | 1.66 | 1.98 | 3.52 | 2.36 | 4.95 | 0.53 | 0.53 |

***Table 4.6:*** *Percentage classified and unclassified reads for each metagenome at the genus level, following read based taxonomic annotation in Kaiju.*

| **Sample** | **% Classified** | **% Unclassified** |
|---|---|---|
| **1** | 57.95 | 42.05 |
| **2** | 40.31 | 59.69 |
| **3** | 56.89 | 43.11 |
| **4** | 52.44 | 47.56 |
| **5** | 53.17 | 46.83 |
| **6** | 53.13 | 46.87 |
| **7** | 50.99 | 49.01 |
| **8** | 52.44 | 47.56 |
| **9** | 52.5 | 47.5 |
| **10** | 49.44 | 50.56 |
| **11** | 49.65 | 50.35 |
| **12** | 49.7 | 50.3 |
| **13** | 34.82 | 65.18 |
| **14** | 47.73 | 52.27 |
| **15** | 47.8 | 52.2 |
| **16** | 45.86 | 54.14 |
| **17** | 34.56 | 65.44 |
| **18** | 46.04 | 53.96 |
| **19** | 47.02 | 52.98 |
| **20** | 46.75 | 53.25 |
| **21** | 46.34 | 53.66 |
| **22** | 36.71 | 63.29 |
| **23** | 53.17 | 46.83 |

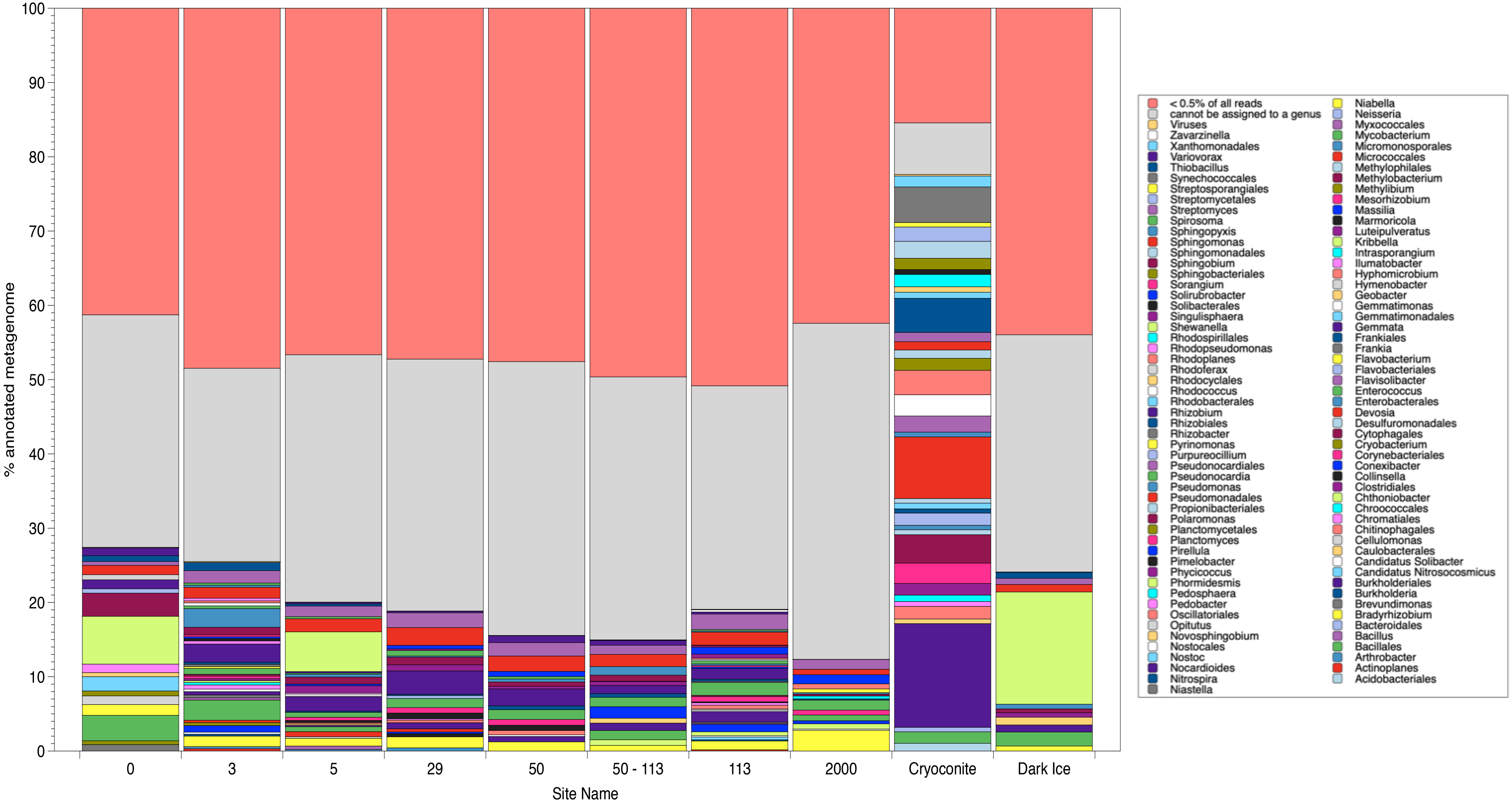


***Figure 4.4:** Read based community composition, expressed as the percentage of annotated reads for Midtre Lovénbreen sample sites at the genus level. Replicate samples for each soil age have been combined. The legend is ordered top to bottom, in line with the graph bars. A tabular format is available in Appendix 3 Table 2.*

### 4.3.4 Genome bins

In total, 83 genome bins were extracted from the metagenomes (Table 4.7). Genome bins were extracted from all metagenomes, except for sample 13. The percentage identity of the extracted bins ranged between 70.5 – 100 % identity to NCBI GenBank (Table 4.7). In total, 21% of bins fell below 80% identity to cultured representatives on NCBI GenBank, with only 8.5% with over 95% identity (Table 4.7). The bins were assigned to organisms with nitrogen metabolisms (16% of bins), sulfur metabolisms (11%) and also cyanobacteria (5%), anaerobes (12%) and plant associated organisms (4%) (Table 4.7). Some BLAST assignments were more prevalent than others, such as *Polaromonas* (5% of bins), *Sphingomonas* (5%), *Conexibacter woesei* (6%), *Sulfuricaulis limicola* (4%) and *Thiobacillus denitrificans* (3%). However, due to the low percentage identity of BLAST matches to publicly available sequences, we cannot be confident in the specificity of the assignments. The extracted bins were annotated with functional genes in RAST, with each bin containing between 144 - 9902 genes (Table 4.7). The genes were matched to known functions using the SEED database, with each bin containing between 126 – 3005 distinct functions (Table 4.7). Binned genomes were grouped with their field replicate samples and a function profile created, based on the SEED ontology. A subset of SEED functions was extracted for investigation, focused on nitrogen cycling pathways.

For each sample (1-23), extracted bins were used in maximum likelihood phylogenies based on a subset of COG groups (Appendix 3, Figure 1 – 22). This was used to audit the BLAST results and to investigate how well bins aligned to KBASE genomes. The trees show 77% of sample bins to be uniquely branching, clustering independently from publicly available KBASE genomes, in agreement with the low percentage identity of BLAST match results (Appendix 3, Figure 1-22).

***Table 4.7:*** *Genome binning results for the 23 assembled metagenomes. For each metagenome, the extracted bins are listed with the top BLASTn hit to NCBI GenBank, along with the RAST annotation results.*

| Blastn Results | | | | | | RAST annotation results | |
|---|---|---|---|---|---|---|---|
| **Sample** | **Acession No** | **% identity** | **E value** | **Max Score** | **Species** | **Number of genes** | **SEED Functions** |
| **1** | | | | | | | |
| Bin 1 | CP002355.1 | 94.81 | 0 | 5243 | *Sulfuricurvum kujiense* | 2849 | 1070 |
| Bin 2 | CP024785.1 | 92.15 | 0 | 13771 | *Nostoc flagelliforme* | 3440 | 1157 |
| Bin 3 | CP003178.1 | 76.26 | 0 | 1629 | *Niastella koreensis* | 1421 | 648 |
| Bin 4 | CP011131.1 | 79.79 | 0 | 1321 | *Lysobacter gummosus* | 1362 | 884 |
| Bin 5 | CP026692.1 | 88.52 | 0 | 3860 | *Nostoc* sp. | 806 | 373 |
| Bin 6 | CP000116.1 | 86.93 | 0 | 3068 | *Thiobacillus denitrificans* | 1992 | 867 |
| Bin 7 | CP019508.1 | 91.72 | 0 | 2747 | *Brevundimonas* | 1194 | 867 |
| Bin 8 | CP003614.1 | 89.43 | 0 | 2645 | *Oscillatoria nigro-viridis* | 1177 | 515 |
| Bin 9 | CP027482.1 | 81.53 | 0 | 4942 | *Aeromicrobium* | 2224 | 960 |
| Bin 10 | CP000116.1 | 83.22 | 0 | 3788 | *Thiobacillus denitrificans* | 3553 | 1161 |
| **2** | | | | | | | |
| Bin 1 | CP025581.1 | 80.17 | 0 | 3472 | *Nocardioides* | 3005 | 1050 |
| Bin 2 | CP001854.1 | 92.76 | 0 | 3864 | *Conexibacter woesei* | 162 | 73 |
| **3** | | | | | | | |
| Bin 1 | AP014879.1 | 84.39 | 0 | 4165 | *Sulfuricaulis limicola* | 2625 | 1140 |
| Bin 2 | CP012573.1 | 84.04 | 0 | 6135 | *Clavibacter capsici* | 1329 | 751 |
| Bin 3 | AP012057.1 | 80.85 | 0 | 2261 | *Ilumatobacter coccineus* | 2022 | 815 |
| Bin 4 | CP010554.1 | 87.37 | 0 | 2030 | *Rugosibacter aromaticivorans* | 1060 | 563 |
| Bin 5 | CP012371.1 | 80.27 | 0 | 2462 | *Nitrosospira briensis* | 1535 | 798 |
| Bin 6 | CP015079.1 | 83.56 | 0 | 2970 | *Nocardioides dokdonensis* | 2748 | 992 |
| Bin 7 | AP014879.1 | 86.33 | 0 | 3921 | *Sulfuricaulis limicola* | 1194 | 593 |
| Bin 8 | CP000116.1 | 84.61 | 0 | 1947 | *Thiobacillus denitrificans* | 841 | 468 |
| **4** | | | | | | | |
| Bin 1 | CP015732.1 | 94.33 | 0 | 3085 | *Arthrobacter* | 3212 | 1057 |
| Bin 2 | CP000529.1 | 90.13 | 0 | 3720 | *Polaromonas naphthalenivorans* | 1144 | 612 |
| **5** | | | | | | | |
| Bin 1 | AP014879.1 | 76.92 | 0 | 2534 | *Sulfuricaulis limicola* | 2609 | 996 |
| Bin 2 | CP031145.1 | 92.32 | 0 | 4575 | *Intrasporangium calvum* | 1574 | 581 |
| Bin 3 | CP009111.1 | 85.94 | 0 | 1930 | *Rhodococcus opacus* | 1574 | 1073 |
| Bin 4 | CP001854.1 | 78.26 | 0 | 2145 | *Conexibacter woesei* | 2031 | 787 |
| **6** | | | | | | | |
| Bin 1 | CP002399.1 | 83.06 | 0 | 3013 | *Micromonospora* | 5386 | 11259 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| Bin 2 | CP004036.1 | 70.05 | 2.00E-45 | 196 | *Sphingomonas* | 4866 | 1200 |
| Bin 3 | CP002199.1 | 90.2 | 2.00E-06 | 67.6 | *Cyanothece* | 3388 | 940 |
| Bin 4 | CP006005.1 | 93.33 | 9.00E-07 | 67.6 | *Vibrio parahaemolyticus* | 2224 | 272 |
| **7** |  |  |  |  |  |  |  |
| Bin 1 | CP002994.1 | 82.83 | 2.00E-13 | 89.8 | *Streptomyces violaceusniger* | 2678 | 861 |
| Bin 2 | CP021181.1 | 74.87 | 1.00E-78 | 307 | *Sphingomonas wittichii* | 3426 | 1086 |
| Bin 3 | CP012184.1 | 77.19 | 7.00E-157 | 566 | *Pseudonocardia* | 3089 | 905 |
| Bin 4 | CP011339.1 | 81.37 | 2.00E-11 | 84.2 | *Microcystis panniformis* | 2315 | 726 |
| **8** |  |  |  |  |  |  |  |
| Bin 1 | CP015732.1 | 94.33 | 0 | 3085 | *Arthrobacter* | 3212 | 1057 |
| Bin 2 | CP000529.1 | 90.13 | 0 | 3720 | *Polaromonas naphthalenivorans* | 1130 | 611 |
| **9** |  |  |  |  |  |  |  |
| Bin 1 | LT827010.1 | 95.24 | 0 | 2918 | *Actinoplanes* | 2996 | 1041 |
| Bin 2 | CP001854.1 | 92.76 | 0 | 3864 | *Conexibacter woesei* | 153 | 67 |
| **10** |  |  |  |  |  |  |  |
| Bin 1 | CP014989.1 | 83.94 | 0 | 3241 | *Serinicoccus* | 1303 | 742 |
| Bin 2 | CP010954.1 | 79.31 | 0 | 1493 | *Sphingobium* | 1966 | 924 |
| Bin 3 | CP000316.1 | 83.2 | 0 | 2499 | *Polaromonas* | 4632 | 791 |
| Bin 4 | CP030865.1 | 85.71 | 5.00E-46 | 198 | *Micromonospora* | 1900 | 760 |
| Bin 5 | CP001124.1 | 81.21 | 0 | 1123 | *Geobacter bemidjiensis Bem* | 2919 | 1098 |
| Bin 6 | CP000533.1 | 91.78 | 0 | 3690 | *Streptomyces lunaelactis* | 1619 | 552 |
| Bin 7 | CP026304.1 | 94.06 | 0 | 3241 | *Polaromonas naphth.* | 4633 | 1049 |
| Bin 8 | CP000531.1 | 96.7 | 0 | 1358 | *Polaromonas naphth.* | 852 | 381 |
| **11** |  |  |  |  |  |  |  |
| Bin 1 | FO117623.1 | 91.54 | 0 | 2510 | *Blastococcus saxobsidens* | 2136 | 791 |
| Bin 2 | CP006644.1 | 92.08 | 0 | 2643 | *Sphingomonas sanxanigenens* | 1291 | 516 |
| **12** |  |  |  |  |  |  |  |
| Bin 1 | FO117623.1 | 91.65 | 0 | 2521 | *Blastococcus saxobsidens* DD2 | 2142 | 772 |
| Bin 2 | CP006644.1 | 92.08 | 0 | 2643 | *Sphingomonas sanxanigenens* | 1285 | 515 |
| **14** |  |  |  |  |  |  |  |
| Bin 1 | CP029343.1 | 91.03 | 0 | 4739 | *Massilia oculi* | 3661 | 1109 |
| Bin 2 | CP029343.1 | 93.05 | 0 | 5033 | *Massilia oculi* | 2665 | 1124 |
| **15** |  |  |  |  |  |  |  |
| Bin 1 | CP027775.1 | 95.55 | 0 | 2837 | *Clostridium botulinum* | 427 | 136 |
| Bin 2 | LT827010.1 | 95.41 | 0 | 2942 | *Actinoplanes* | 2138 | 798 |
| **16** |  |  |  |  |  |  |  |
| Bin 1 | CP027775.1 | 95.55 | 0 | 2837 | *Clostridium botulinum* | 429 | 135 |
| Bin 2 | CP002479.1 | 75.86 | 2.00E-21 | 117 | *Geobacter* | 2137 | 795 |
| **17** |  |  |  |  |  |  |  |
| Bin 1 | CP009241.1 | 91.3 | 0.00004 | 62.1 | *Paenibacillus* | 2394 | 1006 |
| Bin 2 | CP009241.1 | 91.3 | 0.00004 | 62.1 | *Paenibacillus* | 3534 | 891 |
| Bin 3 | CP009571.1 | 81.75 | 0 | 734 | *Sphingomonas taxi* | 6927 | 1083 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Bin 4 | CP031968.1 | 80 | 9E-10 | 78.7 | *Chromobacterium rhizoryzae* | 1495 | 600 |
| **18** | | | | | | | |
| Bin 1 | CP031146.1 | 92.86 | 5E-11 | 82.4 | *Pseudomonas plecoglossicida* | 2094 | 830 |
| Bin 2 | AP014946.1 | 76.36 | 0 | 2464 | *Variibacter gotjawalensis* | 2839 | 1258 |
| Bin 3 | CP001854.1 | 76.88 | 9E-143 | 520 | *Conexibacter woesei* | 5058 | 958 |
| Bin 4 | CP021904.1 | 100 | 0.000003 | 65.8 | *Alkalitalea saponilacus* | 1170 | 499 |
| Bin 5 | CP020569.1 | 85.71 | 0 | 2311 | *Streptomyces gilvosporeus* | 5924 | 1015 |
| Bin 6 | CP026952.1 | 81.25 | 5E-11 | 82.4 | *Aeromicrobium* | 5313 | 1191 |
| Bin 7 | CP011271.1 | 78.03 | 0 | 983 | *Gemmata* | 5397 | 1016 |
| Bin 8 | CP011132.1 | 82.42 | 3E-10 | 80.5 | *Citrobacter amalonaticus* | 3734 | 910 |
| **19** | | | | | | | |
| Bin 1 | CP001854.1 | 87.68 | 0 | 1101 | *Conexibacter woesei* | 331 | 138 |
| Bin 2 | CP001389.1 | 76.65 | 1E-27 | 137 | *Sinorhizobium fredii* | 188 | 93 |
| **20** | | | | | | | |
| Bin 1 | CP022522.1 | 97.56 | 6E-08 | 71.3 | *Pseudoalteromonas* | 144 | 56 |
| Bin 2 | CP018171.1 | 89.74 | 7E-32 | 150 | *Mesorhizobium oceanicum* | 181 | 83 |
| **21** | | | | | | | |
| Bin 1 | CP011491.1 | 78.61 | 3E-55 | 228 | *Mycolicibacterium vaccae* | 673 | 243 |
| Bin 2 | CP001197.1 | 80.31 | 1E-44 | 193 | *Desulfovibrio vulgaris* | 432 | 172 |
| **22** | | | | | | | |
| Bin 1 | AP017308.1 | 80.64 | 0 | 2817 | *Leptolyngbya* sp. | 2239 | 1026 |
| Bin 2 | CP016768.2 | 90.99 | 0 | 4482 | *Candidatus Nanopelagicus limnes* | 9902 | 636 |
| Bin 3 | CP000494.1 | 82.08 | 0 | 1205 | *Bradyrhizobium* | 2663 | 894 |
| Bin 4 | CP016768.2 | 79.64 | 0.00E+00 | 2198 | *Candidatus Nanopelagicus limnes* | 1948 | 707 |
| Bin 5 | CP016282.1 | 73.68 | 3.00E-127 | 468 | *Cryobacterium arcticum* | 1266 | 534 |
| **23** | | | | | | | |
| Bin 1 | AP014879.1 | 76.92 | 0 | 2534 | *Sulfuricaulis limicola* | 2608 | 995 |
| Bin 2 | CP031145.1 | 92.32 | 0.00E+00 | 4575 | *Intrasporangium calvum* | 1584 | 583 |
| Bin 3 | CP009111.1 | 85.94 | 0 | 1930 | *Rhodococcus opacus* | 3320 | 1072 |
| Bin 4 | CP001854.1 | 78.26 | 0.00E+00 | 2145 | *Conexibacter woesei* | 2029 | 784 |

For the binned genomes at each sampling site, a SEED function profile was created for nitrogen cycling pathways. This profile highlights the percentage of protein coding genes attributed to each functional classification. Whist the functional profiles were not exhaustive of all microbes (due to inadequate sequencing depth in some samples and incomplete metagenome assembly), it does show the most abundant nitrogen cycling pathways in the genomes that were recovered from the binning process. Additionally, whist the binning did not sample all genes in each binned genome due to inadequate coverage, it exemplifies the potential of this approach for deeply sequenced samples, highlighting genome functionality.

The SEED classifications for nitrogen cycling pathways are available in Figure 4.5. The pathways include nitrogen fixation, nitrification, denitrification and ammonia assimilation (Figure 4.5). Nitrogen fixation was present in extracted genomes up to 50 years and was not detected in binned genomes from cryoconite and dark ice samples (Figure 4.5). However, ammonia assimilation was found in all samples, reaching 0.63% of recovered protein coding genes in the cryoconite sample (Figure 4.5). Denitrification, nitrite and nitrate ammonification and dissimilatory nitrite reductase were largely found in genome bins throughout the forefield, however were not detected in bins from 2000 year soils (Table 4.3; Figure 4.5). Ammonia assimilation and ammonification were the most prevalent nitrogen cycling pathways, accounting for up to 0.63% and 0.29% of protein coding genes, respectively (Figure 4.5). Whist SEED genome annotation can't be used to profile the complete nitrogen cycling pathways in the metagenomes (due to inadequate sequencing and assembly), it does provide an insight into the functional potential of abundant genomes.

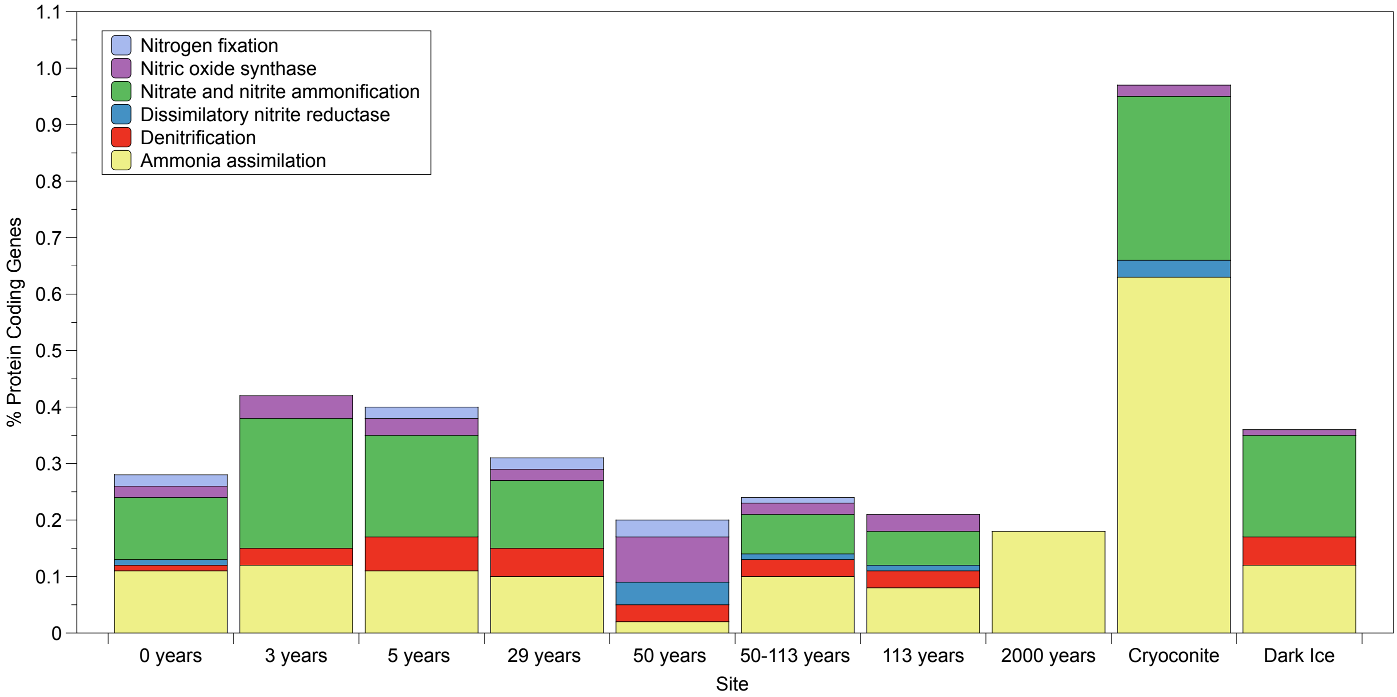


***Figure 4.5:*** *SEED function profile for nitrogen metabolism in genome bins across sampling sites. The percentage of protein coding genes attributed to each pathway is shown.*

## 4.4 Discussion

### 4.4.1 Total nitrogen and total organic carbon

Soil carbon and nitrogen content were investigated to understand changes to soil nutrient content along the chronosequence. This provides important contextual information from which hypotheses regarding community composition and functional shifts can be investigated. The soil total nitrogen and total organic carbon content revealed changes in soil nutrient content with succession. As shown by Figure 4.3, the highest TN and TOC values were found in the soils which were furthest from the glacier terminus and therefore had been exposed for the longest duration. This is in agreement with research by Turpin-Jelfs *et al.,* (2019) who found both carbon and nitrogen content to increase in soils along the Midtre Lovénbreen forefield. This may be because these soils have been exposed by glacier retreat and have been subject to soil development, in line with our hypothesis (Bradley *et al*., 2014). TOC accumulates in forefield soils from allochthonous and autochthonous sources, such as through deposition of soot and organic matter, alongside the action of autotrophic microbial communities (Guelland *et al.,* 2013; Bradley *et al.,* 2014). The increased TOC content observed here in older soils is consistent with studies from other forefields, such as the Damma Glacier (Switzerland) and has been represented in SHIMMER, a numerical model of glacier forefield succession (Guelland *et al*., 2013; Bradley *et al*., 2016). The increase in TOC content with soil age is a crucial factor in soil development and facilitates colonization by plants and heterotrophic microbes (Schutte *et al*., 2009). This largely explains why initial (newly exposed) soils typically support minimal vegetation and a greater plant density can be observed in the later more developed soils (Tscherko *et al*., 2005). Variance within the forefield may relate to disturbances such as water flow pathways or variance in soil type (Bekku *et al.,* 2004; Frey *et al.,* 2013). Whilst older soils contained the highest TOC content, initial soils (0 – 5 years) were also found to contain detectable levels of TOC, between 3.89 – 6.68 mg $g^{-1}$ (Table 4.3). The presence of TOC in newly exposed soils has been previously attributed to ancient subglacial organic carbon and aeolian deposition, which may be crucial for supporting initial heterotrophic microbial communities, prior to the establishment of autotrophs (Hodkinson *et al.,* 2002; Schulz *et al*., 2013).

The majority of sites across the forefield had TN levels below detectable limits (Table 4.3). As nitrogen is crucial for microbial and plant protein synthesis, minimal levels are indicative of low nutrient conditions and potential nitrogen limitation (Turpin-Jelfs *et al.,* 2019). Typically, nitrogen is supplied to forefield soils through autochthonous sources such as microbial nitrogen fixation or remineralization, alongside allochthonous sources such as snowmelt or aeolian deposition (Bradley *et al.,* 2014). In general, studies have shown that microbial nitrogen fixation (diazotrophy) is the most important source in early sites, helping to facilitate the establishment of higher microbes and plants (Brankatschk *et al.,* 2011). The lack of detectable nitrogen in the early soils may relate to an inactivity of diazotrophs or a tight coupling between nitrogen fixation and consumption in soils. Low TN values are often observed in recently deglaciated forefields due to a lack of soil development, with studies across forefields from Canada, Antarctic, Svalbard, Austria, Italy and Switzerland finding values ranging between 0.1 – 2 mg $g^{-1}$ (Bradley *et al.,* 2014). However, at a soil age of 2000 years a TN content of 3.89 – 4.89 mg $g^{-1}$ was detected (Figure 4.3). This is consistent with the expected increase in TN with soil development, in line with the TOC results. Whilst the source of the additional TN cannot be isolated, it may come from a combination of diazotrophy, nitrogen remineralization or increased aeolian deposition (Bradley *et al.,* 2014).

### 4.4.2 Microbial community composition

A chronosequence of metagenomes were evaluated along the forefield, to test our hypothesis that microbial community composition and function would change alongside soil succession and nutrient pool development. We hypothesized autotrophic organisms would be dominant in early soils, with more heterotrophic nutrient cycling appearing later in the chronosequence, relating to nutrient availability.

Throughout all metagenomes, the read based taxonomic annotation shows both aerobic and anaerobic organisms can be identified, with no sites containing exclusively anaerobes or microaerophilic organisms (Figure 4.4). Aerobes identified include organisms such as *Niabella, Variovorax, Rhodococcus* and *Pseudomonas* (Figure 4.4). Anaerobic organisms recovered from the metagenomes include *Rhodoferax, Opitutus, Geobacter* and *Bacillus* (Figure 4.4). The presence of strict and facultative anaerobes in forefield soils has been previously attributed to the periodic flushing by subglacial meltwater (Duc *et al.,* 2009). However, as anaerobic strains have been identified in each site, this may indicate anoxic or microaerophilic micro environments, alongside the aerobic soil surface layer. This could be attributed to poor drainage of glacial meltwater. Anoxic environments are often found in poorly drained subglacial systems, and therefore these anaerobes may be a legacy from the subglacial environment (Wadham *et al.,* 2008; Boyd *et al.,* 2010).

Extremophilic and psychrophilic organisms were recovered from the metagenomes, alongside those with traits to aid colonization and survival in the cold, high UV, oligotrophic conditions typical of Arctic forefields (Figure 4.4). Sequences attributed to *Bacillus* were identified in soils at 5 years following glacier retreat (Figure 4.4). This extremophile is endospore forming and is resistant to both cold and desiccation, with psychrophilic, acidophilic, alkaliphilic, halotolerant, and halophilic properties (Joan *et al*., 2011). The production of endospores allows Bacillus to remain in a dormant state under stressful conditions (Nicholson *et al.,* 2000). *Bacillus* are typically found in extreme environments, such as deserts and Arctic soils, due to their high resistance to environmental stressors (Rüger *et al.,* 2000). In addition, *Candidatus Solibacter* and *Pseudomonas* were recovered from sites ranging from 3 – 2000 years since exposure (Figure 4.4). These organisms secrete extracellular polymeric substances (EPS), consisting of polysaccharides, proteins, lipids and DNA (Yang *et al.,* 2011). This EPS creates a biofilm surrounding the microbes, helping to reduce temperature and nutrient fluctuations (Flemming and Wingender, 2010). This is a crucial mechanism for environmental tolerance, particularly to cold and desiccation stressors in the Arctic. Bacteria with cold-tolerance mechanisms were also identified in the forefield metagenome samples and dark ice, such as *Cryobacterium* and *Polaromonas* (Figure 4.4). Both species are psychrophiles and found in cold environments such as the Arctic, which may be due to cold tolerance mechanisms such as EPS production and cell membrane fluidity (Irgens *et al.,* 1996; Suzuki *et al.,* 1997).

The cyanobacteria, *Nostoc, Oscillatoria, Phormidesmis* and *Synechococcales* were found in recently exposed soils (0 years) and cryoconite samples, in support for our hypothesis (Figure 4.4). Cyanobacteria have previously been proposed as early colonizers of forefield soils and are often found in cryoconite (Zumsteg *et al.,* 2013; Christner *et al.,* 2003; Edwards *et al*., 2011). Cyanobacteria are photosynthetic and many are diazotrophic, and therefore they do not rely on fixed sources of carbon (and nitrogen) for growth (Mitusi *et al.,* 1986). This is crucial in newly exposed soils and cryoconite on the glacier surface, as labile carbon and nitrogen stocks are typically limited (Christner *et al.,* 2003). Furthermore, cyanobacteria are resilient to environmental stressors such as sub-freezing temperatures, attributed to the production of protective EPS, which buffers temperature, desiccation and pH (de los Rios *et al.,* 2015). The recovery of the mat forming cyanobacteria, *Oscillatoria*, from the cryoconite metagenome is in agreement with the surrounding literature, which highlights it as one of the most abundant cryoconite cyanobacteria (Müller *et al*.,2015; Edwards *et al.,* 2011). Cyanobacterial sequences were not recovered from more developed soil sites, however this may relate to inadequate sampling, as including more sample sites and DNA extraction from rock biofilms may highlight the presence of cyanobacteria. Additional deep sequencing of samples would help to highlight less abundant microbes in the metagenome samples, however the results show that cyanobacteria are not a dominant fraction of microbial communities in older soil samples. This reflects the results of Zumsteg *et al.,* (2013), who found cyanobacteria to decline with increasing soil age across the Damma Glacier forefield, Switzerland.

The microbial community composition of newly exposed soils has previously been the subject of scientific debate. The presence of heterotrophic microbial populations in early colonizer communities is not consistent between forefields. Some studies have identified heterotrophic colonizers in newly exposed soils, whilst others suggest autotrophic bacteria comprise the bulk of the initial community composition (Hodkinson *et al.,* 2002; Kastovska *et al.,* 2005). The composition of colonizer communities in this study (Site 0 years) contained both autotrophs (*Nostoc, Phormedesmis, Rhodoferaz, Thiobacillus*) and heterotrophs (Gemmata, Rhodospirillales, Sphingopyxis) (Figure 4.4). This may relate to the presence of low, but detectable TOC at this site, of 6.68 mg $g^{-1}$ (Table 4.3). This TOC may have been sourced from aeolian deposition or from the subglacial environment (Kastovska *et al.,* 2005; Bardgett *et al.,* 2007). The presence of TOC can provide a labile carbon pool for heterotrophic colonization (Kastovska *et al.,* 2005). However, as the TOC content of newly exposed soils will vary between glaciers, based on the content of overridden material and deposited material, the presence of heterotrophs is likely to be inconsistent between glaciers.

Furthermore, runoff from the ice surface and cryoconite may be a key fertilizing mechanism for newly exposed soils (Kastovska *et al.,* 2005; Stibal *et al.,* 2006). This dispersal may be due to the drainage of supraglacial meltwater through moulins to the glacier terminus (Stibal *et al*., 2006). In our study site, autotrophic and heterotrophic microbial populations found in the cryoconite or dark ice sample were also found in newly exposed soils (0 years), including *Nostoc, Thiobacillus, Sphingomonas, Gemmata, Frankia, Corynebacteriales, Bacteriodiales* and *Rhodospirillales* (Figure 4.4). Consequently, ice surface microbial populations may be important for facilitating the colonization and composition of pioneer microbial communities. Again, this is largely a function of aeolian deposition, as glacier ice is most often fertilized by microbial populations in wind driven material (Marshall and Chalmers, 1997). Additionally, the composition of subglacial microbial communities may mediate the composition of initial colonizer communities, for example supplying anaerobic or sulfur cycling organisms, however subglacial samples were not available for validation in this analysis. Consequently, the composition of pioneer microbial communities is likely to vary between glaciers, influenced by the content of depositional material, ice microbial surface populations, subglacial communities and the nutrient content of initial soils.

### 4.4.3 Microbial metabolisms – Carbon

A wide range of microbial metabolic potential was identified in the forefield and glacier metagenomes, including chemolithoautotrophs, organoheterotrophs, chemoheterotrophs and chemolithotrophs, involved in carbon, nitrogen and sulfur cycling (Figure 4.4). This identifies microbial communities to have a potential influence on a multitude of biogeochemical cycles in forefield soils. Heterotrophic organisms potentially involved in microbial carbon cycling were found in soils 3 years following glacier retreat (Figure 4.4). This includes organisms such as *Micromonospora*, which grows off decaying organic matter in soils (White *et al.,* 1996). Additionally, bacteria capable of utilizing multiple carbon sources, including aromatic compounds, were recovered, such as *Rhodococcus opacus*, *Sphingomonas* and *Rugosibacter aromaticivorans* (Figure 4.4). The presence of carbon cycling microbes in sites above 3 years may relate to the minimal carbon content in newly exposed soils (Figure 4.3). Although TOC does not increase significantly from 0 years to 3 years, deposition of aromatic hydrocarbons from fossil fuel combustion and carbon fixation by phototrophic cyanobacteria, may provide sufficient labile carbon for heterotrophic carbon cycling bacteria to colonize (Margesin *et al.,* 2003; Frey *et al.,* 2013). The importance of phototrophs in facilitating the colonization of heterotrophic microbes and plants in forefield soils is supported by the surrounding literature, highlighting the important source of carbon phototrophs provide (Kastovska *et al.,* 2005; Frey *et al.,* 2013). Consequently, this provides support for our hypothesis that heterotrophic cycling would appear in later sites, due to nutrient availability.

### 4.4.4 Microbial metabolisms – Nitrogen

Microbes with soil nitrogen cycling potential were also highlighted using the read based taxonomic annotation. Nitrogen-fixing cyanobacteria were recovered from 0 year soils and cryoconite metagenomes using read annotation and genome binning, including *Nostoc, Oscillatoria* and *Leptolyngbya* (Figure 4.4; Table 4.7). Nitrogen-fixingcyanobacteria may be fundamental for the build-up of labile nitrogen in nutrient deplete initial soils, as they convert nitrogen gas ($N^2$) to ammonium ($NH_4^+$) (Duc *et al.,* 2009; Figure 4.1). These cyanobacteria may therefore be important facilitators for the colonization of heterotrophs, considering the low nitrogen stocks observed at site 0 years (Figure 4.3). This is supported by SEED function profiles, which identified the nitrogen fixation pathway to be present in soils up to 50 years (Figure 4.5). This provides support for our hypothesis that autotrophic nitrogen cycling microbes would be found in early soils, due to low nutrient availability. The findings are supported by Turpin-Jelfs *et al.,* (2019) who found biological nitrogen fixation to occur in newly exposed soils in the Midtre Lovénbreen forefield, and decreased long the chronosequence.

Our results show the presence of cyanobacteria in cryoconite samples.  As these microbes can fix carbon and often nitrogen, they are able to withstand the oligotrophic conditions on the glacier surface (Stibal *et al.,* 2006; Cameron *et al.*, 2012). The flushing of meltwater from cryoconite, through moulins to the glacier terminus, may be an important dispersal mechanism, alongside wind drift, for cyanobacteria to reach newly exposed soils (Mueller *et al.,* 2001). Interestingly, the nitrogen fixation pathway was not detected in genome bins from the cryoconite sample, however this may relate to poor recovery of diazotrophic (nif) genes during sequencing due to EPS interference with DNA extraction, or a reduced need for nitrogen fixation (Figure 4.5). However, data on sediment nitrogen content was not available for this site. Root associated diazotrophs were found in metagenome samples exposed for 113 years or more (Figure 4.4). These organisms are symbiotic, fixing nitrogen in return for organic carbon substrates from plant roots (Franche *et al.,* 2009). The root associated diazotrophs recovered from read annotation and genome binning included *Mesorhizobium oceanicum*, *Bradyrhizobium, Sinorhizobium fredii* and *Chromobacterium rhizoryzae* (Figure 4.4; Table 4.7). The identification of rhizobia in older soil sites is consistent with the colonization of plants at later stages of soil development (Knelman *et al.,* 2012). However, *Bradyhizobium* were detected in cryoconite samples, despite a lack of vascular plants (Figure 4.4). This may be the result of wind dispersal and bird droppings, common mechanisms for fertilizing the glacier surface with microorganisms (Kastovska *et al.,* 2005).

Bacteria related to nitrification (*Nitrospira briensis*) and denitrification (*Conexibacter woesei*, *Thiobacillus denitrificans*) were also found in soils after 3 years of exposure (Figure 4.4; Table 4.7). *Nitrospira briensis* are involved in ammonia oxidation, the first step in nitrification, oxidizing ammonia ($NH_3$) to Nitrite ($NO_2^-$) (Figure 4.1; Teske *et al.,* 1994; Daims *et al.,* 2015). As nitrifying bacteria rely on labile ammonia, they may only occur after 3 years of soil exposure, when sufficient stocks have accumulated from deposition, ammonification and nitrogen fixation. This is supported by Brankatschk *et al.,* (2011) who found low potential nitrification and denitrification rates at initial forefield sites, attributed to the lack of available nitrate and ammonium (Brankatschk *et al.,* 2011). However, if this is the case, the production or deposition of labile ammonia must be tightly coupled to uptake, given the minimal TN values recorded in our sampling sites (Table 4.3). Denitrifying bacteria are largely responsible for the loss of labile nitrogen from soil environments. Denitrification involves the reduction of fixed nitrate ($NO_3^-$) to nitrogen gas ($N_2$) by anaerobic or facultatively anaerobic bacteria (Figure 4.1; Aulakh *et al.,* 1992). Through denitrification, nitrate/nitrite is used as the terminal electron acceptor in respiration, in the absence of oxygen (Simon *et al.,* 2009). Denitrifying bacteria were recovered from soils exposed for 3 years or more and from the dark ice sample (Figure 4.4). This may relate to the requirement of labile nitrate, hindering the occurrence in newly exposed soils (0 years) (Brankatschk *et al.,* 2011). This supports the research of Kandeler *et al.,* (2006), who found evidence of denitrifying communities developing under soil succession, driven by an increasing availability of organic substrates. Denitrification is common in soils and may occur along the Midtre Lovénbreen forefield in wet or waterlogged conditions, where the oxygen supply is limited (Christensen *et al.,*1990). The presence of *Conexibacter woesei* in the dark ice metagenome indicates the potential for denitrification on the glacier surface (Figure 4.4). Denitrifying bacteria have been found in anoxic glacier ice and basal samples where oxygen is limited, including the Midtre Lovénbreen subglacial environment (Hodson *et al.,* 2005; Simon *et al.,* 2009; Ansari *et al.,* 2013). This is also supported by the SEED function profile for genome bins, which found denitrification to occur throughout the forefield and in the dark ice sample (Figure 4.5). Interestingly, the nitrification and denitrification pathways were not found in soils aged 2000 years in the SEED profiles. This may be due to the genome sampling that was carried out (i.e. based on assembled, binned, genomes) and therefore is not as comprehensive in terms of diversity as the read based taxonomic annotation. The results presented here therefore support the hypothesis that heterotrophic nitrogen cycling would appear in later soil sites, requiring an initial buildup of labile nitrogen stocks.

### 4.4.5 Microbial metabolisms – Sulfur

Evidence for the potential of microbial sulfur cycling was also found in the forefield metagenomes and cryoconite samples. Sulfur oxidizing microbes were found in early soils, aged 0 and 5 years, alongside the cryoconite sample by both read annotation and genome binning and included *Thiobacillus denitrificans, Sulfuricurvum kujinse* and *Sulfuricaulis limicola* (Figure 4.4; Table 4.7). Sulfur oxidation is the process by which elemental ($S^0$) or reduced ($H_2S$, $HS^-$) sulfur is oxidized to sulfate ($SO_4^{2-}$), coupled with the reduction of oxygen (aerobic) or nitrate (anaerobic) (Eriksen *et al.,* 1998). This mechanism is used for energy production by chemolithotrophic sulfur oxidizers and has been identified in subglacial systems (Wainright 1978; Bottrell and Tranter, 2002). In addition, evidence for microbial sulfate reduction was provided by the recovery of *Desulfovibrio vulgari*s and *Geobacter bemidijensis Bem* sequences from more developed soil sites (aged 50 – 2000 years) (Figure 4.4; Table 4.7). Sulfate reducers use sulfate ($SO_4^{2-}$) as a terminal electron acceptor for anaerobic respiration, reducing it to hydrogen sulfide ($H_2S$) and aid the degradation of organic matter (Eriksen *et al.,* 1998). The presence of anaerobic sulfur reducers is indicative of anoxic micro-environments along the chronosequence, for example waterlogged soils, where oxygen availability is limited. Again, the presence of sulfate reducers may be a legacy from the subgalcial environment, however, additional sampling of subglacial sediments would be needed to validate this. The location of sulfur reducing bacteria in more developed forefield soils compared to sulfur oxidizers, is interesting. An explanation for this is that sulfate reduction requires the availability of $SO_4^{2-}$, produced by oxidation, and therefore cannot occur in newly exposed nutrient deplete soils. However, additional measurements of sulfur species would be needed to validate this. Whilst sulfur reducing bacteria may be present in early soil samples, they were not recovered by the read based annotation carried out. Consequently, sulfur reducing microbes do not constitute a major fraction of the microbial community composition at early stages of soil succession in the Midtre Lovénbreen forefield. Additional research on the transcription of sulfur cycling genes (using transcriptomics) would be beneficial to highlight if the sulfur oxidation and reduction pathways were indeed active in the forefield.

### 4.4.6 Unique genomes

Genome binning was carried out to investigate the presence of novel genomes in forefield soils, as expected by our hypothesis. The results of this analysis have indicated a degree of novelty of genomes extracted during the binning process, with 21% of bins falling below 80% identity to NCBI GenBank (Table 4.7). Of these bins, three matched most closely to the denitrifying bacteria, *Conexibacter woesei* and two were associated with the sulfur oxidizing *Sulfuricaulis limicola (*Table 4.7*).* This indicates the samples may contain unique species with ecological importance in local biogeochemical cycles, for example in nitrogen and sulfur cycling. The low percentage identity of these bins is related to using incomplete genome databases to perform taxonomy assignments (Albertsen *et al.,* 2013). As the complete global microbial diversity has not been captured on genome databases, it is likely that some samples may not match closely to the genomes that are available (Albertsen *et al.,* 2013). In the future, developments in NGS technologies, analysis pipelines and methodologies such as metagenomics will allow more environmental microbial diversity to be profiled in online databases. However, the results of this analysis do provide scope for further single cell sequencing and culture based studies, to fully ascertain the novelty of genomes found during the binning process.

## 4.5 Conclusions

This study has used metagenomics to explore the microbial community composition along a chronosequence of soil succession in the Midtre Lovénbreen forefield, Svalbard. Prior to this study, most research on forefields had focused on plant succession, changes to soil physicochemical properties or utilized 16s rRNA or single gene sequencing. For the first time, this study implements metagenomics to understanding the succession of microbial communities in a glacial forefield. We aimed to test several hypotheses: (1) community composition and function would modify with soil development; (2) autotrophic microbes would be present in newly exposed soils, with heterotrophic nutrient cycling occurring in later sites; and (3) extremophilic or novel microbes may be identified due to the harsh environmental conditions in Arctic forefields.

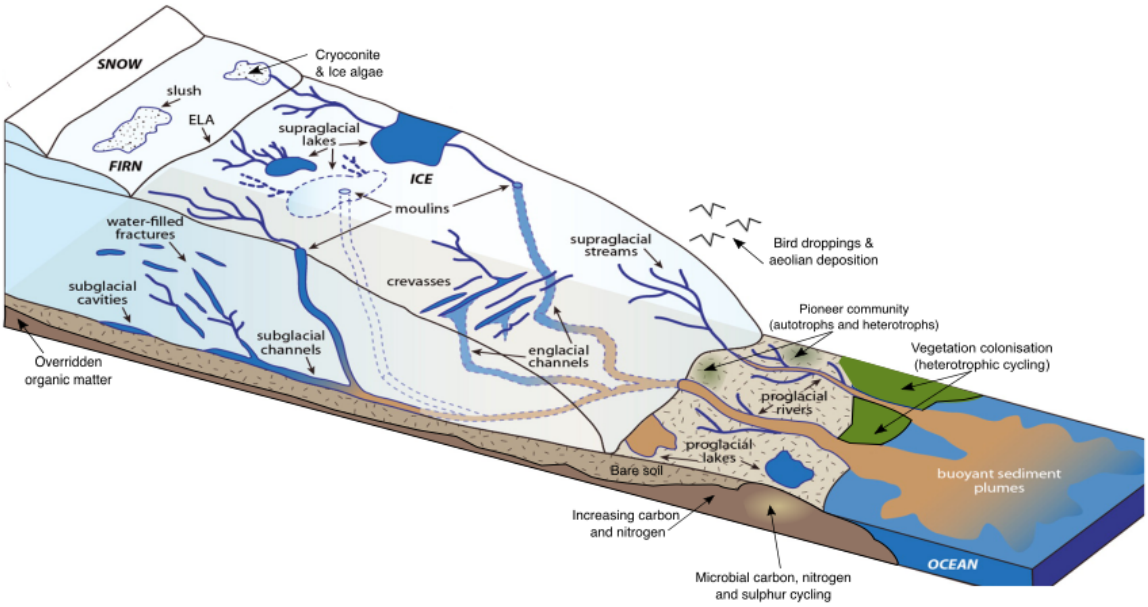


***Figure 4.6:*** *Scematic outlining proglacial features and processes described in this chapter. Key drainage features include: englacial, subglacial and proglacial pathways. Microbial features include: cryoconite and ice algae, the pioneer community and plant colonization. Soil nutrient content is displayed graphically, highlighting overridden organic matter and the buildup of labile organic carbon and nitrogen with succession. Nutrient sources are autochthonous (microbial fixation by autotrophs) and allochthonous (deposition). Source: The figure is an adaptation of Chu et al., (2014).*

The key pathways and processes discussed in this analysis are highlighted in Figure 4.6. This study finds that along the successional chronosequence, both total organic carbon (TOC) and total nitrogen (TN) increased, in line with soil development. This is consistent with previous findings from glacier forefields, which identify TN and TOC pools to rise with soil development due to aeolian deposition, microbial fixation and the establishment of plants. This study also finds a range of extremophilic microbes to be present along the forefield, which may be adapted to the high UV, oligotrophic, cold conditions present, in line with our hypothesis. Additionally, we recovered cyanobacterial sequences from newly exposed soils and cryoconite, in line with previous forefield literature. This may be related to the ability of cyanobacteria fix carbon (and often nitrogen) in low nutrient conditions, alongside the secretion of protective exopolymeric substances (EPS) to buffer the harsh environmental conditions. The composition of the initial pioneer microbial community has previously been the subject of debate, in particular, the prevalence of heterotrophs in newly exposed soils. Our findings conclude that both autotrophic and heterotrophic bacteria were present in the initial pioneer community. This may be attributed to the availability of an initial TOC pool, from overridden soils or aeolian deposition, which heterotrophic bacteria could utilize. Additionally, heterotrophic and autotrophic microbes identified in the pioneer community were also recovered from glacier ice samples. Consequently, we conclude that the community composition of newly exposed soils may be mediated by the microbes supplied in runoff from the glacier surface or from subglacial systems. In response to the debate on pioneer community composition, we suggest the presence of heterotrophic bacteria in newly exposed soils is likely to vary between glaciers, due to differences in the ice surface microbial communities, aeolian deposition and overridden TOC stocks.

The use of metagenome sequencing in this study enabled an investigation into the carbon, nitrogen and sulfur cycling pathways along the chronosequence, through read annotation, genome binning and SEED functional pathway annotation. The use of functional genes provides evidence for microbial biogeochemical cycling potential, which is not possible with taxonomic markers such as 16s rRNA. Our study concludes that forefield soils contain the potential for microbial carbon fixation, heterotrophic carbon cycling, nitrogen fixation, nitrification, denitrification, alongside sulfur oxidation and reduction. We show that microbes associated with heterotrophic carbon, nitrogen and sulfur cycling were recovered from older soils, as these pathways require labile nutrient stocks which may not be present in newly exposed soils, in agreement with our hypothesis. The use of metagenomics in our study also allowed the recovery of discrete genome bins from the samples. Whist the genomes extracted were not comprehensive of the complete diversity (due to incomplete assembly and read coverage), they provide an insight into the functional potential of microbes in forefield samples. Several of these genome bins did not align closely to those on NCBI GenBank, indicating a degree of genome novelty. In particular, several genomes related to denitrifying and sulfur oxidizing microbes were recovered, highlighting these pathways as potentially prevalent in the forefield soils. Whilst extracting complete draft genomes was not the aim of this analysis, due to sequencing limitations, the results demonstrate the potential of metagenomics to explore functionality on a single genome scale, given deep sequencing.

Overall, this study has provided an insight in to the diversity and metabolic potential of microbial communities along a successional chronosequence. It is hoped that this work will help to stimulate further research exploring functional activity rates, gene expression using transcriptomics and probe deeper into genome novelty using single cell sequencing and culture based studies.

## 4.6 Limitations and Future Work

To build on the results of this study, additional deep metagenome sequencing would be beneficial in under sampled sites (Figure 4.2). This would enable less abundant microbes to be recovered and thus fully profile the microbial community composition in each site. Additionally, deep sequencing would provide more read coverage for each microbe, thus enabling a greater number (and more complete) genomes to be extracted during the binning process (Albertsen *et al.,* 2013). Consequently, this may provide the opportunity to build draft metagenome-assembled genomes, enabling an insight into genome structure, diversity and novelty (Albertsen *et al.,* 2013; Hugerth *et al.,* 2015). In particular, building near complete draft genomes would enable a comparison of the average nucleotide identity (ANI) between sample sequences and previously sequenced genomes (Konstantinidis and Tiedje, 2005). This information would be beneficial to fully ascertain how unique our genomes are in comparison to those available on public databases. Draft microbial genomes may also be extracted using culture based studies and single cell sequencing, focused on those genomes highlighted in the binning process (Table 4.7). The benefit of single cell sequencing would be an increased coverage for the target genome, increasing the likelihood of a fully closed genome assembly (Blainey 2013; Shapiro *et al.,* 2013). Additionally, culture based studies may be beneficial, to understand the life cycle and growth conditions of any newly isolated genomes (Marx 2017).

Furthermore, (meta) transcriptomics could also be applied to each sampling site, to highlight which metabolic pathways are active. As transcriptomics involves sequencing the transcribed mRNA, it goes one step further from metagenomics, as it identifies which processes are active at a snapshot in time (Moran *et al.,* 2013). This would be beneficial to understand which aspects of biogeochemical cycles are functioning and any changes along the forefield. Finally, metagenomics supplemented with 16s rRNA analysis may also be useful. Not only would this enable a comparison between the taxonomy conclusions of each method, but the addition of 16s analysis would enable the less abundant fractions of the microbial community to be profiled (Franzosa *et al.,* 2015). The reduced cost and ability of this method to fully profile the community composition enables further diversity metrics and gene abundance calculations to be applied. Finally, a subglacial sample from Midtre Lovénbreen would have been beneficial for this analysis. This would allow a comparison between newly exposed soils, to ice surface and subglacial systems, to identify how influential supraglacial or subglacial microbes are on the composition of the pioneer microbial community.

# Chapter 5: Sequencing the deep - microbial communities recovered from benthic fjord sediments, Chilean Patagonia

Nash, M.V[1]., Anesio, A.M[2]., Barker, G[3]., Wadham, J[1]., Tranter, M[1]., Hawkings, J[4]., Beaton, A[5]., Chin Ng, H[6]., and Sánchez-Baracaldo, P[1].

[1]School of Geographical Sciences, University of Bristol, BS8 1SS, UK
[2] Department of Environmental Science, Aarhus University, PO box 358, Denmark
[3] School of Life Sciences, University of Bristol, BS8 1TQ, UK
[4] Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL 32304, USA
[5] National oceanography Centre, European Way, Southampton, SO14 3ZH, UK
[6] Department of Earth Sciences, University of Bristol, BS8 1RL, UK

### Contributions and acknowledgements

MN carried out the sampling, analysis and interpretation of the sediments data, alongside the writing of the manuscript. AB supplied the underway sampling data. JH, JW and HCN assisted with funding, fieldwork logistics and sampling. PSB and AA assisted with the planning and refinement of the manuscript.

### Funding

This work was carried out as part of the NERC GW4+ Doctoral Training Partnership, awarded to MN. Funding support was provided through a NERC grant NE/J02399X/1, awarded to AA and a Royal Society University Research Fellowship, awarded to PSB. Additional funding support came from the PISCES project and a PISCES Knowledge Exchange Scholarship, awarded to MN.

### 5.1 Introduction

The fjord systems of Chilean Patagonia present a unique habitat for microbial life, draining glacial melt water into marine fjord systems. These fjords have been hypothesized as a hotspot of primary productivity at the land-ocean interface and are crucial for understanding coastal ecosystem functioning and diversity (Iriarte *et al.,* 2007). The fjord systems of Chilean Patagonia host three UNESCO bio-reserves and support commercially important Salmon fisheries (Haussemann and Forsterra, 2009; Niklitscheck *et al.,* 2013). High rates of primary productivity have been observed in Chilean fjords, alongside harboring diverse and unique ecosystems, such as the Patagonian cold-water corals (Iriarte *et al.,* 2007). However, the microbiology of the fjord sediments, crucial for understanding the wider ecosystem functioning, has received limited research attention. These fjords are particularly of interest due to the interaction of glacial meltwater from Patagonia ice fields with marine waters, harboring a range of physico-chemical conditions for microbial life.

The Patagonian ice fields drain into downstream fjords, influencing the fjord's physical characteristics, with implications on primary productivity (PP) (Iriarte *et al.,* 2014). The fjords are also supplied by marine Sub-Antarctic water, which mixes in the inner fjords with freshwater runoff from terrestrial and glacial rivers (Palma and Silva, 2004; Iriarte *et al.,* 2014). The greater density of the marine waters means that a vertical stratification of the water column is often observed (Palma and Silva, 2004). Consequently, Patagonia fjords often contain a surface freshwater lens, over more dense marine waters, promoting a range of dynamic conditions for endemic and novel organisms (Iriarte *et al.,* 2010). Corresponding vertical gradients in oxygen, salinity, nutrients and light have been observed, all of which have implications on PP (Iriarte *et al.,* 2014). Oxygen can reduce to near a hypoxic state with depth, caused by the oxidation of organic matter in the inner fjords (Gonzalez *et al.,* 2013; Iriarte *et al.,* 2014). Horizontal gradients in oxygen have also been observed, reducing towards the inner fjords and glacier outflows (Davila *et al.,* 2002; Silva and Vargas, 2014). This relates to the supply of oxygen from the Sub-Antarctic water, which is depleted by microbial consumption towards the fjord head (Silva and Vargas, 2014). Consequently, deep waters in the inner fjords are more likely to experience anoxic conditions.

Research on the hydrochemistry of Patagonia fjords indicates that the glacial meltwater is low in nutrients, and that mixing plumes from the Sub-Antarctic water may help to stimulate phytoplankton blooms (Arancena *et al.,* 2011; Gonzalez *et al.,* 2013; Montero *et al.,* 2017). Additionally, the inner fjord regions, dominated by glacial runoff, typically have lower nitrate and phosphate than the distal marine waters (Aracena *et al.,* 2011). The glacial runoff is sediment dense, largely attributed to inorganic matter from glacial weathering (Aracena *et al.,* 2011). Alongside the low nutrient content of glacial runoff, the high sediment yield may block out light for photosynthetic microorganisms (Aracena *et al.,* 2011; Landaeta *et al.,* 2012; Meerhoff *et al.,* 2013). Consequently, the biological export from inner fjord regions can be reduced (Silva, 2008). However, the fjords are also fed by terrestrial rivers which contain increased loads of organic matter, silicon, nitrate and phosphate to fjord surface waters (Mayer *et al.,* 1998; Tréguer *et al.,* 2013; Gonzalez *et al.,* 2013).

On average, Patagonian glaciers discharge 70 km$^3$ of freshwater a year into downstream fjords and this flux is increasing with glacier thinning (Lenaerts *et al.,* 2014; Schaefer *et al.,* 2017). The implications of a greater meltwater flux on the microbial communities is hard to ascertain, particularly because fjord PP is a result of complex interactions between light, nutrients and temperature, all of which may be influenced by glacial runoff (Landaeta *et al.,* 2012). For example, increased sediment flux may reduce water column and benthic microbial activity due to increased light attenuation, with implications propagating up the food chain to fish populations (Landaeta *et al.,* 2012; Gonzalez *et al.,* 2013). Additionally, a study by Gutiérrez *et al.,* (2015) suggests that a change in hydrographic conditions from increased glacial meltwater may influence the microbial community structure, favoring freshwater dominant species such as nano and pico plankton. Consequently, a greater understanding of the current microbial community composition and function in these fjords in needed, to help understand how the ecology and ecosystem services may change in the future.

Biological analysis of the fjords has focused on the water column, and in particular, on plankton activity and diversity (Iriarte *et al.,* 2007; Gonzalez *et al.,* 2013). Ecological indicators have classified benthic microbial communities at a ‘good’ status, however communities may be unbalanced due to the high outflow of glacial meltwater into the fjords (Quiorga *et al.,* 2013). The overall water column net primary productivity (NPP) and export production has been shown to increase when moving out from the inner fjords, largely related to the high sediment flux and freshwater in inner fjords (Aracena *et al.,* 2011; Gonzalez *et al.,* 2013). Research using 16s rRNA amplicon sequencing has been carried out on the water column microbial communities (Gutiérrez *et al.,* 2015). This research indicates that the glacial meltwater influences the microbial community structure in inner fjords, increasing the dominance of freshwater and cold adapted communities (Gutiérrez *et al.,* 2015). Consequently, meltwater drainage has been proposed as a key influence on biological activity and diversity in Patagonian fjords.

Chilean fjords present an interesting opportunity to study the influence of glacial meltwater on marine habitats, as they form the intersection between the land and open ocean. Currently, little is known about the diversity and function of microbial communities in these fjord environments. This information will aid our understanding of microbial diversity, fjord biogeochemical cycling and may help inform predictions to how these functions may change with increased ice melt in future years. This is especially important in Chilean Patagonia, given the presence of commercially important Salmon fisheries, which rely on fjord ecosystem functioning. Additionally, the interaction of saline and fresh waters, alongside dark and cold conditions within the sediments may harbor novel and ecologically significant microbes. Techniques such as metagenomics and genome binning, which enable both the functional and taxonomic diversity of uncultured microbial communities to be uncovered, have yet to be applied to this environment. These approaches may be highly beneficial for answering questions on the diversity, importance and uniqueness of communities in this region.

This study applied metagenomics to understand the taxonomic diversity of uncultured benthic sediment metagenomes from Patagonian fjords. The uncovered taxonomic diversity was used to drive genome binning and functional analysis, providing insights into the novelty, metabolic potential and ecological diversity of these microbial communities. The findings of Chapter 2 are used to inform metagenome assembly for functional annotation, however the limitations of assembly highlighted by Chapter 4 are also considered. The findings aim to provide an insight into this unique environment, whist highlighting areas where targeted single cell genome sequencing and culture-based studies may be beneficial.

## 5.2 Methods

### 5.2.1 Field Sampling

Sampling was undertaken to investigate the microbial community composition of the benthic fjord sediments along the Steffen fjord and the outflow to the Baker channel, Chilean Patagonia. A total of 17 sites were investigated by the PISCES project for temperature, oxygen, salinity, pH and turbidity with depth using a CTD with a microfluidic colorimetric analyzer (Figure 5.1). From these sites, 5 were selected for analysis of benthic sediment microbial community composition (sites 1,2,4,5 and 7, Table 5.1). For these sites, integrated sediment grab samples were taken in a Van Veer grab sampler and stored in sterile Eppendorf tubes. Samples were frozen at -20°C for analysis at the University of Bristol. Temperature, oxygen, salinity, pH and turbidity of the fjord network surface waters were measured using underway sampling. This data was recorded using an EXO 2 sonde during the boat cruise track, which provided continuous measurements of the fjords. This data is used here to supplement the microbial community composition data obtained from the sediment samples from Sites 1,2,4,5 and 7.

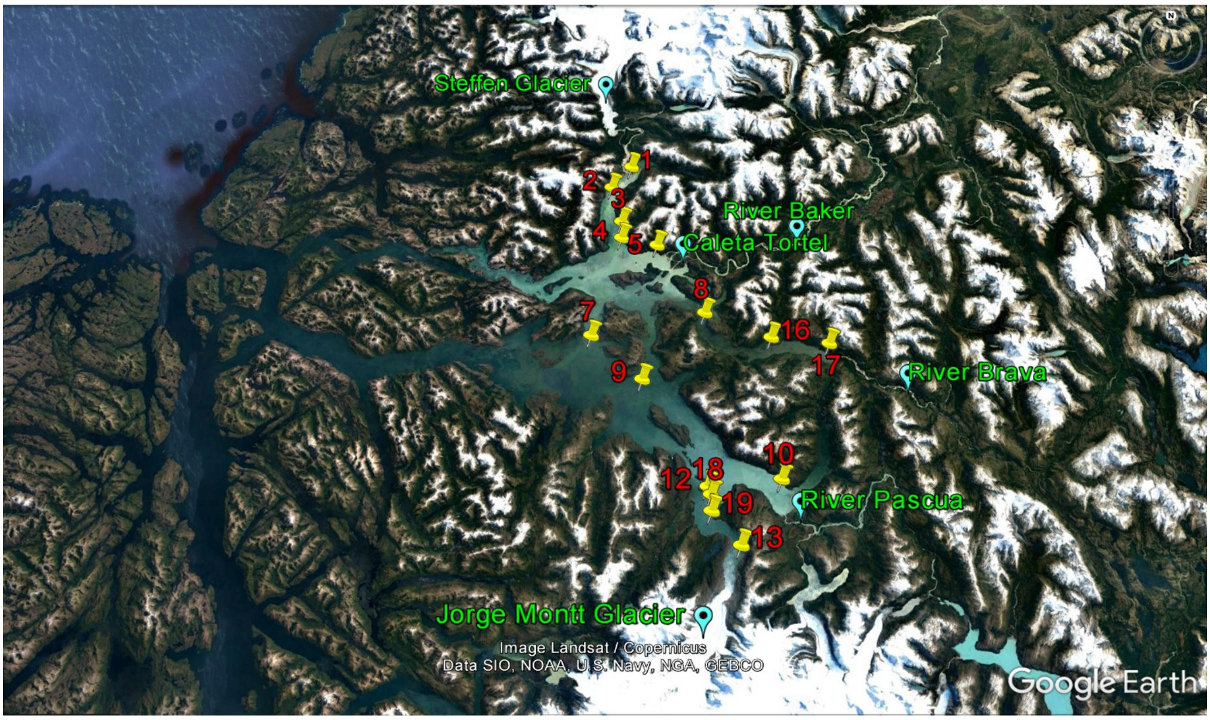


***Figure 5.1:*** *Sampling sites for aqueous nutrients, salinity, pH and dissolved oxygen. Sampling sites for sediment microbial community composition were 1,2,4,5 and 7. Source: Google Earth*

***Table 5.1:*** *Benthic sediment sampling site locations*

| Sample | Coordinates South | Coordinates West | Date sampled |
|---|---|---|---|
| 1 | S47 38.373 | W73 40.046 | 18/02/2017 |
| 2 | S47 40.748 | W73 42.897 | 20/02/2017 |
| 4 | S47 46.396 | W73 41.770 | 20/02/2017 |
| 5 | S47 47.263 | W73 36.512 | 16/02/2017 |
| 7 | S47 56.756 | W73 45.839 | 17/02/2017 |

5.2.2 Metagenome DNA extraction and sequencing

DNA extraction and sequencing was carried out in line with the protocols detailed in Chapters 3 and 4 (Nash *et al.,* 2018). DNA for metagenomics was extracted using the same protocol applied in Chapter 3 and 4 (section 3.2.3; Nash *et al.,* 2018). Metagenomes were sequenced using an Illumina Next-Seq 500, with a TruSeq library prep kit at the University of Bristol Genomics facility (Nash *et al.,* 2018). A total of 5 metagenomes were sequenced (one metagenome per site) using 2x 150bp reads.

### 5.2.3 Metagenome read annotation

The 5 sequenced metagenomes were quality trimmed using Trimmomatic V0.38, with quality checks carried out using FASTQC v0.11.7 (Bolger *et al.,* 2014; Babraham Bioinformatics). Read based taxonomic annotation was carried out in Kaiju v1.5.0 (Menzel *et al.,* 2016). Kaiju is a kmer based approach which utilises protein sequences, which are more highly conserved when compared to nucleotide sequences (Menzel *et al.,* 2016). This method has been shown to be well suited to novel and divergent metagenome samples (Menzel *et al.,* 2016). Interactive hierarchical microbial community structures of the annotated samples were obtained in Krona, using the KBASE platform (Ondov *et al.,* 2015; Arkin *et al.,* 2016).

### 5.2.4 Metagenome assembly, genome binning and annotation

The sequenced reads for each metagenome were assembled using metaSPAdes V3.11.1 with kmer length optimisation and reads mapped back to the metagenomes using Bowtie 2 v2.3.2 (Langmead and Salzberg, 2012; Nurk *et al.,* 2017). The assembled metagenomes were imported in to JGI IMG/ MER for annotation and gene calling (Chen *et al.,* 2017). Genome binning was used to extract novel genomes from the assembled metagenomes, using MaxBin2 v2.2.4, which clusters genomes based on sequence coverage and tetra-nucleotide frequencies of assembled contigs (Bankevich *et al.,* 2012; Wu *et al.,* 2015). The genomes were extracted using BinUtil v1.0.1 on KBASE (Arkin *et al.,* 2016). The taxonomic identity of the extracted bins were assigned using a BLAST-n search against all complete NCBI RefSeq genomes (e-value $10^{-5}$). The top BLAST hit for each bin was used to assign the taxonomy, based on the BLAST max score value, with an E-value below 1x $10^{-5}$. Thresholds of 95 and 85 % identity or better to classify bins to species and genus level were used, in line with analysis carried out by Camparano *et al.,* (2016). Bins classified at below 80% identity were proposed as potentially unique, as they cannot be accurately placed given the current available DNA sequences.

### 5.2.5 16s rRNA and dsrAB phylogenies

To demonstrate the potential capabilities of metagenome analysis, 16s rRNA and dissimilatory sulfite reductase (dsrAB) phylogenies were carried out, based on the results from genome binning. Whilst other explorative analyses and phylogenies could be carried out, these were selected to provide an example for how metagenome sequencing can be used.

The archaeon genus *Nitrosopumilus* was selected for phylogenetic analysis based on the 16s SSU rRNA marker gene. This is because the genus was abundant, accounting for 10-26% of genome bins across the samples and may have biogeochemical significance in the sediment nitrogen cycle. 16s rRNA sequences from reference *Nitrosopumilus* genomes were obtained from NCBI GenBank, based on the phylogeny of Qin *et al.,* (2017). *Nitrosopumilus* 16s sequences were extracted from the assembled metagenome samples using JGI IMG/MER, using BLASTn, with an e-value of $10^{-5}$ (Chen *et al.,* 2017). As the assembled metagenomes incorporated unamplified DNA from complex sediment microbial communities, the 16s sequences obtained may not cover the complete set of *Nitrosopumilus* genomes in the samples (Rodriguez and Konstantinidis, 2014). However, those obtained through BLAST searching are likely to be those most prevalent in the metagenomes, as these have been recovered by sequencing (Rodriguez and Konstantinidis, 2014). Metagenome and reference sequence alignments were created in SATé 2.2.7, using MAFT, MUSCLE and FASTTREE with the GTR+CAT model, in line with Chapter 3 (Liu *et al.,* 2011; Nash *et al.,* 2018). Manual inspection of the 16s rRNA sequence alignment and production of Phylip files was carried out using Mesquite 3.2 (Maddison and Maddison, 2017). A 16s maximum likelihood phylogeny was generated using RAXML-HPC2 8.2.10 on XSEDE through the CIPRES Science Gateway, with 1000 bootstrap iterations, implementing the GTR+G nucleotide substitution model (Stamatakis, 2014; Nash *et al.,* 2018). Final trees were produced using Figtree 1.4.3 and visual modifications were made in Inkscape 0.91. Similarity evaluation between *Nitrosopumilus* 16s sample sequences and NCBI GenBank relatives were made using NCBI BLASTn.

Additionally, genome binning and taxonomic analysis of the metagenomes highlighted the presence of sulfur cycling microbes. As metagenomes contain functional genes (alongside taxonomic markers such as 16s rRNA), the functional gene dsrAB (dissimilatory sulfite reductase) for sulfite reduction was explored phylogenetically to demonstrate the potential of using metagenomes for functional exploration. This gene encodes the reduction of sulfite ($SO_3^{2-}$) to sulfide ($S^{2-}$) in anaerobic respiration by both bacteria and archaea (Müller *et al.,* 2015). Reference dsrAB sequences were obtained from NCBI GenBank based on the phylogeny of Moreau *et al.,* (2010). Sequences were obtained from assembled metagenomes with IMG/MER using BLASTn with an e-value of $10^{-5}$, with the alignment and phylogeny generated as outlined above. The metagenome dsrAB sequences were compared to nearest relatives using BLASTn against NCBI GenBank.

### 5.2.6 Sediment organic carbon and total nitrogen content

Sediment total nitrogen (TN) and total organic carbon (TOC) were determined in triplicate for each sample site, using mass spectrometry. The protocol demonstrated in Chapters 3 and 4 was applied (section 3.2.2; Nash *et al.,* 2018).

## 5.3 Results and Discussion

### 5.3.1 Fjord metadata and influence of glacial meltwater

The results of underway sampling the surface waters of the Steffen fjord and Baker channel for temperature, oxygen, salinity, pH and turbidity are shown in Table 5.2 and Figure 5.2. The results of depth profiles for each site are shown in Table 5.2 and Figure 5.3.

***Table 5.2:*** *Metadata for benthic sediment sampling sites, courtesy of the PISCES project.*

| Site | Water depth (m) | Water temperature (°C) | Surface water dissolved oxygen (% sat) | Bottom water salinity | Surface water pH | Turbidity (FNU) |
|---|---|---|---|---|---|---|
| 1 | 61 | 10 | 104 | 32 | 6.76 | 40 |
| 2 | 199 | 8 | 104 | 34 | 7.65 | 38 |
| 4 | 268 | 8 | 102 | 34 | 7.98 | 50 |
| 5 | 78 | 8 | 104 | 34 | 7.65 | 50 |
| 7 | 360 | 9 | 104 | 34 | 7.71 | 30 |

For sediment sampling sites, water depth ranged between 61 - 360m, between sites 1 and 7. This reflects the movement from shallower inner waters of the Steffen Fjord, to more distal, marine dominated sites (Iriarte *et al.,* 2014). Surface water temperature ranged between 8 - 10°C and displayed limited variation with depth using a CTD sensor (Table 5.2; Figure 5.2). Surface water temperatures reduced at sites 1 and 13 due to the outflow of the Steffen and Jorge Montt glaciers (Figure 5.1; Figure 5.3). Surface water pH ranged between 6.8 – 8 at sediment sampling sites, with the lowest pH found at sample site 1, at the outflow of the Steffen glacier into the fjord (Table 5.2). Surface water salinity ranged between 0.4 – 1.3 and increased to 32 – 34 in bottom waters (Table 5.2; Figure 5.3; Figure 5.4). This reflects the distinct salinity stratification in the Steffen fjord, whereby fresh terrestrial and glacial runoff largely remained at the fjord surface, over a layer of dense saline marine water (Pickard, 1971; Iriarte *et al.,* 2014). This indicates a lack of mixing between the two water masses (Iriarte *et al.,* 2014).

Surface water turbidity ranged between 30 – 50 FNU, however reduced to 0.4 – 1.2 FNU at the sediment bed (Table 5.2; Figure 5.2). This again reflects the distinct fjord stratification, whereby surface waters contain more suspended sediment from terrestrial runoff than marine bottom waters (Iriarte *et al.,* 2014). The turbidity of surface waters showed distinct variation in the fjord network, increasing from 0 FNU in the Baker channel to 40 FNU in the Steffen fjord (Figure 5.3). This may relate to sediment flux from the terrestrially fed River Baker and glacial flour from the Steffen glacier, which is largely retained in the fresh surface waters of the Steffen fjord (Figure 5.1). Surface waters were saturated with oxygen, however this reduced with depth to 51 – 71% saturation at the sediment bed (Table 5.2; Figure 5.2). This is due to air-sea gas exchange with oxygen in surface waters maintaining atmospheric equilibrium, influenced by wind velocity (Broecker and Peng, 1974). This is not maintained with depth due to inadequate mixing and microbial consumption of oxygen (Leon-Munoz *et al.,* 2013).

The freshwater glacial runoff in the Steffen fjord and wider Baker channel was therefore shown to have distinct influences on the fjord physicochemical characteristics. Glacial and terrestrial runoff provided a less dense freshwater lens over more saline bottom waters throughout the fjord (Iriarte *et al.,* 2014; Figure 5.2). This freshwater lens had greater turbidity, oxygen content and lower salinity than benthic waters (Table 5.2; Figure 5.2). This information provides the physicochemical background for understanding the microbial community structure in the benthic sediment metagenomes sampled.

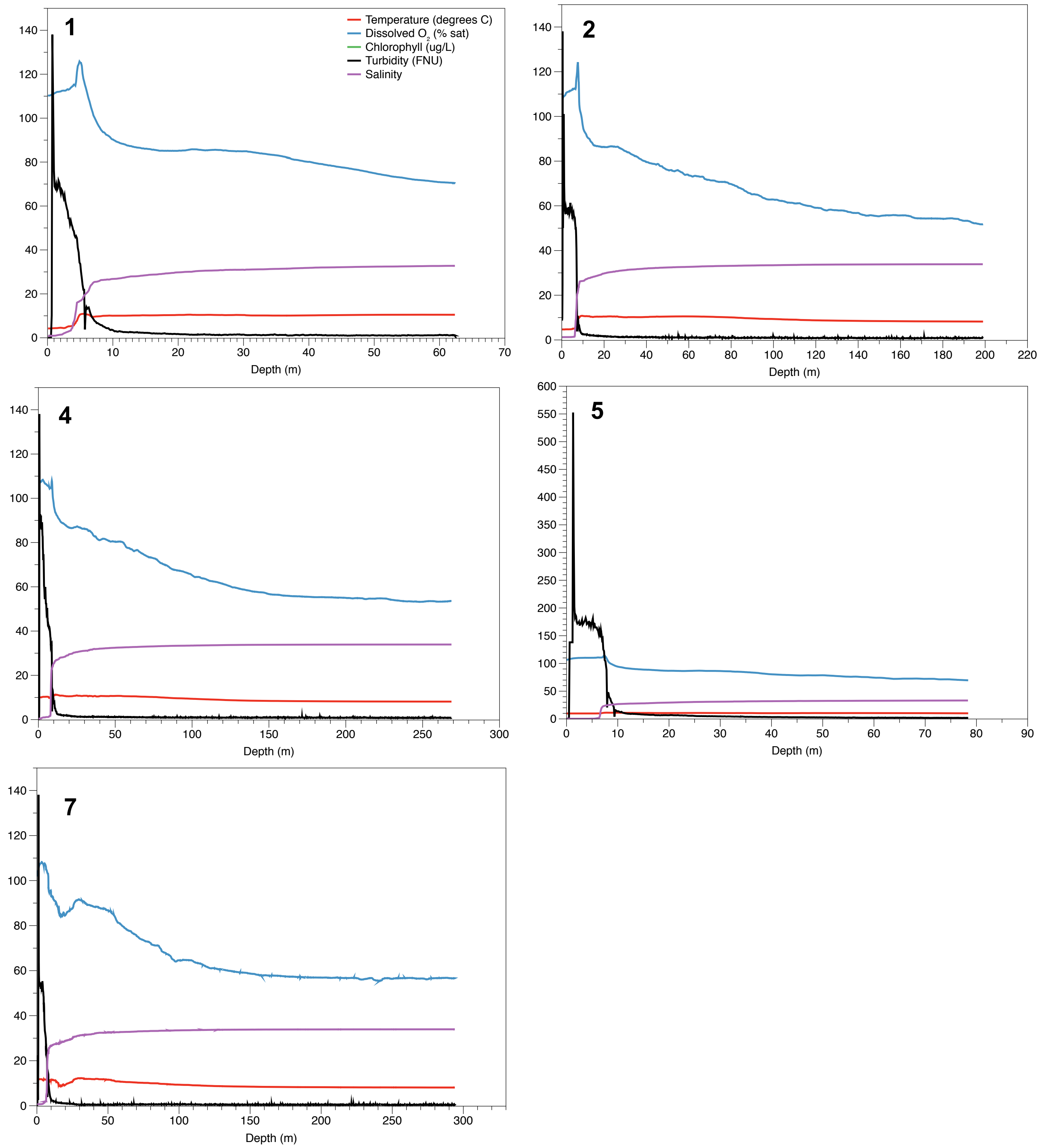


***Figure 5.2:*** *CTD depth profiles for sediment sampling cites 1,2,4,5,7. Temperature, dissolved oxygen, chlorophyll, turbidity and salinity are plotted with depth for each sampling site.*

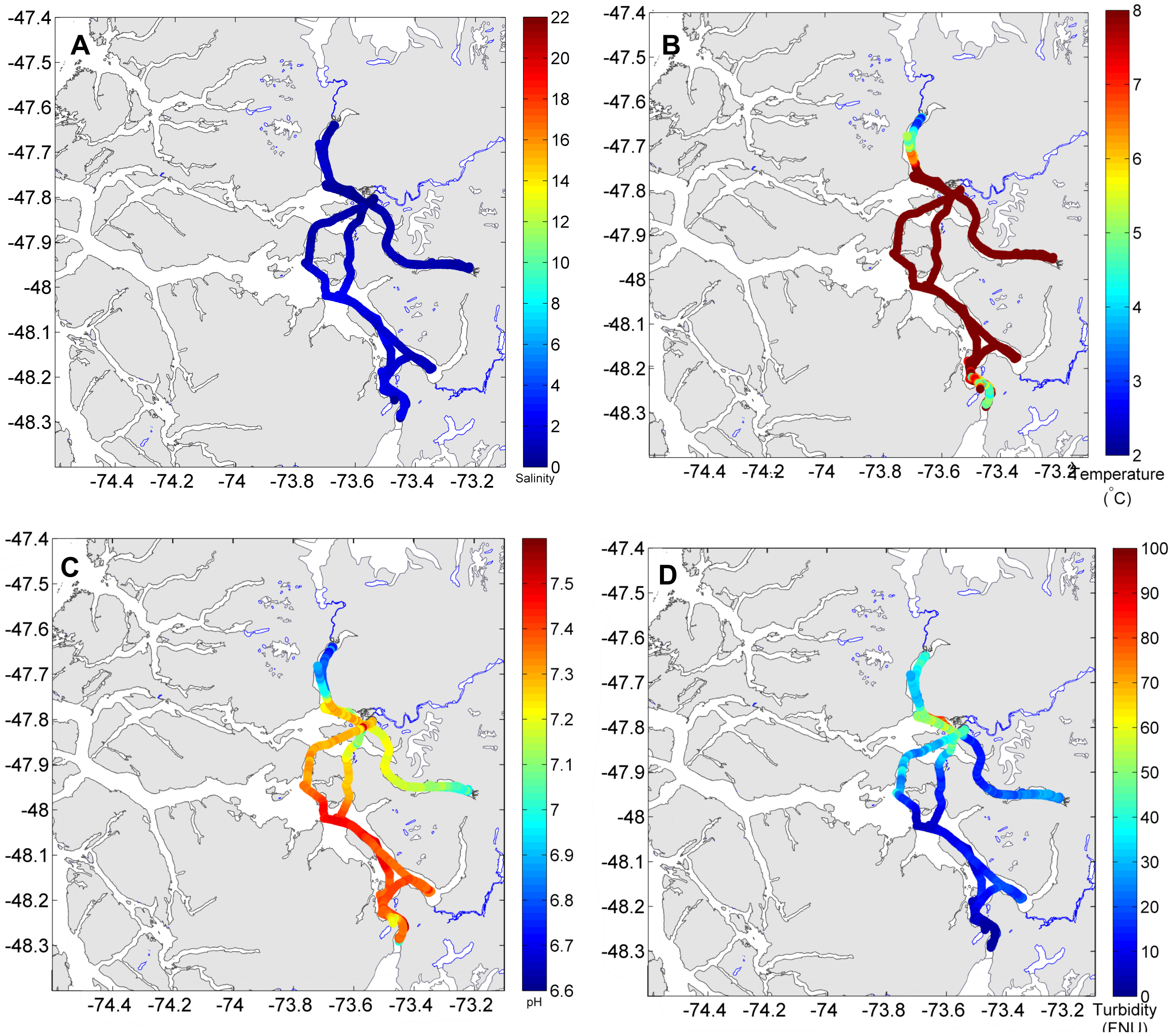


***Figure 5.3:*** *Underway sampling data of surface waters for (A) Salinity, (B) Temperature, (C) pH and (D) Turbidity. Data curtesy of Alex Beaton, PISCES project.*

### 5.3.2 Sediment TN and TOC

The results for sediment total organic carbon (TOC) and total nitrogen (TN) are displayed in Figure 5.4. TOC ranged between 3 mg g$^{-1}$ to 13.5 mg g$^{-1}$, from Site 2 to Site 17 respectively. The sites sampled for metagenomics along the Steffen Fjord (Sites 1-7) display a smaller range of values, between 3 – 5.8 mg g$^{-1}$ TOC (Figure 5.4). The range of TN values from sediments was more constrained, ranging between 0.2 – 1 mg g$^{-1}$ (Figure 5.4). The limit of detection was 1 mg g$^{-1}$ for both TOC and TN. These values are indicative of low nutrient conditions in the benthic fjord sediments, which may mediate the microbial communities which they can sustain. In particular, the minimal values of TN indicate nitrogen may be limiting, and thus may influence the diversity and structure of microbial life in the sediments.

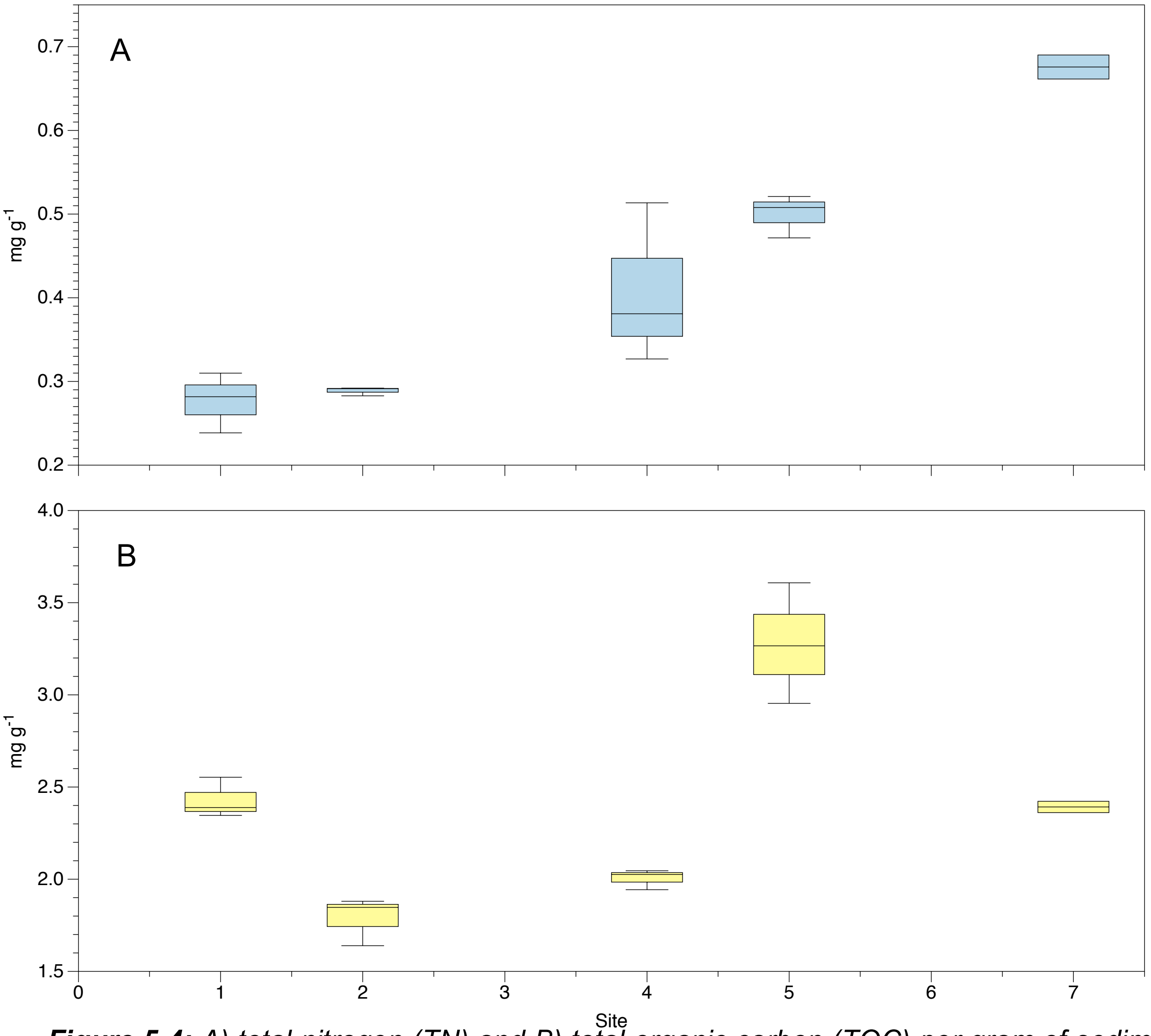


***Figure 5.4:*** *A) total nitrogen (TN) and B) total organic carbon (TOC) per gram of sediment for metagenome sampling sites (1,2,4,5,7). The limit of detection was 1 mg g$^{-1}$ for both TOC and TN. The relative standard deviation of measurement was 1.44% and 4.86% for TOC and TN respectively, calculated by measuring a series of 9 standards.*

### 5.3.3 Metagenome assembly and annotation

Following sequencing, the five benthic sediment metagenomes were assembled to create longer contigs for functional annotation. Metagenome assembly has been shown to improve the annotation quality of the datasets, due to longer DNA sequence lengths for alignment-based interpretation (Nagarajan and Pop, 2013, Chapter 2). The output quality for metagenome assembly and read mapping are displayed in Table 5.3. The number of contigs ranged between 25,112 – 64,579, with contig size between 10,000 – 1,000,000 bp (Table 5.3).

The read coverage of the metagenomes ranged between 87.1- 95.4%, highlighting the majority of the raw read data was used in the assemblies (Table 5.3). However, following taxonomic and functional annotation in JGI IMG/MER 98.76 – 99.33% of reads were not assigned to a taxonomy at 90% identity (Table 5.4). This may highlight a substantial novelty of genomes in the samples which cannot be identified using current databases, as these do not fully profile current global microbial diversity (Ferrer *et al.,* 2005).

**Table 5.3:** Metagenome assembly statistics for microbial community metagenomes, highlighting assembly size and contig length, produced in Quast (Gurevich *et al.,* 2013). Output quality metrics for raw read mapping against the metagenomes are also provided.

| | Site 1 | Site 2 | Site 4 | Site 5 | Site 7 |
|---|---|---|---|---|---|
| **Assembly parameters** | | | | | |
| # contigs | 47703 | 44713 | 42395 | 64579 | 25112 |
| # contigs (>= 0 bp) | 47703 | 44713 | 42395 | 64579 | 25112 |
| # contigs (>= 1000 bp) | 47703 | 44713 | 42395 | 64579 | 25112 |
| # contigs (>= 10000 bp) | 2374 | 1581 | 2050 | 3227 | 946 |
| # contigs (>= 100000 bp) | 6 | 14 | 14 | 10 | 0 |
| # contigs (>= 1000000 bp) | 0 | 0 | 0 | 0 | 0 |
| Largest contig | 169641 | 297725 | 360451 | 227878 | 82647 |
| Total length | 197488635 | 174470052 | 175632561 | 268241056 | 95362614 |
| Total length (>= 0 bp) | 197488635 | 174470052 | 175632561 | 268241056 | 95362614 |
| Total length (>= 1000 bp) | 197488635 | 174470052 | 175632561 | 268241056 | 95362614 |
| Total length (>= 10000 bp) | 42735865 | 33490431 | 41852693 | 62069788 | 15473582 |
| Total length (>= 100000 bp) | 762550 | 1960800 | 2074273 | 1341724 | 0 |
| Total length (>= 1000000 bp) | 0 | 0 | 0 | 0 | 0 |
| N50 | 4300 | 3842 | 4189 | 4236 | 3774 |
| N75 | 2731 | 2602 | 2691 | 2725 | 2598 |
| L50 | 11553 | 11316 | 9829 | 15263 | 6821 |
| L75 | 26303 | 25409 | 23230 | 35496 | 14572 |
| GC (%) | 51.32 | 47.16 | 45.67 | 46.93 | 50.36 |
| **Mapping parameters** | | | | | |
| Total Reads | 174621110 | 161768224 | 164078036 | 155907364 | 164920816 |
| Unmapped Reads | 160588571 | 148029200 | 149463098 | 136710343 | 157328781 |
| Mapped Reads | 14032539 | 13739024 | 14614938 | 19197021 | 7592035 |
| Singletons | 958195 | 981434 | 853002 | 1273877 | 798535 |
| % coverage | 91.96 | 91.5 | 91.1 | 87.1 | 95.4 |

**Table 5.4:** Percentage of taxonomically unassigned assembled sequences at 60 and 90% identity, using JGI IMG/MER

| Sample | % unassigned sequences (90% identity) | % unassigned sequences (60% identity) |
|---|---|---|
| 1 | 98.97 | 68.35 |
| 2 | 99.15 | 80.8 |
| 4 | 99.33 | 83.46 |
| 5 | 99.24 | 75.94 |
| 7 | 98.76 | 65.19 |

Taxonomic annotation of quality trimmed (unassembled) sequencing reads was carried out following the limited scope of assembled metagenome annotation (Table 5.4). The read classifier Kaiju was selected, as this classifier has been shown to have a higher sensitivity to underrepresented genera than other kmer based classifiers (Menzel *et al.,* 2016). Furthermore, the read classification is based upon protein sequences, which are more conserved and resilient to sequencing errors than DNA based analysis (Menzel *et al.,* 2016).

The overall microbial community composition of the five metagenomes, at the class level, can be identified in Figure 5.5. The full classification from class to species level is available in Appendix 4, Tables 1-5. In total, only 25% of the sequenced reads could be annotated with a taxonomy at the class level (Figure 5.5). Of these reads, 20% could not be annotated at the class level, or were at a very low abundance (Figure 5.5). This indicates a presence of either novel, and/or highly fragmented DNA in the datasets. As current annotation techniques rely on the use of known sequences, any novel or unique organisms will not be assigned a taxonomy (Menzel *et al.,* 2016). This is particularly the case for environmental samples, as known sequenced genomes are often biased towards those which have a medical application (Menzel *et al.,* 2016). The use of binning techniques may help to delineate the presence of novel genomes in the samples. However, as the samples were derived from low nutrient sediments, it is possible that a large fraction of the DNA is degraded (i.e. not living) due to the hash environmental conditions and exposure to water (Taberlet *et al.,* 2012; Bohmann *et al.,* 2014). As DNA is generally fragmented during the degradation process, the shorter length hinders the ability to annotate proteins, either by sequence alignment, sequence composition or tetranucleotide frequencies (Taberlet *et al.,* 2012; Menzel *et al.,* 2016). To reveal the composition of this more degraded DNA, amplification using short primers would be needed, for example using metabarcoding (Taberlet *et al.,* 2012; Leray and Knowlton, 2015) however this is beyond the scope of the current analysis.

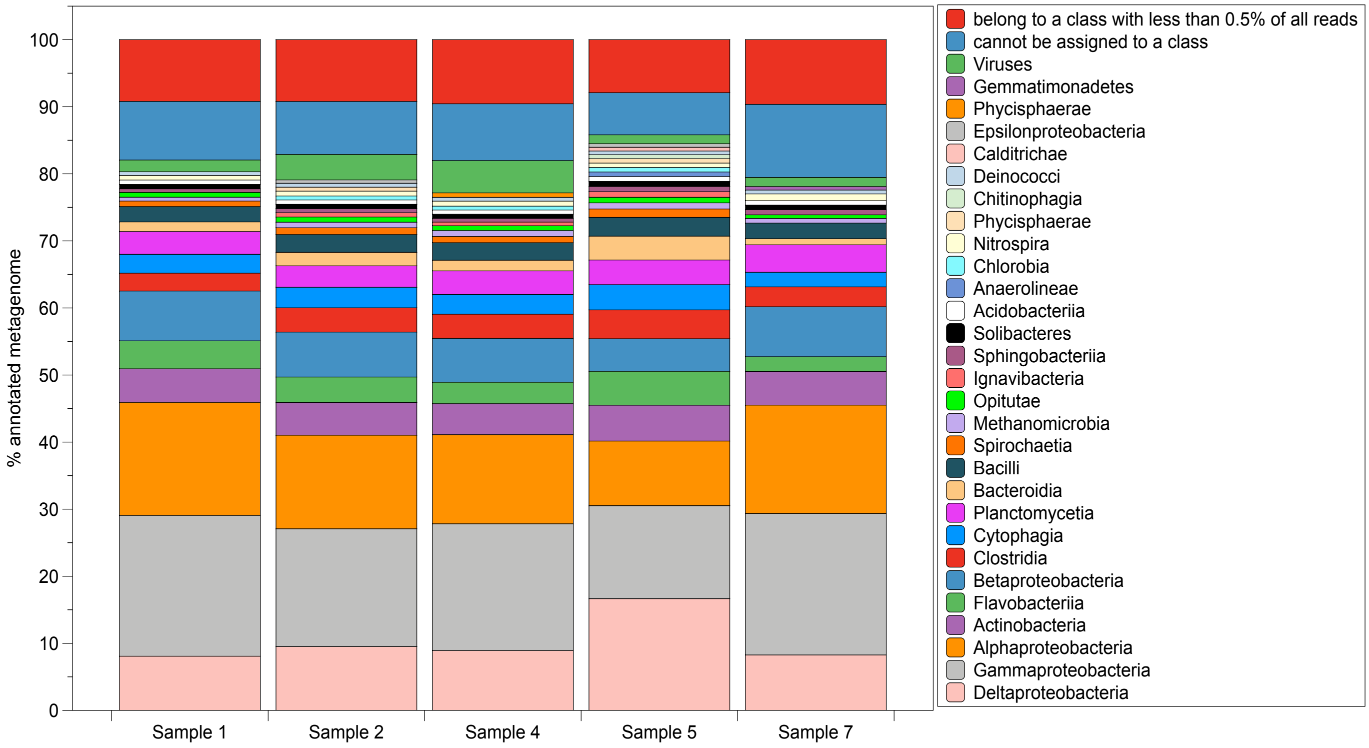


***Figure 5.5:*** *Kaiju taxonomic classification of reads for the five sequenced samples, at the Class level. Only the annotated sequences are shown, which account for 25% of the total metagenome sequenced for each site. The percentage of the classified metagenome attributed to each Class is shown. Sequences that have been classified but cannot be assigned to a class and very low abundance reads, are also provided. The legend is ordered in line with the sequence of bars in the main plot. The full taxonomic classification from class – species level is available in Appendix 4, Tables 1-5.*

From the classified reads, the five samples show a similar overall bacterial community composition, with Deltaproteobacteria, Gammaproteobacteria, Alphaproteobacteria and Actinobacteria dominant in all samples (Figure 5.5). The presence of organisms related to *Methanomicrobia, Chlorobia, Clostridia, Nitrospira* and *Anaerolineae* is indicative of anoxic conditions in the surface sediments, thus eliciting the detection of anaerobic organisms and the potential for methanogenesis (Eisen *et al.,* 2002; Luker *et al.,* 2010; Mackelprang *et al.,* 2011; Matsuura *et al.,* 2015). This aligns with the reduction in oxygen saturation with depth, which may continue below the sediment surface due to microbial activity in sediments and lack of oxygen replenishment (Figure 5.2; Oschmann, 2001). The microbial communities in the benthic fjord sediments were subjected to low nutrient, saline and dark conditions (Figure 5.4; Figure 5.5). Consequently, there may be evidence for adaptations within the bacterial community, as the environment may be highly selective for the organisms which can survive (Figure 5.5). For example, *Acidobacteria* (1.4 - 1.6% of metagenomes) have been shown to produce large amounts of exopolysaccharide (EPS), which is both protective and adhesive (Ward *et al.,* 2009). The adhesive properties of EPS can increase the nutrient uptake, which is beneficial in a low TN environment such as the Steffen fjord (Weiner *et al.,* 1995; Ward *et al.,* 2009). Additionally, *Acidobacteria* are facultative anaerobes and use ferric iron reduction in the absence of oxygen for respiration (Blothe *et al.,* 2008). Facultative anaerobic respiration may be beneficial in the Steffen fjord given the potential for low oxygen conditions. Furthermore, taxa related to the aerobic chemoorganotroph, *Deinococci,* were identified in sediment samples (Copeland *et al.,* 2012). This organism is a known extremophile and can survive ionizing radiation, cold and oligotrophy (Copeland *et al.,* 2012). Additionally, evidence for the selection of resilient microbes is supported by the identification of the extremophile, *Epsilonproteobacteria* (Figure 5.5). These bacteria are commonly found in extreme environments such as hydrothermal vents and cold seeps and obtains energy through chemolithotrophy, thus can survive without organic compounds and light (Takai *et al.,* 2005). This may be significant given the low TOC in the sediment samples (Figure 5.4). This consequently provides additional evidence for the selection of well adapted resilient organisms to the environmental conditions in the benthic sediment of the Steffen fjord, if active.

The hierarchical structure of the Kaiju microbial community composition was visualized in interactive Krona plots for each site, to derive more detail from the classifications (Appendix 4 Tables 1-5; Table 5.5). Overall, Proteobacteria accounted for 46 - 55% of bacteria in the metagenomes and provided evidence for sulfur cycling potential within the fjord sediments (Table 5.5). The presence of *Desulfovibrionales* and *Desulfobacterales* indicates the potential for anaerobic sulfate reduction, producing sulfide ($H_2S$) from sulfate ($SO_4^{2-}$) (Kuever, 2014, Table 5.5). Sulfate reducing bacteria are commonly found in seawater due to the abundance of sulfate, alongside in anaerobic sediments, as they contribute to the degradation of organic matter (Goldharber and Kaplan, 1974). Additional evidence for sulfur cycling potential is provided through the identification of the purple sulfur bacteria *Chromatiales,* accounting for 3-6% of the metagenomes (Table 5.5). *Chromatiales* are typically anaerobic, and utilise the waste $H_2S$ from sulfate reduction, and oxidise it to elemental sulfur (S) (Imhoff, 2005). Consequently, the sediment metagenomes provide evidence for microbial sulfur cycling potential within the sediments. In order to validate the activity of sulfur cycling, further work using transcriptomics could be implemented to identify the transcription of key marker genes, such as dsrAB and apsA for sulfate reduction (Wagner *et al.,* 2005).

Archaea accounted for 3-5% of cellular organisms within the metagenomes (Table 5.5). The depth of sequencing used in shotgun metagenomics is often suitable to profile bacterial communities, however may not have the coverage to isolate the archaea and eukaryotes present (Hugenholtz and Tyson, 2008). Deep sequencing is often required to reveal the full breath of these communities (Narasingarao *et al.,* 2012). However, the archaea identified do provide support for the presence of anaerobic conditions at the sediment bed. The presence of the methanogens *Methanomicrobia*, *Methanobacteriales, Methanococcales* and

*Thermoplasmata* indicate the potential production of methane ($CH_4$), using carbon dioxide ($CO_2$) as the terminal electron acceptor in anaerobic respiration (Table 5.5; Valentine, 2002; Bonin and Boone, 2006). Methanogenesis is significant for carbon cycling, acting as the final stage in the degradation of organic material following the preferential use of other electron acceptors such as oxygen, sulfate and nitrate (Zeikus, 1977; Valentine, 2002). Methane has been shown to be a substantial contributor to global climate change, having a warming potential greater than that of $CO_2$ (Valentine, 2002). The presence of methanogenesis may therefore have implications on both the local and global carbon cycle over long timescales, for both the degradation and recycling of organic matter, and production of $CH_4$.

***Table 5.5:*** *Extract of Kaiju read based classification for the 5 metagenome samples, based on those discussed in this analysis. The classification name and percentage of metagenome assigned is provided. Full classification for each metagenome is available in Appendix 4 Tables 1-5.*

| | Classification rank | Sample 1 (%) | Sample 2 (%) | Sample 4 (%) | Sample 5 (%) | Sample 7 (%) |
|---|---|---|---|---|---|---|
| Bacteria | Domain | 93 | 92 | 90 | 96 | 91 |
| Archaea | Domian | 5 | 4 | 5 | 3 | 7 |
| Viruses | Domain | 2 | 4 | 5 | 1 | 2 |
| | | | | | | |
| Proteobacteria | Phylum | 55 | 49 | 49 | 46 | 54 |
| Terrabacteria | Phylum | 15 | 17 | 17 | 19 | 16 |
| FCB group/ Sphingobacteria | Phylum | 12 | 13 | 11 | 17 | 8 |
| PVC group | Phylum | 5 | 5 | 6 | 6 | 6 |
| Thaumarchaeota | Phylum | 3 | 2 | 3 | 3 | 5 |
| Other | - | 10 | 14 | 14 | 9 | 11 |
| | | | | | | |
| Methanomicrobia | Class | 0.6 | 0.8 | 0.9 | 0.9 | 0.7 |
| Epsilonproteobacteria | Class | 0.4 | 0.5 | 0.5 | 0.5 | 0.4 |
| Clostridia | Class | 3 | 4 | 4 | 4 | 3 |
| Bacilli | Class | 2 | 3 | 3 | 3 | 2 |
| Thermoplasmata | Class | 0.08 | 0.1 | 0.2 | 0.1 | 0.1 |
| Other | - | 93.92 | 91.6 | 91.4 | 91.5 | 93.8 |
| | | | | | | |
| Desulfovibrionales | Order | 0.7 | 0.9 | 0.9 | 1 | 0.8 |
| Desulfobacterales | Order | 1 | 2 | 2 | 7 | 2 |
| Chromatiales | Order | 5 | 4 | 4 | 3 | 6 |
| Micrococcales | Order | 0.08 | 0.8 | 0.7 | 0.8 | 0.7 |
| Oscillatoriales | Order | 0.7 | 0.5 | 0.5 | 0.5 | 0.6 |
| Methanobacteriales | Order | 0.1 | 0.2 | 0.3 | 0.2 | 0.2 |
| Methanococcales | Order | 0.09 | 0.2 | 2 | 0.1 | 0.1 |
| Other | - | 92.33 | 91.4 | 89.6 | 87.4 | 10.4 |
| | | | | | | |
| Nitrobacter | Genus | 0.09 | 0.08 | 0.07 | 0.09 | 0.08 |
| Rhizobium | Genus | 0.7 | 0.3 | 0.3 | 0.3 | 0.4 |
| Pseudomonas | Genus | 1 | 1 | 1 | 1 | 1 |
| Other | - | 98.21 | 98.62 | 98.63 | 98.61 | 98.52 |

Finally, the metagenomes in this study provide evidence for nitrogen cycling potential within the sediments. The presence of organisms relating to *Thaumarchaeota* (3 – 5 % of the metagenomes), chemolithotrophic ammonia oxidisers, suggests that microbial nitrification may be taking place within the sediments (Table 5.5; Park *et al.,* 2012). Ammonia oxidising archaea convert ammonia ($NH_3$) to nitrite ($NO_2^-$), which can be utilised by nitrite oxidisers in such as *Nitrobacter*, to convert $NO_2^-$ to nitrate ($NO_3^-$) for assimilation (Jetten, 2008). In addition to this, the presence of *Epsilonproteobacteria* provides evidence for denitrification potential within the sediments, converting $NO_3^-$ to nitrogen gas ($N_2$) and is a key pathway for fixed nitrogen removal from sediments (Murdock and Juniper, 2017). Nitrogen-fixing bacteria such as *Clostridia*, *Rhizobium* and $N_2$ fixing cyanobacteria such as *Micrococcales*, and *Oscillatoriales* were identified in the sediment metagenomes (Table 5.5). These organisms may be instrumental in providing labile fixed nitrogen into the sediments, to be utilized by heterotrophic bacteria for protein assimilation (Bergman *et al.,* 1997). If the nitrogen fixers were active in the sediments, they may be crucial for facilitating heterotrophic microbial communities, given the oligotrophic conditions identified (Figure 5.4). Finally, ammonification is performed by bacteria to convert organic nitrogen to more bioavailable ammonium ($NH_4^+$) (Gruber, 2008). Evidence for the potential of this pathway is shown through the presence of organisms relating to *Pseudomonas* and *Bacilli* in the sediments (Table 5.5). Consequently, the metagenomes sampled contain organisms capable of the complete nitrogen cycle, despite the low nitrogen concentrations sampled (Figure 5.4). If active, the nitrification and denitrification pathways may be tightly coupled to microbial nitrogen fixation to obtain sufficient nitrogen stocks.

### 5.3.4 Genome binning

Genome binning was carried out using assembled metagenomes, to isolate discrete genomes which may be unique compared to those currently available in reference datasets. The results for each metagenome are shown in Tables 5.6 – 5.10, identifying the genome bins (discrete genomes) alongside the BLASTn match of each genome to NCBI GenBank. Metagenome 1 contained 36 genome bins, with 67% of bins containing less than 80% identity to cultured relatives on NCBI GenBank (Table 5.6). This indicates that the metagenomes may contain several novel species and/or strains, which have yet to be cultured or fully sequenced. This pattern was continued within the remainder of the metagenomes, with 58%, 55%, 71% and 63% of genome bins below 80% identity to GenBank for samples 2,4,5 and 7 respectively (Tables 5.6 – 5.10). This highlights a need for further culture-based studies or single cell sequencing, to isolate complete genome sequences for these potentially novel organisms. The presence of novel sequences may relate to the dark, cold and low nutrient conditions within the fjord sediments, lack of previous exploration of these sediments and the incomplete nature of current DNA reference databases.

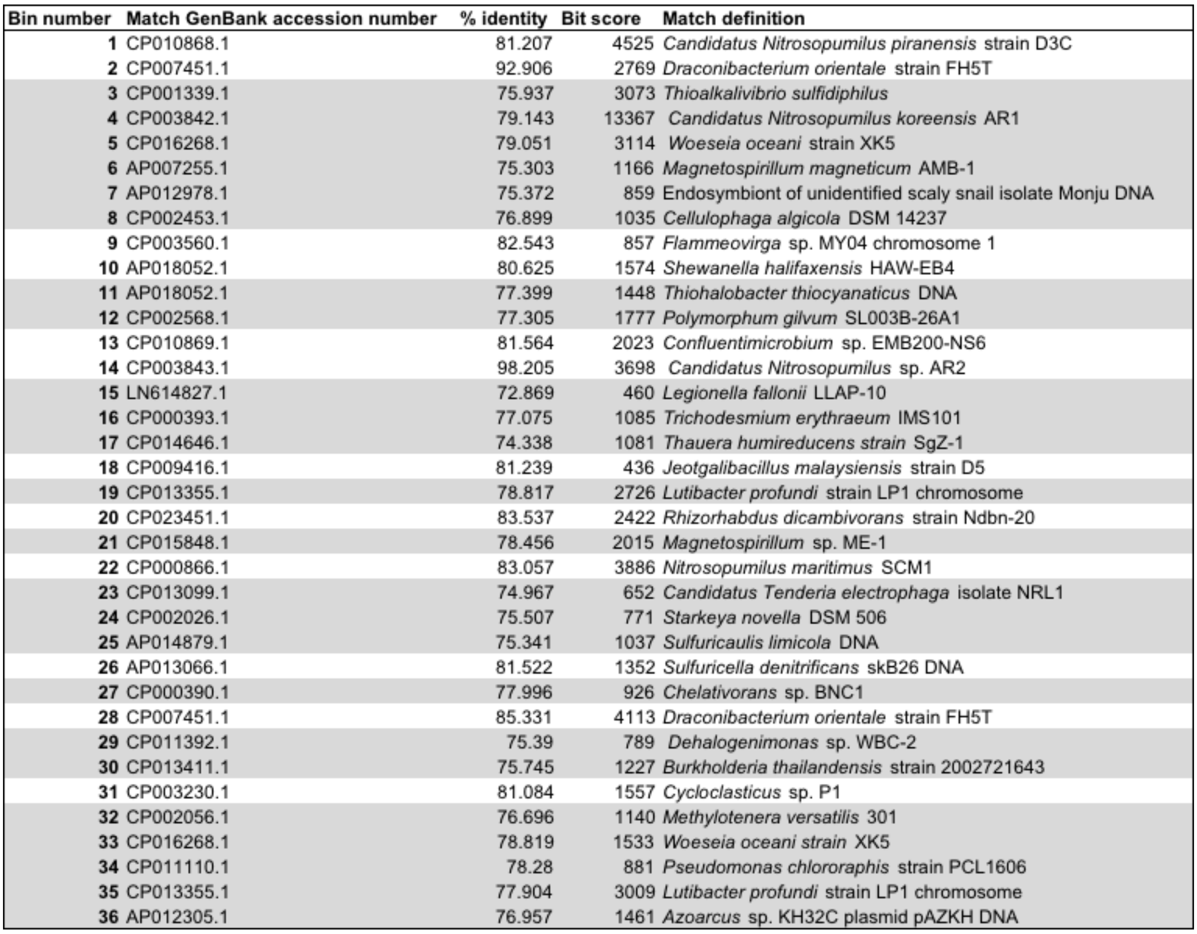
In line with the taxonomy, the genome bins obtained were broadly similar between the metagenomes, however Sample 5 contained a wider range of genomes than Sample 7 (Tables 5.6 – 5.10). The genome bins predominately spanned the bacterial phyla Proteobacteria, Cyanobacteria, Bacteroidetes, Chloroflexi, Firmicutes and Verrucomicrobia, alongside the archeal phyla Thaumarchaeota and Euryarchaeota (Tables 5.6 – 5.10). The Thaumarchaeota *Nitrosopumilus* was common between metagenomes, accounting for between 10 – 26% of genome bins for each site. The binned *Nitrosopumilus* sequences contained between 75 – 98% identity to NCBI GenBank reference genomes, highlighting the presence of both unique and previously isolated genomes in the samples (Tables 5.6 – 5.10). *Nitrosopumilus* are ammonia-oxidising archaea, which use $CO_2$ as a carbon source to covert ammonia ($NH_3$) to nitrite ($NO_2^-$) during nitrification (Banning *et al*., 2015). These organisms may therefore be significant in sediment nitrogen cycling, helping to provide nitrite (and subsequently nitrate) for biological uptake by heterotrophic organisms (Francis *et al.,* 2005). This may be substantial for facilitating the activity of heterotrophic microbes, given the potential for nitrogen limitation in the sediments (Figure 5.4).

Interestingly, sequences relating to the Proteobacterium *Magnetospirillum* were found in samples 1, 4 and 7 (Tables 5.6 – 5.10). This organism is microaerophilic, magneto tactic and capable of producing high quality magnetite from low iron aquatic environments (Matsunaga *et al.,* 2005). However, sequences ranged between 74 – 78% identity to *Magnetosprillum*, highlighting the potential for a novel species in the metagenomes (Tables 5.6 – 5.10). As *Magnetosprillium* has a wide range of commercial applications, such as bioremediation, manufacture and pharmaceuticals, the isolation of a new novel species may provide additional commercial applications to those already identified (Safarik and Safarikova, 2004). This again highlights the need for further culture-based analysis and single cell sequencing of environmental genomes from unique environments, such as those sampled here.

The binned genomes also displayed evidence of sulfur cycling, in line with the results of the read based taxonomic annotation (Figure 5.5). The presence of the chemolithotrophic *Thioalkalivibrio*, in metagenomes 1, 2 and 7 indicates the potential of sulfur oxidation in the sediments, whereby elemental sulfur (S) is converted to sulfate ($SO_4^{2-}$) (Sorokin *et al.,* 2001). In turn, the presence of the anaerobic *Desulfococcus, Desulfuromonas, Desulfobacter* and *Desulfobacterium* highlight the presence of sulfur and sulfate reduction (Taylor and Parkes, 1983; Kleindienst *et al.,* 2014). These organisms have been shown to oxidise organic compounds such as acetate and pyruvate using sulfate and sulfur, to reduced forms of sulfur (Brysch *et al.,* 1987).

Consequently, the genome binning has identified the potential for novel species with ecological importance in the sediment metagenomes. Deep re-sequencing of the metagenomes may help to recover longer more complete metagenome assembled genomes. This may be beneficial to gain a better understating of the strain level diversity and metabolisms of unique species. This, alongside genome sequencing of cultured stains would help provide new draft genomes to publicly available databases.

***Table 5.6:*** *Metagenome Sample 1 assembled bins and top BLAST match for each bin, with % NCBI GenBank identity and bit score. In total, 36 assembled bins representing distinct genomes within Sample 1 metagenome were identified. The genomes with an % identity less than 80% to GenBank sequences are highlighted.*

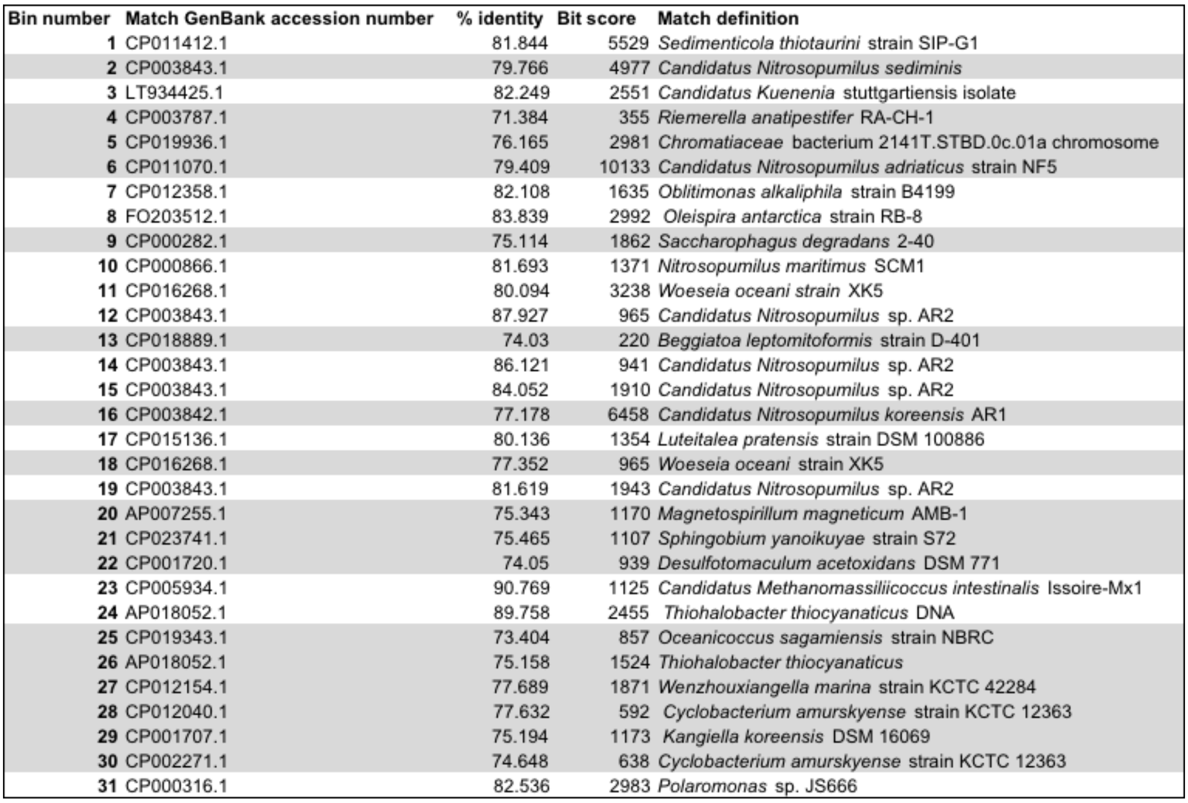


| Bin number | Match GenBank accession number | % identity | Bit score | Match definition |
|---|---|---|---|---|
| 1 | CP010868.1 | 81.207 | 4525 | *Candidatus Nitrosopumilus piranensis* strain D3C |
| 2 | CP007451.1 | 92.906 | 2769 | *Draconibacterium orientale* strain FH5T |
| 3 | CP001339.1 | 75.937 | 3073 | *Thioalkalivibrio sulfidiphilus* |
| 4 | CP003842.1 | 79.143 | 13367 | *Candidatus Nitrosopumilus koreensis* AR1 |
| 5 | CP016268.1 | 79.051 | 3114 | *Woeseia oceani* strain XK5 |
| 6 | AP007255.1 | 75.303 | 1166 | *Magnetospirillum magneticum* AMB-1 |
| 7 | AP012978.1 | 75.372 | 859 | Endosymbiont of unidentified scaly snail isolate Monju DNA |
| 8 | CP002453.1 | 76.899 | 1035 | *Cellulophaga algicola* DSM 14237 |
| 9 | CP003560.1 | 82.543 | 857 | *Flammeovirga* sp. MY04 chromosome 1 |
| 10 | AP018052.1 | 80.625 | 1574 | *Shewanella halifaxensis* HAW-EB4 |
| 11 | AP018052.1 | 77.399 | 1448 | *Thiohalobacter thiocyanaticus* DNA |
| 12 | CP002568.1 | 77.305 | 1777 | *Polymorphum gilvum* SL003B-26A1 |
| 13 | CP010869.1 | 81.564 | 2023 | *Confluentimicrobium* sp. EMB200-NS6 |
| 14 | CP003843.1 | 98.205 | 3698 | *Candidatus Nitrosopumilus* sp. AR2 |
| 15 | LN614827.1 | 72.869 | 460 | *Legionella fallonii* LLAP-10 |
| 16 | CP000393.1 | 77.075 | 1085 | *Trichodesmium erythraeum* IMS101 |
| 17 | CP014646.1 | 74.338 | 1081 | *Thauera humireducens strain* SgZ-1 |
| 18 | CP009416.1 | 81.239 | 436 | *Jeotgalibacillus malaysiensis* strain D5 |
| 19 | CP013355.1 | 78.817 | 2726 | *Lutibacter profundi* strain LP1 chromosome |
| 20 | CP023451.1 | 83.537 | 2422 | *Rhizorhabdus dicambivorans* strain Ndbn-20 |
| 21 | CP015848.1 | 78.456 | 2015 | *Magnetospirillum* sp. ME-1 |
| 22 | CP000866.1 | 83.057 | 3886 | *Nitrosopumilus maritimus* SCM1 |
| 23 | CP013099.1 | 74.967 | 652 | *Candidatus Tenderia electrophaga* isolate NRL1 |
| 24 | CP002026.1 | 75.507 | 771 | *Starkeya novella* DSM 506 |
| 25 | AP014879.1 | 75.341 | 1037 | *Sulfuricaulis limicola* DNA |
| 26 | AP013066.1 | 81.522 | 1352 | *Sulfuricella denitrificans* skB26 DNA |
| 27 | CP000390.1 | 77.996 | 926 | *Chelativorans* sp. BNC1 |
| 28 | CP007451.1 | 85.331 | 4113 | *Draconibacterium orientale* strain FH5T |
| 29 | CP011392.1 | 75.39 | 789 | *Dehalogenimonas* sp. WBC-2 |
| 30 | CP013411.1 | 75.745 | 1227 | *Burkholderia thailandensis* strain 2002721643 |
| 31 | CP003230.1 | 81.084 | 1557 | *Cycloclasticus* sp. P1 |
| 32 | CP002056.1 | 76.696 | 1140 | *Methylotenera versatilis* 301 |
| 33 | CP016268.1 | 78.819 | 1533 | *Woeseia oceani strain* XK5 |
| 34 | CP011110.1 | 78.28 | 881 | *Pseudomonas chlororaphis* strain PCL1606 |
| 35 | CP013355.1 | 77.904 | 3009 | *Lutibacter profundi* strain LP1 chromosome |
| 36 | AP012305.1 | 76.957 | 1461 | *Azoarcus* sp. KH32C plasmid pAZKH DNA |

***Table 5.7:*** *Metagenome Sample 2 assembled bins and top BLAST match for each bin, with % identity and bit score. In total, 36 assembled bins representing distinct genomes within Sample 2 metagenome were identified. The genomes with an % identity less than 80% to GenBank sequences are highlighted.*

| Bin number | Match GenBank accession number | % identity | Bit score | Match definition |
|---:|---|---:|---:|---|
| **1** | CP021431.1 | 100 | 150 | *Loktanella vestfoldensis* strain SMR4r |
| **2** | CP003843.1 | 97.522 | 3856 | *Candidatus Nitrosopumilus* sp. AR2 |
| **3** | CP011412.1 | 81.859 | 5535 | *Sedimenticola thiotaurini* strain SIP-G1 |
| **4** | CR522870.1 | 78.462 | 457 | *Desulfotalea psychrophila* LSv54 |
| **5** | CP001147.1 | 82.116 | 1735 | *Thermodesulfovibrio yellowstonii* DSM 11347 |
| **6** | CP000478.1 | 77.095 | 1048 | *Syntrophobacter fumaroxidans* MPOB |
| **7** | CP001339.1 | 75.921 | 3068 | *Thioalkalivibrio sulfidiphilus* HL-EbGr7 |
| **8** | CP018632.1 | 72.188 | 893 | *Granulosicoccus antarcticus* IMCC3135 |
| **9** | CP000252.1 | 80.675 | 737 | *Thioalkalivibrio sulfidiphilus* HL-EbGr7 |
| **10** | CP011070.1 | 78.046 | 8626 | *Thioalkalivibrio sulfidiphilus* HL-EbGr7 |
| **11** | CP016268.1 | 80.768 | 4023 | *Woeseia oceani strain* XK5 |
| **12** | CP017478.1 | 71.649 | 292 | *Lutibacter* sp. LPB0138 |
| **13** | CP003843.1 | 81.137 | 3921 | *Candidatus Nitrosopumilus* sp. AR2 |
| **14** | CP011070.1 | 85.923 | 2769 | *Candidatus Nitrosopumilus adriaticus* strain NF5 |
| **15** | CP011036.1 | 86.065 | 2265 | *Pseudoalteromonas nigrifaciens* strain KMM 661 |
| **16** | CP013355.1 | 77.878 | 2198 | *Lutibacter profundi* strain LP1 chromosome |
| **17** | CP003843.1 | 85.51 | 2846 | *Candidatus Nitrosopumilus* sp. AR2 |
| **18** | CP011412.1 | 84.806 | 1216 | *Sedimenticola thiotaurini* strain SIP-G1 |
| **19** | CP001131.1 | 75.762 | 1206 | *Anaeromyxobacter* sp. K |
| **20** | CP001032.1 | 84.372 | 1886 | *Opitutus terrae* PB90-1 |
| **21** | CP011070.1 | 75.438 | 983 | *Candidatus Nitrosopumilus adriaticus* strain NF5 |
| **22** | CP010869.1 | 81.564 | 2023 | *Confluentimicrobium* sp. EMB200-NS6 |
| **23** | CP000142.2 | 79.661 | 1079 | *Pelobacter carbinolicus* DSM 2380 |
| **24** | CP011070.1 | 84.783 | 2320 | *Candidatus Nitrosopumilus adriaticus* strain NF5 |
| **25** | CP020892.1 | 75.18 | 761 | *Pseudomonas* sp. M30-35 chromosome |
| **26** | CP011412.1 | 74.583 | 1262 | *Sedimenticola thiotaurini* strain SIP-G1 chromosome |
| **27** | CP000083.1 | 74.989 | 1014 | *Colwellia psychrerythraea* 34H |
| **28** | CP018889.1 | 77.537 | 712 | *Beggiatoa leptomitoformis* strain D-401 chromosome |
| **29** | AP018042.1 | 76.407 | 1153 | *Marinifilaceae bacterium* SPP2 DNA |
| **30** | CP012398.1 | 76.8 | 987 | *Chelatococcus* sp. CO-6 |
| **31** | CP016268.1 | 77.68 | 2095 | *Woeseia oceani* strain XK5 |
| **32** | CP020555.1 | 79.577 | 1653 | *Streptomyces* sp. Sge12 |
| **33** | CP013355.1 | 78.791 | 2067 | *Lutibacter profundi* strain LP1 |
| **34** | CP000859.1 | 78.682 | 1262 | *Desulfococcus oleovorans* Hxd3 |
| **35** | LN890655.2 | 90.747 | 1195 | *Ardenticatena* sp. Cfx-K strain Cfx-K |
| **36** | CP023439.1 | 79.824 | 1461 | *Thauera* sp. K11 chromosome |

***Table 5.8:*** *Metagenome Sample 4 assembled bins and top BLAST match for each bin, with % identity and bit score. In total, 31 assembled bins representing distinct genomes within Sample 4 metagenome were identified. The genomes with an % identity less than 80% to GenBank sequences are highlighted.*

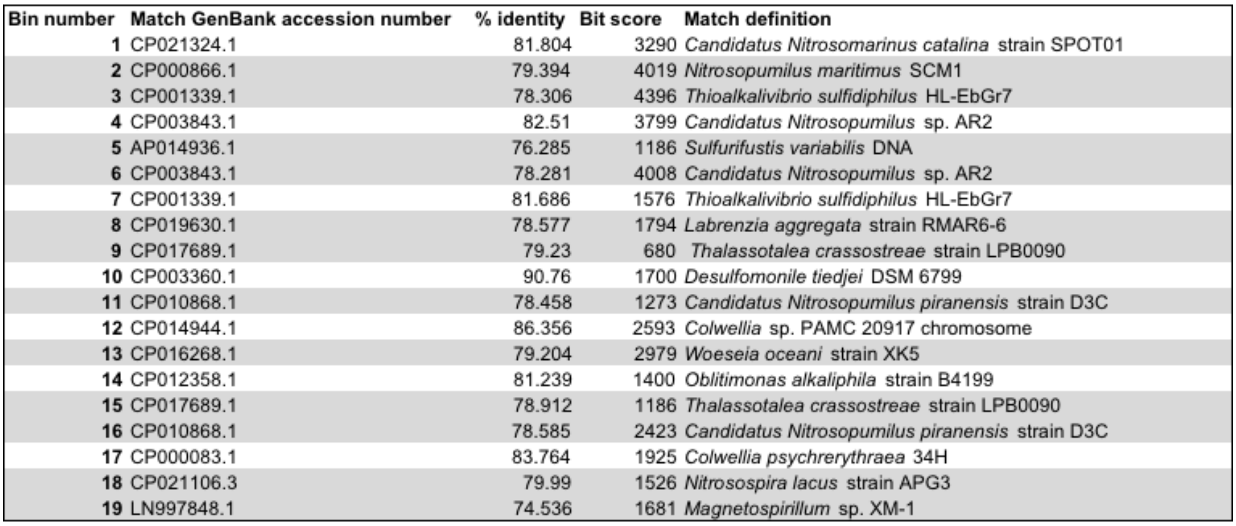


| Bin number | Match GenBank accession number | % identity | Bit score | Match definition |
|---|---|---|---|---|
| 1 | CP011412.1 | 81.844 | 5529 | *Sedimenticola thiotaurini* strain SIP-G1 |
| 2 | CP003843.1 | 79.766 | 4977 | *Candidatus Nitrosopumilus sediminis* |
| 3 | LT934425.1 | 82.249 | 2551 | *Candidatus Kuenenia* stuttgartiensis isolate |
| 4 | CP003787.1 | 71.384 | 355 | *Riemerella anatipestifer* RA-CH-1 |
| 5 | CP019936.1 | 76.165 | 2981 | *Chromatiaceae* bacterium 2141T.STBD.0c.01a chromosome |
| 6 | CP011070.1 | 79.409 | 10133 | *Candidatus Nitrosopumilus adriaticus* strain NF5 |
| 7 | CP012358.1 | 82.108 | 1635 | *Oblitimonas alkaliphila* strain B4199 |
| 8 | FO203512.1 | 83.839 | 2992 | *Oleispira antarctica* strain RB-8 |
| 9 | CP000282.1 | 75.114 | 1862 | *Saccharophagus degradans* 2-40 |
| 10 | CP000866.1 | 81.693 | 1371 | *Nitrosopumilus maritimus* SCM1 |
| 11 | CP016268.1 | 80.094 | 3238 | *Woeseia oceani strain* XK5 |
| 12 | CP003843.1 | 87.927 | 965 | *Candidatus Nitrosopumilus* sp. AR2 |
| 13 | CP018889.1 | 74.03 | 220 | *Beggiatoa leptomitoformis* strain D-401 |
| 14 | CP003843.1 | 86.121 | 941 | *Candidatus Nitrosopumilus* sp. AR2 |
| 15 | CP003843.1 | 84.052 | 1910 | *Candidatus Nitrosopumilus* sp. AR2 |
| 16 | CP003842.1 | 77.178 | 6458 | *Candidatus Nitrosopumilus koreensis* AR1 |
| 17 | CP015136.1 | 80.136 | 1354 | *Luteitalea pratensis* strain DSM 100886 |
| 18 | CP016268.1 | 77.352 | 965 | *Woeseia oceani* strain XK5 |
| 19 | CP003843.1 | 81.619 | 1943 | *Candidatus Nitrosopumilus* sp. AR2 |
| 20 | AP007255.1 | 75.343 | 1170 | *Magnetospirillum magneticum* AMB-1 |
| 21 | CP023741.1 | 75.465 | 1107 | *Sphingobium yanoikuyae* strain S72 |
| 22 | CP001720.1 | 74.05 | 939 | *Desulfotomaculum acetoxidans* DSM 771 |
| 23 | CP005934.1 | 90.769 | 1125 | *Candidatus Methanomassiliicoccus intestinalis* Issoire-Mx1 |
| 24 | AP018052.1 | 89.758 | 2455 | *Thiohalobacter thiocyanaticus* DNA |
| 25 | CP019343.1 | 73.404 | 857 | *Oceanicoccus sagamiensis* strain NBRC |
| 26 | AP018052.1 | 75.158 | 1524 | *Thiohalobacter thiocyanaticus* |
| 27 | CP012154.1 | 77.689 | 1871 | *Wenzhouxiangella marina* strain KCTC 42284 |
| 28 | CP012040.1 | 77.632 | 592 | *Cyclobacterium amurskyense* strain KCTC 12363 |
| 29 | CP001707.1 | 75.194 | 1173 | *Kangiella koreensis* DSM 16069 |
| 30 | CP002271.1 | 74.648 | 638 | *Cyclobacterium amurskyense* strain KCTC 12363 |
| 31 | CP000316.1 | 82.536 | 2983 | *Polaromonas* sp. JS666 |

***Table 5.9:*** *Metagenome Sample 5 assembled bins and top BLAST match for each bin, with % identity and bit score. In total, 51 assembled bins representing distinct genomes within Sample 5 metagenome were identified. The genomes with an % identity less than 80% to GenBank sequences are highlighted.*

| Bin number | Match GenBank accession number | % identity | Bit score | Match definition |
|---|---|---|---|---|
| **1** | CP000724.1 | 83.259 | 2180 | *Alkaliphilus metalliredigens* QYMF |
| **2** | CP016268.1 | 79.953 | 3103 | *Woeseia oceani strain* XK5 |
| **3** | CP012851.1 | 76.274 | 1037 | *Persicobacter* sp. JZB09 |
| **4** | LT981265.1 | 77.833 | 1290 | *Candidatus Nitrosocaldus cavascurensis* strain SCU2 |
| **5** | CP013118.1 | 73.115 | 878 | *Salinivirga cyanobacteriivorans* strain L21-Spi-D4 |
| **6** | CP002031.1 | 76.244 | 1387 | *Geobacter sulfurreducens* KN400 |
| **7** | CP001032.1 | 76.652 | 1676 | *Opitutus terrae* PB90-1 |
| **8** | CP009788.1 | 73.92 | 682 | *Geobacter pickeringii* strain G13 |
| **9** | CP010802.1 | 75.07 | 2220 | *Desulfuromonas soudanensis* strain WTL chromosome |
| **10** | CP002271.1 | 78.881 | 737 | *Stigmatella aurantiaca* DW4/3-1 |
| **11** | CP003843.1 | 91.176 | 3602 | *Candidatus Nitrosopumilus* sp. AR2 |
| **12** | CP003360.1 | 78.68 | 4414 | *Desulfomonile tiedjei* DSM 6799 |
| **13** | CP003350.1 | 74.774 | 1461 | *Frateuria aurantia* DSM 6220 |
| **14** | CP013355.1 | 75.697 | 590 | *Lutibacter profundi* strain LP1 |
| **15** | CP009505.1 | 83.006 | 1927 | *Methanosarcina* sp. MTP4 |
| **16** | FO203503.1 | 73.2 | 1469 | *Desulfobacula toluolica* Tol2 |
| **17** | LT934425.1 | 75.995 | 815 | *Candidatus Kuenenia stuttgartiensis* |
| **18** | CP012358.1 | 82.159 | 1640 | *Oblitimonas alkaliphila* strain B4199 chromosome |
| **19** | CP013355.1 | 74.638 | 1397 | *Lutibacter profundi* strain LP1 chromosome |
| **20** | CP003985.1 | 73.498 | 669 | *Desulfocapsa sulfexigens* DSM 10523 |
| **21** | CP003843.1 | 86.747 | 15400 | *Candidatus Nitrosopumilus* sp. AR2 |
| **22** | CP006900.2 | 73.958 | 484 | *Pandoraea pnomenusa* 3kgm |
| **23** | CP003273.1 | 76.135 | 1574 | *Desulfotomaculum gibsoniae* DSM 7213 |
| **24** | CP019913.2 | 75 | 928 | *Desulfococcus multivorans* strain DSM 2059 |
| **25** | CP016268.1 | 79.47 | 1724 | *Woeseia oceani* strain XK5 |
| **26** | CP000859.1 | 83.796 | 1672 | *Desulfococcus oleovorans* Hxd3 |
| **27** | CP000934.1 | 74.105 | 1664 | *Cellvibrio japonicus* Ueda107 |
| **28** | CP025791.1 | 75.197 | 994 | *Flavivirga eckloniae* strain ECD14 chromosome |
| **29** | CU207366.1 | 73.988 | 130 | *Gramella forsetii* KT0803 |
| **30** | CP013457.1 | 80.632 | 1099 | *Burkholderia s*p. MSMB617WGS |
| **31** | CP013118.1 | 74.824 | 734 | *Salinivirga cyanobacteriivorans* strain L21-Spi-D4 |
| **32** | CP003843.1 | 86.724 | 2320 | *Candidatus Nitrosopumilus* sp. AR2 |
| **33** | CP003843.1 | 94.042 | 9022 | *Candidatus Nitrosopumilus* sp. AR2 |
| **34** | CP001087.1 | 76.225 | 1709 | *Desulfobacterium autotrophicum* HRM2 |
| **35** | CP003985.1 | 74.536 | 881 | *Desulfocapsa sulfexigens* DSM 10523 |
| **36** | CP003843.1 | 88.372 | 1705 | *Desulfobacterium autotrophicum* HRM2 |
| **37** | CP003843.1 | 88.209 | 4833 | *Desulfobacterium autotrophicum* HRM2 |
| **38** | CP016268.1 | 79.01 | 2320 | *Woeseia oceani strain* XK5 |
| **39** | CP013355.1 | 81.96 | 1448 | *Lutibacter profundi* strain LP1 chromosome |
| **40** | CP000473.1 | 77.081 | 1903 | *Candidatus Solibacter usitatus* Ellin6076 |
| **41** | CP003380.1 | 89.854 | 1404 | *Methylophaga frappieri* strain JAM7 |
| **42** | CP011454.1 | 78.822 | 931 | *Gemmatimonas phototrophica* strain AP64 |
| **43** | CP010904.1 | 75.319 | 669 | *Kiritimatiella glycovorans* strain L21-Fru-AB |
| **44** | CP013355.1 | 78.896 | 2442 | *Lutibacter profundi* strain LP1 chromosome |
| **45** | CP011125.1 | 73.835 | 1144 | *Sandaracinus amylolyticus* strain DSM 53668 |
| **46** | CP013355.1 | 77.89 | 5387 | *Lutibacter profundi* strain LP1 chromosome |
| **47** | CP003389.1 | 77.644 | 992 | *Corallococcus coralloides* DSM 2259 |
| **48** | CP003843.1 | 87.817 | 3410 | *Candidatus Nitrosopumilus* sp. AR2 |
| **49** | CP013355.1 | 84.56 | 1770 | *Lutibacter profundi strain* LP1 chromosome |
| **50** | CP015080.1 | 78.171 | 1742 | *Desulfuromonas* sp. DDH964 |
| **51** | CP006587.1 | 78.762 | 1112 | *Hymenobacter* sp. APR13 |

***Table 5.10:*** *Metagenome Sample 7 assembled bins and top BLAST match for each bin, with % identity and bit score. In total, 19 assembled bins representing distinct genomes within Sample 7 metagenome were identified. The genomes with an % identity less than 80% to GenBank sequences are highlighted.*



| Bin number | Match GenBank accession number | % identity | Bit score | Match definition |
|---|---|---|---|---|
| 1 | CP021324.1 | 81.804 | 3290 | *Candidatus Nitrosomarinus catalina* strain SPOT01 |
| 2 | CP000866.1 | 79.394 | 4019 | *Nitrosopumilus maritimus* SCM1 |
| 3 | CP001339.1 | 78.306 | 4396 | *Thioalkalivibrio sulfidiphilus* HL-EbGr7 |
| 4 | CP003843.1 | 82.51 | 3799 | *Candidatus Nitrosopumilus* sp. AR2 |
| 5 | AP014936.1 | 76.285 | 1186 | *Sulfurifustis variabilis* DNA |
| 6 | CP003843.1 | 78.281 | 4008 | *Candidatus Nitrosopumilus* sp. AR2 |
| 7 | CP001339.1 | 81.686 | 1576 | *Thioalkalivibrio sulfidiphilus* HL-EbGr7 |
| 8 | CP019630.1 | 78.577 | 1794 | *Labrenzia aggregata* strain RMAR6-6 |
| 9 | CP017689.1 | 79.23 | 680 | *Thalassotalea crassostreae* strain LPB0090 |
| 10 | CP003360.1 | 90.76 | 1700 | *Desulfomonile tiedjei* DSM 6799 |
| 11 | CP010868.1 | 78.458 | 1273 | *Candidatus Nitrosopumilus piranensis* strain D3C |
| 12 | CP014944.1 | 86.356 | 2593 | *Colwellia* sp. PAMC 20917 chromosome |
| 13 | CP016268.1 | 79.204 | 2979 | *Woeseia oceani* strain XK5 |
| 14 | CP012358.1 | 81.239 | 1400 | *Oblitimonas alkaliphila* strain B4199 |
| 15 | CP017689.1 | 78.912 | 1186 | *Thalassotalea crassostreae* strain LPB0090 |
| 16 | CP010868.1 | 78.585 | 2423 | *Candidatus Nitrosopumilus piranensis* strain D3C |
| 17 | CP000083.1 | 83.764 | 1925 | *Colwellia psychrerythraea* 34H |
| 18 | CP021106.3 | 79.99 | 1526 | *Nitrosospira lacus* strain APG3 |
| 19 | LN997848.1 | 74.536 | 1681 | *Magnetospirillum* sp. XM-1 |

### 5.3.5 Phylogenetic analysis – 16s rRNA

To explore the findings from genome binning further and to highlight the potential of metagenome data analysis, the genus *Nitrosopumilus* was selected for 16s rRNA analysis, as this archaeon was identified in multiple genome bins and may be ecologically important in the sediment nitrogen cycle, if active. The 16s sequences were obtained from the metagenomes, however as amplification was not used, some less abundant strains may not have been recovered through sequencing. However, it is likely that the *Nitrosopumilus* sequences obtained were those which were the most dominant in the samples (Rodriguez and Konstantinos, 2014). The metagenome 16s sequences relating to *Nitrosopumilus* were aligned with and subsequently evaluated in a maximum likelihood phylogeny with GenBank cultured relatives (Figure 5.6). The results show that the sample sequences align most closely with *Nitrosopumilus HCE1* (*Nitrosopumilus oxclinae*) and *Nitrosopumilus HCA1* (*Nitrosopumilus cobalaminigenes)* (Figure 5.6). When the 16s sample sequences were BLAST searched, Nitrosopumilus sequence 1, 2 and 3 aligned at 99% sequence similarity to *Nitrosopumilus oxclinae, Nitrosopumilus cobalaminigenes* and *Candidatus Nitrosopumilus sediminis* (Park *et al.,* 2012; Qin *et al.,* 2017; Figure 5.6). This supports the genome bins and phylogeny in identifying these sequences as belonging to *Nitrosopumilus* genomes. However, whilst 99% similarity was shown using the 16s sequences, work by Park *et al.,* (2012) has shown that when looking at overall genome nucleotide identity, the nucleotide similarity can reduce below 80%, providing evidence for new species or strains, in line with the results from genome binning. Consequently, this highlights the need for culturing and full genome sequencing of these organisms, as these may constitute new genomes to the *Nitrosopumilus* genus, as found by Park *et al.,* (2012).

As highlighted above, the presence of *Nitrosopumilus* genomes provided evidence for ammonia oxidising archaea (AOA) in the sediments, which if active, would contribute to nitrogen cycling (Banning *et al.,* 2015). These AOA are well suited to low nitrogen environments such as the fjord sediments sampled, as they have a high affinity to ammonia (Qin *et al.,* 2017; Figure 5.4). Additionally, they are suited to the low organic carbon contents of the sediments as they efficiently use $CO_2$ as a carbon source (Qin *et al.,* 2017). This could explain why these genomes were consistently identified using genome binning, as these genomes may be tolerant to low nutrient conditions (Tables 5.6 – 5.10). Furthermore, the 99% similarity with *Nitrosopumilus cobalaminigenes* indicates the genomes may have additional adaptive mechanisms to the extreme conditions in the fjord sediments. *Nitrosopumilus cobalaminigenes* is psychotolerant and can survive in a wide salinity range (10 – 40 PSU) (Qin *et al.,* 2017). This suggests these genomes may be able to survive preferentially in the cold fjord conditions and withstand fluctuating salinities during mixing of stratified waterbodies (Qin *et al.,* 2017).

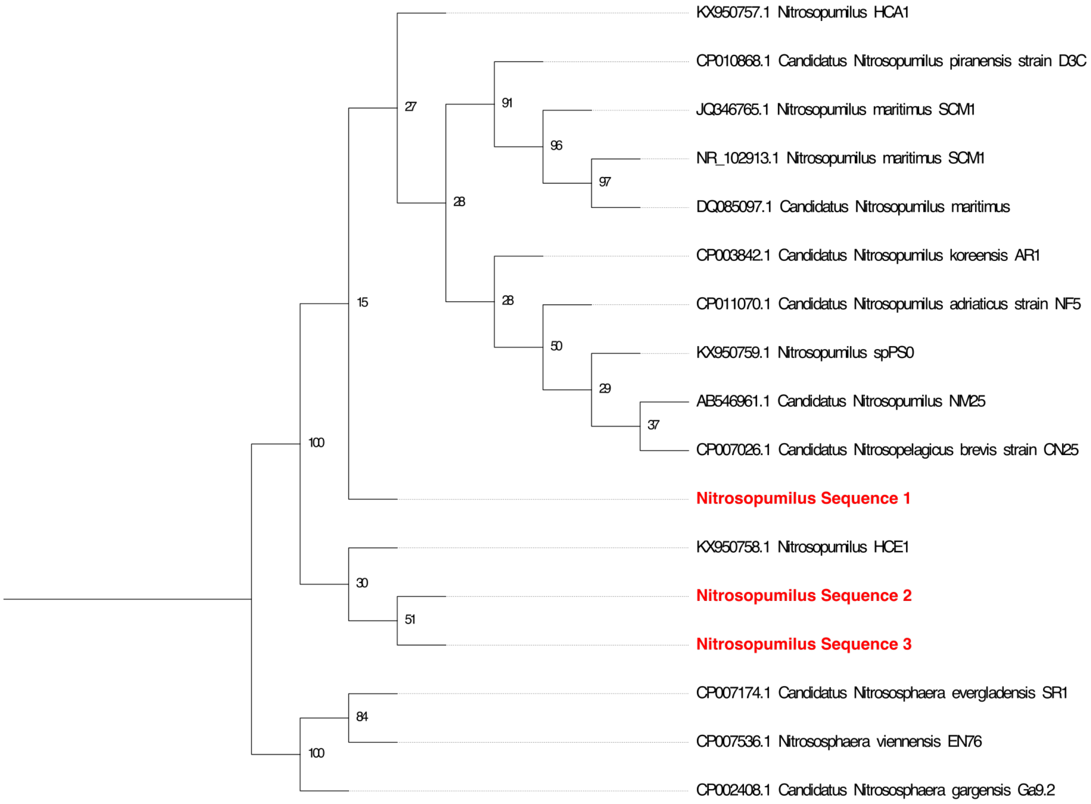

KX950757.1 Nitrosopumilus HCA1
CP010868.1 Candidatus Nitrosopumilus piranensis strain D3C
JQ346765.1 Nitrosopumilus maritimus SCM1
NR_102913.1 Nitrosopumilus maritimus SCM1
DQ085097.1 Candidatus Nitrosopumilus maritimus
CP003842.1 Candidatus Nitrosopumilus koreensis AR1
CP011070.1 Candidatus Nitrosopumilus adriaticus strain NF5
KX950759.1 Nitrosopumilus spPS0
AB546961.1 Candidatus Nitrosopumilus NM25
CP007026.1 Candidatus Nitrosopelagicus brevis strain CN25
Nitrosopumilus Sequence 1
KX950758.1 Nitrosopumilus HCE1
Nitrosopumilus Sequence 2
Nitrosopumilus Sequence 3
CP007174.1 Candidatus Nitrososphaera evergladensis SR1
CP007536.1 Nitrososphaera viennensis EN76
CP002408.1 Candidatus Nitrososphaera gargensis Ga9.2
27
91
96
97
28
15
28
50
29
37
100
30
51
84
100

***Figure 5.6:*** *16s rRNA Maximum likelihood phylogeny of Nitrosopumilus sample sequences with cultured relatives, obtained from NCBI GenBank. The accession numbers for cultured relatives are shown and metagenome sample sequences highlighted in bold. Bootstrap*

*support values are provided for each branch node, resulting from 1000 tree iterations. The tree is rooted on Nitrosophaera, in line with the phylogeny of Quin et al., (2017).*

### 5.3.6 Phylogenetic analysis – dissimilatory sulfite reductase

A key advantage of metagenomics, in comparison to 16s rRNA sequencing, is that functional genes can be investigated (Wooley *et al.,* 2010). This means that functional pathways can be searched for, and the taxonomy subsequently assigned (Handelsman, 2004). The benefit of this is that it provides more accurate evidence for biogeochemical cycling in the sediments, as the genes encoding the functional pathways are present (Handelsman, 2004). Whilst this does not show activity of the pathways, it highlights the potential for these to occur (Wooley *et al.,* 2010).

The findings of both the read based taxonomy and genome bins (Figure 5.5, Tables 5.6 – 5.10) identified the presence of sulfur cycling microbes in the sediments. To investigate this further, and demonstrate the potential of metagenome sequencing, the dsrAB gene (dissimilatory sulfite reductase) was investigated. This gene encodes the reduction of sulfite ($SO_3^{2-}$) to sulfide ($H_2S$) (Müller *et al.,* 2015). This is a key reduction step in the sulfur cycle, whereby prokaryotes reduce sulfite for energy in combination with the oxidation of organic matter, during anaerobic respiration (Müller *et al.,* 2015). A maximum likelihood phylogeny of metagenome sample dsrAB genes and those from GenBank relatives is shown in Figure 5.7. Here, cultured relatives have been grouped in to the classes of Clostridia, Archaeglobi, Nitrospira and Deltaproteobacteria (Figure 5.7), based on the phylogeny of Moreau *et al.,* (2010). The sample sequences (Sf dsr 1 – 7) are grouped together, branching between Clostridia and Deltaproteobacteria (Figure 5.7). This may highlight a unique set of sulfite reducers in the samples. To investigate this further, the sample dsrAB sequences were BLASTn searched against cultured isolates in NCBI GenBank, with results shown in Table 11. The sequences were matched with several sulfite reducing microbes, with 75% of samples identified under the class of Gammaproteobacteria (Table 8). Samples SF dsr 1 – 2 were classified as Deltaproteobacteria and Chlorobia, respectively. However, as the matches received a low score, and identity of 82%, there is limited confidence in these assignments. Overall, the identity of matches ranged between 76 – 93%, suggesting that these sequences may be unique (Table 11). Additional single genome sequencing would help isolate the full genomes of these prokaryotes, thereby validating the suggestions made here. Furthermore, future work with meta-transcriptomics would help highlight if these sulfite reducers were active, as this technique investigates the transcribed mRNA (Carvalhais *et al.,* 2013).

The identification of dsrAB sequences in the samples supports the potential of sulfur cycling highlighted by the read based taxonomy and genome bins (Figure 5.5; Table 5.5; Tables 5.6

– 5.10). This contributes to the evidence indicating the potential for biogeochemical cycling within these oligotrophic sediments. The sulfite may be lost by volatilisation or mineral formation, however may also be oxidised back to sulfite and sulfate by chemo-lithotrophic prokaryotes such as *Chromatiales,* identified in the read based taxonomic annotation (Holmer and Stockholm, 2001; Imhoff, 2005).

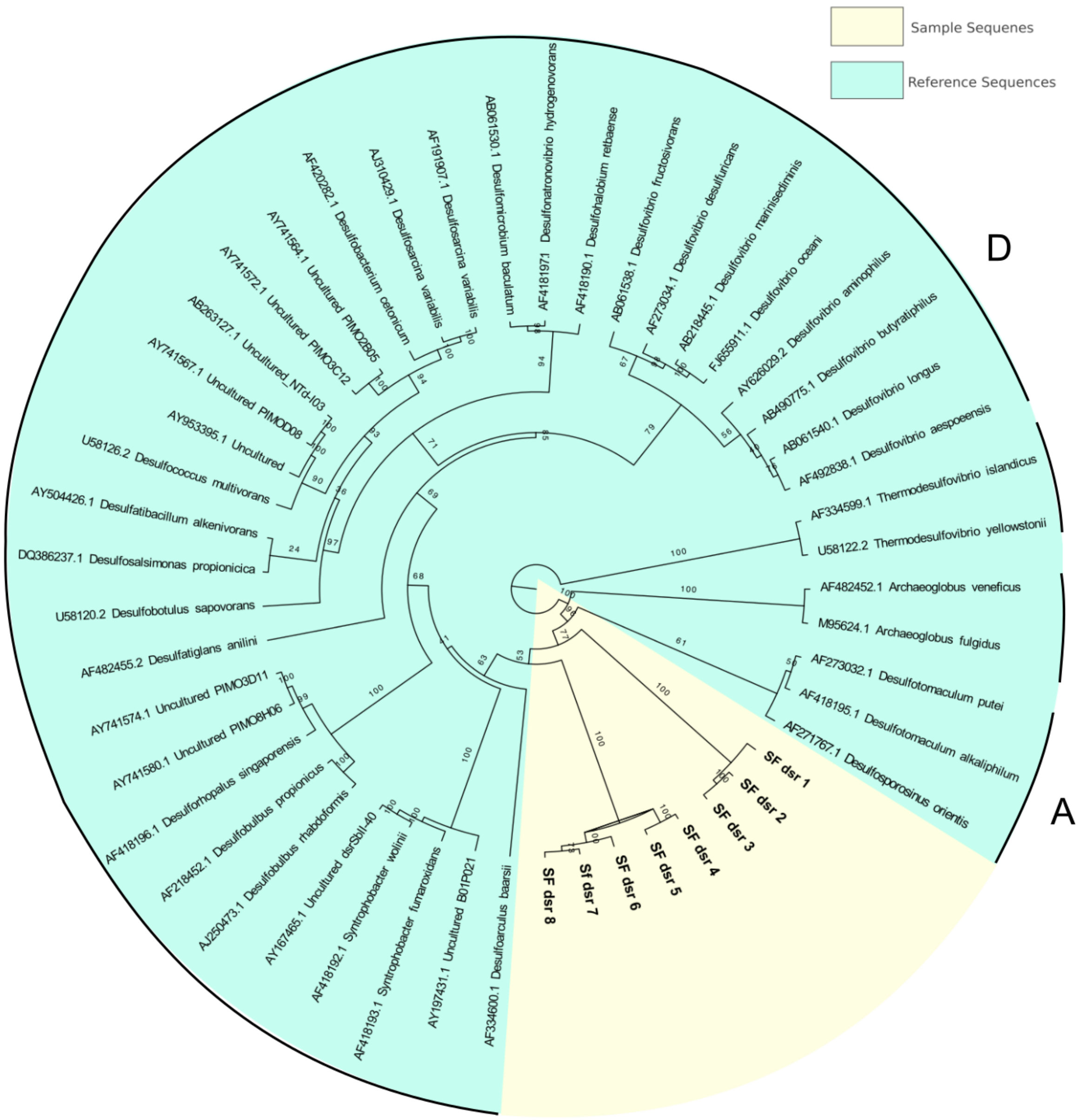

Sample Sequenes
Reference Sequences
D
C
B
A
AB061530.1 Desulfomicrobium baculatum
AF418197.1 Desulfonatronovibrio hydrogenovorans
AF418190.1 Desulfohalobium retbaense
AB061538.1 Desulfovibrio fructosivorans
AF273034.1 Desulfovibrio desulfuricans
AB218445.1 Desulfovibrio marinisediminis
FJ655911.1 Desulfovibrio oceani
AY626029.2 Desulfovibrio aminophilus
AB490775.1 Desulfovibrio butyratiphilus
AB061540.1 Desulfovibrio longus
AF492838.1 Desulfovibrio aespoeensis
AF334599.1 Thermodesulfovibrio islandicus
U58122.2 Thermodesulfovibrio yellowstonii
AF482452.1 Archaeoglobus veneficus
M95624.1 Archaeoglobus fulgidus
AF273032.1 Desulfotomaculum putei
AF418195.1 Desulfotomaculum alkaliphilum
AF271767.1 Desulfosporosinus orientis
SF dsr 1
SF dsr 2
SF dsr 3
SF dsr 4
SF dsr 5
SF dsr 6
Sf dsr 7
SF dsr 8
AF334600.1 Desulfoarculus baarsii
AY197431.1 Uncultured B01P021
AF418193.1 Syntrophobacter fumaroxidans
AF418192.1 Syntrophobacter wolinii
AY167465.1 Uncultured dsrStII-40
AJ250473.1 Desulfobulbus rhabdoformis
AF218452.1 Desulfobulbus propionicus
AF418196.1 Desulforhopalus singaporensis
AY741580.1 Uncultured PIMO8H06
AY741574.1 Uncultured PIMO3D11
AF482455.2 Desulfatiglans anilini
U58120.2 Desulfobotulus sapovorans
DQ386237.1 Desulfosalsimonas propionicica
AY504426.1 Desulfatibacillum alkenivorans
U58126.2 Desulfococcus multivorans
AY953395.1 Uncultured
AY741567.1 Uncultured PIMOD08
AB263127.1 Uncultured_NTd-I03
AY741572.1 Uncultured PIMO3C12
AY741564.1 Uncultured PIMO2B05
AF420282.1 Desulfobacterium cetonicum
AJ310429.1 Desulfosarcina variabilis
AF191907.1 Desulfosarcina variabilis
100
94
71
85
79
67
56
69
68
63
53
97
24
90
61
77
100

***Figure 5.7:*** *Maximum likelihood phylogeny of the Dissimilatory Sulfite Reductase (dsrAB) gene, including metagenome sample dsr sequences (bold and yellow, SF dsr 1 - 7) and reference genes (blue) obtained from NCBI GenBank and the phylogeny of Moreau et al., (2010). Reference sequences display the GenBank accession number, followed by the species name. Reference sequences are grouped in to classes A) Clostridia; B) Archaeglobi; C) Nitrospira and D) Deltaproteobacteria.*

***Table 5.11:*** *BLASTn matches for metagenome dsrAB sequences 1-7 against NCBI GenBank. The BLAST match name, GenBank accession number, % identity and score is provided.*

| Sample Name | Blast Match | Accession Number | % identity | Total Score |
|---|---|---|---|---|
| sf dsr 1 | *Sulfurifustis variabilis* | AP014936.1 | 93 | 686 |
| sf dsr 2 | *Desulfobulbus* sp. ORNL | CP021255.1 | 82 | 95.3 |
| sf dsr 3 | *Chlorobaculum limnaeum* strain DSM 1677 | CP017305.1 | 82 | 149 |
| sf dsr 4 | *Thioalkalivibrio paradoxus* ARh 1 | CP007029.1 | 78 | 431 |
| sf dsr 5 | *Sulfuricaulis limicola* DNA, complete genome | AP014879.1 | 76 | 490 |
| sf dsr 6 | *Thiohalobacter thiocyanaticus* | AP018052.1 | 91 | 854 |
| sf dsr 7 | *Thiohalobacter thiocyanaticus* | AP018052.1 | 82 | 848 |
| sf dsr 8 | *Thioflavicoccus mobilis* 8321 | CP003051.1 | 92 | 81 |

### 5.3.7 Implications for biogeochemical cycling

This analysis has used metagenomics to highlight the microbial community composition and functional potential of benthic fjord sediments. Whilst productivity assays would be needed to measure rates of activity, the presence of functional genes can be used to indicate potential biogeochemical implications of fjord microbial communities.

Interestingly, the fjord waters of Chilean Patagonia had been proposed as a hotspot of primary productivity, hosting three UNESCO bio reserves and support commercially important fisheries (Iriate *et al.,* 2007; Haussemann and Forsterra 2009; Niklitscheck *et al.,* 2013). However, the results from our analysis indicated a low nutrient environment, one which would not typically be classed as productive. In line with previous literature, we confirmed the presence of a surface freshwater lens, supplied by glacial runoff (Arancena *et al.,* 2011). This surface lens was low nutrient and high turbidity, attributed to the outflow of inorganic sediments from glacial runoff (Arancena *et al.,* 2011). We propose that the turbid surface waters may block light, limiting surface productivity and thus the export of organic matter to deep sediments. This is supported by previous research which has shown productivity to increase moving away from inner fjord waters, attributed to light and nutrient availability (Silva 2008; Arancena *et al.,* 2011; Gonzalez *et al.,* 2013). However, nutrient export from terrestrial surface waters may provide some limiting nutrients, helping to stimulate biological productivity to inner fjord waters (Gonzalez *et al.,* 2013).

With rising global temperatures, it is likely that glacial meltwater export will increase in this region. Increased meltwater export may reduce benthic and surface water biological productivity, through limiting light penetration and nutrient availability. Decreased surface productivity will not only limit nutrient export to benthic sediments, but may also have implications on wider ecosystem functioning and fisheries (Landaeta *et al.,* 2012; Gonzalez *et al.*, 2013).

Despite limitations on light and nutrient availability, we were able to detect genomic potential for microbial carbon, sulfur and nitrogen cycling in the sediments investigated. Potentially attributed to the extreme conditions, we found a combination of anaerobes, extremophiles and chemolithotrophs, who are resilient to the low light, oligotrophic and cold conditions. These results lay a foundation for the identification of potential novel genomes within the samples, which could relate to environmental selection for a unique subset of traits for survival. However, more exploration on the genomic level is needed to characterize these genomes and their biogeochemical functionality.

## 5.4 Conclusions

This study aimed to investigate the microbial community composition of the Steffen Fjord, Chilean Patagonia. This fjord sparked interest as it is fed by melt water from the Steffen glacier, and drains into the marine Baker fjord system. The lack of investigation in this region means that new microbial genomes, that have yet to be sequenced, may be in the sediments, with ecological, biotechnological or industrial uses.

The metadata showed the Steffen glacier to have a clear effect on fjord salinity, turbidity and temperature, with glacial freshwater runoff separated at the surface from the more dense marine bottom waters. The benthic sediments displayed low TN and TOC, indicating that nutrient limitation may be taking place. This indicates that increased freshwater runoff in future years may increase the fjord stratification, which could influence biological productivity. Overall, the sediment microbial community composition was largely dominated by bacteria, with archaea accounting for 3-5% of the metagenomes. The bacteria and archaea recovered related to anaerobic (and often methanogenic) taxa, indicating the presence of anoxia, or low oxygen zones within the surface sediments. Additionally, bacteria with adaptations or extremophilic qualities were detected, attributed to the oligotrophic, low oxygen, cold and dark conditions present in the sediments. Despite this, microbes relating to nitrogen and sulfur cycling organisms were found in the sediments, indicating the potential for microbial nutrient cycling in this low nutrient environment.  Genome binning was used to help identify the presence of potentially novel genomes within the samples. Over half the genome bins scored below 80% identity to NCBI GenBank. This indicates that some genomes found during binning could represent new strains or species. Additionally, microbes relating to sulfur cycling organisms were found frequently among the metagenomes, and bins relating to the ammonia oxidizing archaea, *Nitrosopumilus,* were common.

Sequences relating to *Nitrosopumilus* were selected for a 16s rRNA phylogeny with sequenced relatives from the phylogeny of Park *et al.,* (2012). The sample sequences were shown to have 99% similarity to the 16s sequences of sequenced *Nitrosopumilus* relatives *Nitrosopumilus oxclinae, Nitrosopumilus cobalaminigenes* and *Candidatus Nitrosopumilus sediminis* (Park *et al.,* 2012; Qin *et al.,* 2017). The presence of *Nitrosopumilus*-like genomes may be related to the high affinity to ammonia, and thereby resilience to the low TN of the sediments. Additionally, the presence of these genomes highlights the potential for nitrogen cycling in the sediments, despite the oligotrophy detected. Furthermore, the detection of sulfur cycling microbes within the sediments was used to drive a dsrAB phylogeny, for sulfite reducing organisms in the metagenomes. This provides an addition to 16s analysis, as it shows the presence of functional genes involved in sulfur cycling. The sulfite reducers were shown to constitute a discrete group within the phylogeny and had below a 91% identity to cultured relatives.

Overall the results highlight the presence of extremophiles and nutrient cycling microbes, despite the hostile conditions. The results demonstrate the presence of potentially unique microbes within the samples, which may have ecological or commercial implications. It is hoped that this work will stimulate further culture based analysis and single cell sequencing to fully isolate novel genomes, to facilitate understanding of environmental microbial diversity. Additionally, this analysis has shown the scope of metagenomics for investigating microbial communities, as it enables taxonomic and functional potential to be inferred.

### 5.5 Limitations and Further work

Whilst this analysis has provided an insight into the taxonomic diversity and functional potential of Patagonian fjord benthic microbial communities, it has also highlighted areas for future work. Firstly, a large fraction of the sequenced DNA was unannotated, attributed to either novelty or the presence of degraded DNA. To reveal the composition of more degraded DNA, amplification using short primers could be used, for example with metabarcoding (Taberlet *et al.,* 2012; Leray and Knowlton, 2015). This would increase the abundance of shorter DNA fragments for sequencing and annotation. Additionally, 16s rRNA amplicon sequencing and deep metagenome sequencing could also be used to uncover less abundant microbes, which may not have been recovered from this analysis. As sediment metagenomes are diverse, the sequencing carried out in this analysis is likely to focus on the most abundant fraction and may exclude some less abundant organisms (Rodriguez and Konstantinidis, 2014).

Deep metagenome sequencing may also be beneficial to recover complete genome sequences, which could be used to compile metagenome assembled draft genomes (MAGs). This would be largely related to an increase in sequencing coverage which would help resolve repeat regions during the assembly of genomes (Albertsen *et al.,* 2013). The recovery of MAGs would provide a deeper understanding of the metabolic functioning and strain level diversity of novel genomes. Culture based analysis and subsequent complete genome sequencing would also be beneficial for the *Nitrosopumilus* genomes. This would allow comparison of the average nucleotide identity (ANI) to current sequenced strains, to clarify if the species and strains recovered in this analysis are novel, alongside the growth conditions and physiology (Park *et al.,* 2012). Finally, radio-labelled $^{13}$C assays or meta-transcriptomics could also be used to identify if the organisms or functional pathways were active. As metagenomics only highlights the functional potential of the microbial community (rather than the active fraction), it would be interesting to investigate which pathways were more active than others in this unique environment.

# Chapter 6: Concluding discussion

## 6.1 Summary

Through the rise of DNA sequencing, our understanding of global microbial diversity has increased. Advances through 16s rRNA amplicon sequencing has aided understanding of microbial community composition in natural environments, without the need for culturing. More recently, metagenomics has been used in medical and environmental applications to directly sequence community DNA without the need for amplification. This method allows both the functional diversity and taxonomy of communities to be investigated and is especially beneficial for unique microbial samples. The more commonly-used 16s rRNA sequencing relies on publicly-available databases to assign taxa to DNA sequences, however these databases do not capture global microbial diversity. Whist taxonomy assignments can be made using metagenome data, the function of sequences or genomes that cannot be annotated with taxonomic databases can be investigated. Consequently, more information on the ecological and biogeochemical importance of the uncultured fraction of microbial communities can be understood.

Currently, limited guidance is available for metagenomic data analysis in the field of microbial ecology. Whilst raw metagenomic sequencing data can be directly analysed for taxonomy and function, assembling the DNA reads in to longer fragments (contigs) may be beneficial. DNA assembly can be carried out using a range of different assembly algorithms and aims to improve the alignment of DNA sequences to functional or taxonomic databases. DNA assembly is not commonly used in soil microbial ecology, and when it is, the choice of assembler and parameters often requires greater attention.

This study applied metagenomics to several areas in microbial ecology, to improve understanding of microbial function and taxonomy. Prior to this study, metagenomics had not been applied to understanding microbial succession in glacial forefields. Microbes have been proposed as the initial colonisers of newly exposed soil, but more research into how these communities change and the role they play in soil biogeochemical cycles was needed. In addition, the composition of the initial pioneer microbial community has been subject to debate, so metagenomic data may help to contribute to this discussion. This is significant given the continued retreat of glaciers, exposing more land for microbial colonisation. Furthermore, microbial nitrogen fixation (diazotrophy) has been previously identified as important for building up labile nitrogen stocks in oligotrophic forefield soils, facilitating the colonisation of higher microbes and plants during succession. However, little research had been carried out into the diversity of these diazotrophs and how they vary between forefields. In addition to this, a significant fraction of global microbial diversity has yet to be explored, including the glacially fed fjord sediments of Chilean Patagonia. These fjords are hotspots of primary productivity and support commercially important Salmon fisheries. Using metagenomics in this region will help facilitate understanding of microbial diversity, function and biogeochemical significance in these unexplored sediments.

Broadly, this study aimed to contribute to the use of metagenomics in the field of environmental microbial ecology, in terms of providing both methodological advancements and to broaden understanding of microbial diversity. In particular, the objectives of this study were to: 1) evaluate metagenome assemblers for soil microbial ecology, and subsequently apply metagenomics to investigate; 2) microbial nitrogen fixation in Arctic glacier forefields; 3) microbial diversity during succession in an Arctic forefield and 4) the diversity and functional potential of microbial communities in benthic fjord sediments. The following section provides an overview of the key findings:

**Objective 1:** To compare the performance of five publicly available metagenome assemblers for soil bacterial communities

This study found de Bruijn graph based assemblers (CLC and metaSPAdes) to provide the highest coverage and contig lengths during metagenome assembly of artificial soil bacterial datasets. However, due to the increased complexity of this algorithm, these assemblers are more sensitive to parameterisation, in particular, to components such as the kmer length. Therefore, de Bruijn graph-based assemblers can provide high quality assemblies, but also produce the largest spread of values. Consequently, testing parameter values to suit the dataset in question may improve the output quality, over simply using default values. Additionally, this study also showed the importance of evaluating assembly quality using multiple metrics, covering assembly contiguity, size and completeness. Using a spread of metrics helps to avoid large fragmented datasets or long erroneous contigs, which would not be optimal for downstream analysis. Additionally, the use of artificial test datasets was shown to be beneficial in evaluating the accuracy of metagenome assembly in producing the correct taxonomic composition.

Overall, the study found assembler selection to have a significant influence on outcome quality. This has importance for the microbial ecology community, as it shows assembler selection is an important factor to consider during sequence analysis. Fitting the assembler (and parameterisation) to the complexity of the dataset can improve assembly outcome and thus taxonomic and functional annotation. It is therefore recommended that assembler comparison, or justification of assembler choice is made during methodology development.

**Objective 2:** To investigate the similarities and differences in taxonomic composition of diazotrophic bacteria in metagenomes sampled from four Arctic glacier forefields

This study applied the assembler selected in Objective 1 (metaSPAdes) to investigate the diversity of nitrogen-fixingbacteria across four Arctic forefields (Storglaciären, Rabots, Russell and Midtre Lovénbreen). Metagenome assembly was applied to provide longer contigs for taxonomic annotation of nif genes (for nitrogen fixation). The study found a diverse range of diazotrophs across the forefields, including a core group of cyanobacteria, anaerobes and extremophiles, which were identified across sites. The composition of this core group may be related to adaptive mechanisms, including tolerance to the oligotrophic, high UV and cold conditions that are typical in forefield soils. This analysis provided a new nifH phylogeny, demonstrating the phylogenetic distribution of Arctic diazotrophs, in relation to sequenced relatives. The study contributes to our understanding of microbial diversity in the Arctic, including how a range of bacterial species may contribute to local biogeochemical cycling through nitrogen fixation. In addition, the study exemplified how metagenomics can be applied to functional and taxonomic analysis of microbial communities in extreme environments. The use of metagenomics allowed the taxonomy of functional (nitrogen fixation, nif) genes to be interpreted, which would not be possible with 16s analysis or culture-based methodologies. It is hoped this work will stimulate others in the microbial ecology community to apply metagenomics to widen understanding of microbial diversity.

**Objective 3:** To investigate the bulk microbial community composition along a chronosequence of soil succession in the Midtre Lovénbreen forefield, Svalbard, using metagenomics.

Objective 3 applied metagenomics to understand microbial community composition during soil succession along the Midtre Lovénbreen forefield, Svalbard. In this study, unassembled DNA sequencing was used for community taxonomic annotation, due to limitations of assembly resulting from the community complexity and sequencing coverage. However, sequences were assembled with metaSPAdes and used for functional interpretation and genome binning, to benefit from longer contig lengths.

During succession, forefield soil total nitrogen and total organic carbon content increased, attributed to allochthonous and autochthonous sources. This included both aeolian deposition and the fixation of carbon and nitrogen by autotrophic bacteria. Throughout the forefield, aerobic, anaerobic and extremophilic species were identified, attributed to the low nutrient, high UV and cold conditions, in line with Objective 2. Newly-exposed soils were shown to contain both autotrophic cyanobacteria and heterotrophs, providing further information on the current debate surrounding the composition of pioneer communities. The presence of heterotrophic bacteria may relate to the identification of some organic carbon in early soils, which may be sourced from overridden material or aeolian deposition. Furthermore, from 3 years post ice retreat, organisms with carbon, nitrogen and sulfur cycling metabolisms were identified, in line with soil nutrient build up. Genome binning was used to support the taxonomy and also indicated the potential for novel strains and species in the soils.

Overall, this study has used metagenomics to aid understanding of how microbial community structure (and function) modified during soil succession. This is significant for highlighting how microbes may contribute to soil development and local biogeochemical cycles in the Arctic, a region which was once thought to be abiotic. This is especially significant given the expected increase in soil exposure in the Arctic with glacier retreat in future years, and thus, microbial colonisation. Furthermore, this study has also shown how metagenome assembly, whilst beneficial for functional analysis, may not always be the most useful approach. Given datasets that are highly complex or have low read coverage, read-based taxonomic annotation may provide a better overview, than that based on incomplete assemblies.

**Objective 4:** To investigate the composition and potential function of microbial communities sourced from benthic metagenomes in a Chilean fjord.

Objective 4 applied metagenomics to understand the microbial community composition of benthic fjord sediments, fed by glacial meltwater. The glacial outflow from the Steffen glacier has a distinct influence on fjord salinity, turbidity and temperature. However, as this freshwater was stratified as a surface lens, there was limited influence on the benthic sediments. These sediments were in more saline (marine) conditions, with very low total organic carbon and nitrogen content. Alongside metagenome assembly, read-based annotation was also carried out, due to the low read recruitment in assemblies, which may be due to limited DNA quality from this extreme sampling location. This highlights that metagenome assembly may not always be the best approach and the data type/quality in question needs to be considered. A substantial fraction of the DNA reads could not be annotated, which again may be related to DNA quality or the presence of unique genomes in the samples. The majority of the recovered community composition consisted of bacteria, including anaerobes and those with potential extremophilic adaptations, to the dark, cold, oligotrophic conditions. Results from genome binning revealed that over half the genomes did not match to NCBI GenBank accurately, indicating a degree of species novelty may be present in the samples. This provides a focus for further single cell sequencing or culture-based studies, that may wish to isolate these genomes. Additionally, 16s rRNA and drsAB gene phylogenies where applied to demonstrate the potential of metagenomics for both taxonomic and functional analysis.

Overall, this study has shown how metagenomics can provide useful insights into community composition and function of uncultured, potentially unique, samples. This technology can provide insights into microbial diversity, which may help to focus further analysis, such as culturing or deep re-sequencing. However, it has also been shown how metagenome assembly may not always be the most optimal tool, given a dataset that is highly diverse or fragmented. It is therefore suggested that methodologies are constructed with consideration of the dataset at hand.

### 6.2 Limitations and future opportunities

The studies here have demonstrated metagenomics to be a useful tool for understanding microbial diversity and functional potential. However, metagenomics is unable to provide information on the activity of microbes or functional pathways because it is based on DNA sequencing (Wooley *et al.,* 2010). Transcriptomics may be a useful addition to metagenomic analysis, which involves sequencing the microbial community mRNA (Moran *et al.,* 2013). Transcriptomics is therefore able to provide an insight into the functional pathways which are active at a snapshot in time. Whilst this cannot be extrapolated to daily (or yearly) activities, it is useful to gauge if functional pathways are indeed active. Incubation experiments may also be carried out to measure rates of productivity, for example through acetylene reduction assays for microbial nitrogen fixation (Burris *et al*., 1972; Hardy *et al.,*1973).

Furthermore, sequencing coverage can also be an issue, especially for complex microbial communities, as shown in this analysis. It is difficult to determine the level of sequencing required *a priori*, and therefore all genomes in diverse community samples may not be recovered. To recover the less abundant microbial fraction, or to extract metagenome assembled genomes, deep resequencing can be carried out to improve the coverage (Narasingarao *et al.,* 2012). However, this is not necessary if the analysis aims to only gauge an insight into the dominant microbial taxonomy and functions.

Currently, there is a lack of formalised guidance for the use of metagenomics by microbial ecologists. Multiple stages during the analysis require informed judgement, for example; to assemble the metagenome or not, the choice of assembler, to undertake genome binning or to target specific functional genes. If metagenomics is used to a greater degree in future, a methodological consensus or guidance may be achieved. This study has contributed by evaluating the use of assembly and the choice of assemblers for soil microbial ecologists. However, the development of more formalised tutorials, guidance and analysis platforms would be beneficial to improve uptake of metagenomics by the research community. More recently, platforms such as KBASE have moved towards this, by providing interactive user interfaces for scientists without substantial bioinformatic experience.

### 6.3 Implications

The results from this analysis have demonstrated the potential of metagenomics for understanding microbial taxonomy and functional potential within the field of environmental microbial ecology. Not only can metagenomics provide insights into the diversity and potential role of microbes in local biogeochemical cycles, it can also help guide further analysis, such as single cell sequencing, genome assembly and culturing. It is hoped that this work will help to demonstrate the benefit of metagenomics, which can be used as an alternative to, or alongside, 16s or targeted gene sequencing.

Here, we have used metagenomics to contribute to the understanding of microbial ecology in glaciated regions. In particular, the work has improved understanding of microbial function and diversity in glacial forefields, which has particular significance given continued ice retreat with climate change. This analysis has identified distinct changes to microbial diversity and functional potential during forefield succession, and supplied a new nifH phylogeny for forefield diazotrophs. In addition, metagenomics has provided new insights into microbial community structure and function in deep low biomass Chilean fjord sediments, a largely unexplored environment. We have also discussed the benefit of metagenome assembly and how the choice of assembler can impact the output quality and downstream analysis. It has been suggested that studies which use metagenomics consider assembler choice with respect to the dataset, and acknowledge that assembly may not always be the best approach for taxonomic analysis given highly diverse or fragmented metagenomes. This is because low abundance sequences may be excluded from sequencing in complex samples, and therefore the full diversity may not be profiled. This work has identified de Bruijn graph based assemblers as the most appropriate choice for soil metagenome assemblies, however consideration needs to made to the data type, complexity and choice of parameters during assembler selection. Our recommendations include careful assembler selection, testing and parameter optimisation to provide an improved assembly outcome for downstream annotation.

It is hoped that the application of metagenomics in this work will encourage those in the field of soil microbial ecology to explore this technique. However, we suggest that scientists do so with methodological care and caution. Further methodological inter-comparisons and guidance may encourage the use metagenomics in microbial ecology. This is because there is limited formalised structure available for metagenomic analysis, in contrast to the body of literature using 16s rRNA sequencing. Metagenomics is a high useful tool for understanding the uncultured microbial fraction and can help guide further studies and more detailed analysis. This not only has application and benefit for understanding global microbial diversity, but can be beneficial for applied studies, for example in biotechnology or anti-microbial resistance. Whilst metagenomics is a highly beneficial tool for understanding microbial diversity, it is best applied in a holistic approach, incorporating culture based studies, modelling and microscopy, to gain a full understanding of community structure and function.

# References

Abatenh, E., Gizaw, B., Tsegaye, Z., & Tefera, G. (2018). Microbial function on climate change - a review. *Environmental Pollution and Climate Change*, *2*(147), 10-4172.

Akob, D. M., Baesman, S. M., Sutton, J. M., Fierst, J. L., Mumford, A. C., Shrestha, Y., ... & Oremland, R. S. (2017). Detection of diazotrophy in the acetylene-fermenting anaerobe *Pelobacter* sp. strain SFB93. *Applied and Environmental Microbiology*, 83(17).

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology,* 31(6), 533.

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods,* 8(1), 61-65.

Allison, S. D., & Treseder, K. K. (2008). Warming and drying suppress microbial activity and carbon cycling in boreal forest soils. *Global Change Biology,* 14(12), 2898-2909.

Allison, V. J., Condron, L. M., Peltzer, D. A., Richardson, S. J., & Turner, B. L. (2007). Changes in enzyme activities and soil microbial community composition along carbon and nutrient gradients at the Franz Josef chronosequence, New Zealand. *Soil Biology and Biochemistry*, 39(7), 1770-1781

Anesio, A. M., Hodson, A. J., Fritz, A., Psenner, R., & Sattler, B. (2009). High microbial activity on glaciers: importance to the global carbon cycle. *Global Change Biology,* 15(4), 955-960.

Ansari, A. H., Hodson, A. J., Kaiser, J., & Marca-Bell, A. (2013). Stable isotopic evidence for nitrification and denitrification in a High Arctic glacial ecosystem. *Biogeochemistry,* 113(1-3), 341-357.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4), 195-203.

Aracena, C., Lange, C. B., Iriarte, J. L., Rebolledo, L., & Pantoja, S. (2011). Latitudinal patterns of export production recorded in surface sediments of the Chilean Patagonian fjords (41–55 S) as a response to water column productivity. *Continental Shelf Research*, *31*(3-4), 340-355.

Arkin, A. P., Stevens, R. L., Cottingham, R. W., Maslov, S., Henry, C. S., Dehal, P., ... & Sneddon, M. W. (2016). The DOE systems biology knowledgebase (KBase). *bioRxiv,* 096354.

Arumugam, M., Harrington, E.D., Foerstner K.U., Raes, J., Bork, P. (2010) SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics,* 26, 2977–2978

Aulakh, M. S., Doran, J. W., & Mosier, A. R. (1992). 'Soil denitrification—significance, measurement, and effects of management'. In Stewart, B.A. *Advances in Soil Science.* Springer: New York, 1-57.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Meyer, F. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 9(1), 75.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pyshkin, A. V. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5), 455-477.

Banning, N. C., Maccarone, L. D., Fisk, L. M., & Murphy, D. V. (2015). Ammonia-oxidising bacteria not archaea dominate nitrification activity in semi-arid agricultural soil. *Scientific Reports*, 5, 11146.

Barber, S. A. (1995) *Soil nutrient bioavailability: a mechanistic approach*. John Wiley & Sons: USA.

Bardgett, R. D., Richter, A., Bol, R., Garnett, M. H., Bäumler, R., Xu, X., ... & Wanek, W. (2007). Heterotrophic microbial communities use ancient carbon following glacial retreat. *Biology Letters,* 3(5), 487-490.

Bastide, M., & McCombie, W. R. (2007). Assembling genomic DNA sequences with PHRAP. *Current Protocols in Bioinformatics,* 11-4.

Benn, D., & Evans, D. J. (2014). *Glaciers and Glaciation*. Routledge: London.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research*, 33, 34-38.

Benson, D. R., & Silvester, W. B. (1993). Biology of *Frankia* strains, actinomycete symbionts of actinorhizal plants. *Microbiological Reviews,* 57(2), 293-319.

Bergman, B., Gallon, J. R., Rai, A. N., & Stal, L. J. (1997). $N_2$ fixation by non-heterocystous cyanobacteria. *FEMS Microbiology Reviews,* 19(3), 139-185.

Bernasconi, S. M. (2008). Weathering, soil formation and initial ecosystem evolution on a glacier forefield: a case study from the Damma Glacier, Switzerland. *Mineralogical Magazine*, 72(1), 19-22.

Bernasconi, S. M., Bauder, A., Bourdon, B., Brunner, I., Bünemann, E., Chris, I., ... & Frossard, E. (2011). Chemical and biological gradients along the Damma glacier soil chronosequence, Switzerland. *Vadose Zone Journal*, 10(3), 867-883.

Bhatia, M., Sharp, M., & Foght, J. (2006). Distinct bacterial communities exist beneath a high Arctic polythermal glacier. *Applied and Environmental Microbiology*, 72(9), 5838-5845.

Blainey, P. C. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews,* 37(3), 407-427

Błażewicz, J., Formanowicz, P., Kasprzak, M., Schuurman, P., & Woeginger, G. J. (2002*).* DNA sequencing, Eulerian graphs and the exact perfect matching problem'. In Brandstadt, A., *Graph-Theoretic Concepts in Computer Science*. Springer: Berlin Heidelberg, 13-24.

Blothe, M., Akob, D. M., Kostka, J. E., Goschel, K., Drake, H. L. & Kusel, K. (2008). pH gradient-induced heterogeneity of Fe(III)-reducing microorganisms in coal mining-associated lake sediments. *Applied Environmental Microbiology*, 74,1019–1029.

Bohmann, Kristine, Alice Evans, M. Thomas P. Gilbert, Gary R. Carvalho, Simon Creer, Michael Knapp, W. Yu Douglas, and Mark De Bruyn. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29 (6), 358-367.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bonin, A. S., & Boone, D. R. (2006). 'The order methanobacteriales'. In: Dworkin, M., Falkow, S., Schleifer K.H., Stackebrandt, E. (Eds). *The Prokaryotes.* Springer: New York, 231-243.

Boopathy, R., and Kulpa, C. F. (1993). Nitroaromatic compounds serve as nitrogen source for *Desulfovibrio* sp.(B strain). *Canadian Journal of Microbiology,* 39(4), 430-433.

Bottrell, S. H., & Tranter, M. (2002). Sulfide oxidation under partially anoxic conditions at the bed of the Haut Glacier d'Arolla, Switzerland. *Hydrological Processes*, *16*(12), 2363-2368.

Boyd, E. S., Skidmore, M., Mitchell, A. C., Bakermans, C., & Peters, J. W. (2010). Methanogenesis in subglacial sediments. *Environmental Microbiology Reports,* 2(5), 685-692.

Bradley J, Singarayer J, Anesio A. (2014) Microbial community dynamics in the forefield of glaciers. *Proceedings of the Royal Society B*, 281.

Bradley, J. A., Anesio, A. M., Singarayer, J. S., Heath, M. R., and Arndt, S. (2015). SHIMMER (1.0): a novel mathematical model for microbial and biogeochemical dynamics in glacier forefield ecosystems. *Geoscientific Model Development Discussions,* 8(8), 6143-6216.

Bradley, J. A., Arndt, S., Sabacká, M., Benning, L. G., Barker, G. L., Blacker, J. J., ... & Tranter, M. (2016). Microbial dynamics in a High Arctic glacier forefield: a combined field, laboratory, and modelling approach. *Biogeosciences,* 13(19), 5677.

Brankatschk, R., Töwe, S., Kleineidam, K., Schloter, M., & Zeyer, J. (2011). Abundances and potential activities of nitrogen cycling microbial communities along a chronosequence of a glacier forefield. *The ISME Journal,* 5(6), 1025-1037.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... & Jovanovich, S. B. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10), 1146-1153.

Bräutigam, A., Mullick, T., Schliesky, S., & Weber, A. P. (2011). Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. *Journal of Experimental Botany*, 62(9), 3093-3102.

Breen, K., & Levesque, E. (2006). Proglacial succession of biological soil crusts and vascular plants: biotic interactions in the High Arctic. *Botany*, 84(11), 1714-1731.

Bremner, J. M. (1965). Organic nitrogen in soils. *Soil Nitrogen,* 93-149.

Brill, W. J. (1975). Regulation and genetics of bacterial nitrogen fixation. *Annual Reviews in Microbiology*, 29(1), 109-129.

Broecker, W. S., & Peng, T. H. (1974). Gas exchange rates between air and sea. Tellus, 26(1-2), 21-35.

Brunner, I., Plötze, M., Rieder, S., Zumsteg, A., Furrer, G., and Frey, B. (2011). Pioneering fungi from the Damma glacier forefield in the Swiss Alps can promote granite weathering. *Geobiology,* 9(3), 266-279.

Burris, R. H. (1972). 'Nitrogen fixation—Assay methods and techniques'. In: Colowick, S.P. *Methods in enzymology*. Elsevier: Netherlands, 24, 415-431.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., ... & Jaffe, D. B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Research*, 18(5), 810-820.

Caccavo, F., Lonergan, D. J., Lovley, D. R., Davis, M., Stolz, J. F., & McInerney, M. J. (1994). *Geobacter sulfurreducens* sp. nov., a hydrogen-and acetate-oxidizing dissimilatory metal-reducing microorganism. *Applied and Environmental Microbiology,* 60(10), 3752-3759.

Cameron, K. A., Hodson, A. J., & Osborn, A. M. (2012). Carbon and nitrogen biogeochemical cycling potentials of supraglacial cryoconite communities. *Polar Biology*, 35(9), 1375-1393.

Carvalhais, L. C., Dennis, P. G., Badri, D. V., Tyson, G. W., Vivanco, J. M., & Schenk, P. M. (2013). Activation of the jasmonic acid plant defence pathway alters the composition of rhizosphere bacterial communities. *PLoS One*, 8(2), e56457.

Cattonaro, F., Policriti, A., & Vezzi, F. (2010). *Enhanced reference guided assembly.* Bioinformatics and Biomedicine (BIBM). IEEE International Conference.

Chaia, E. E., Wall, L. G., & Huss-Danell, K. (2010). Life in soil by the actinorhizal root nodule endophyte Frankia. A review. *Symbiosis*, 51(3), 201-226.

Challacombe, J. F., Majid, S., Deole, R., Brettin, T. S., Bruce, D., Delano, S. F., ... & Reitenga, K. G. (2013). Complete genome sequence of *Halorhodospira halophila* SL1. *Standards in Genomic Sciences*, 8(2), 206.

Chapin, F. S., Walker, L. R., Fastie, C. L., & Sharman, L. C. (1994). Mechanisms of primary succession following deglaciation at Glacier Bay, Alaska. *Ecological Monographs,* 64(2), 149-175.

Chattopadhyay, M. K. (2006). Mechanism of bacterial adaptation to low temperature. *Journal of Biosciences*, 31(1), 157-165.

Chen M,A., Markowitz, V.M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M. (2017). IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research,* 45.

Chevreux (2014) MIRA [Online] Available from: http://www.chevreux.org/index0.html [Accessed 25-01-2018].

Chien, Y. T., and Zinder, S. H. (1996). Cloning, functional organization, transcript studies, and phylogenetic analysis of the complete nitrogenase structural genes (nifHDK2) and associated genes in the archaeon *Methanosarcina barkeri* 227. *Journal of Bacteriology*, 178(1), 143-148.

Chikhi, R., & Medvedev, P. (2013). Informed and automated k-mer size selection for genome assembly. *Bioinformatics,* 30(1), 31-37.

Childers, Susan E., Stacy Ciufo, and Derek R. Lovley. (2002). *Geobacter metallireducens* accesses insoluble Fe (III) oxide by chemotaxis. *Nature,* 416 (6882), 767-769.

Chrismas, N. A., Barker, G., Anesio, A. M., & Sánchez-Baracaldo, P. (2016). Genomic mechanisms for cold tolerance and production of exopolysaccharides in the Arctic cyanobacterium *Phormidesmis priestleyi* BC1401. *BMC Genomics,* 17(1), 533.

Christner, B. C., Kvitko, B. H., & Reeve, J. N. (2003). Molecular identification of bacteria and eukarya inhabiting an Antarctic cryoconite hole. *Extremophiles,* 7(3), 177-183.

Christner, B. C., Skidmore, M. L., Priscu, J. C., Tranter, M., & Foreman, C. M. (2008). 'Bacteria in subglacial environments'. In: Margesin, R., Schinner, F., Marx, J.C., Gerday, C. (Eds) *Psychrophiles: from biodiversity to biotechnology*. Springer: Berlin Heidelberg, 51-71.

Chu, H., Fierer, N., Lauber, C. L., Caporaso, J. G., Knight, R., & Grogan, P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environmental Microbiology,* 12(11), 2998-3006.

Chu, V. W. (2014). Greenland ice sheet hydrology: A review. *Progress in Physical Geography*, *38*(1), 19-54.

Collins, D. N. (1979). Hydrochemistry of meltwaters draining from an alpine glacier. *Arctic and Alpine Research,* 307-324.

Conen, F., Yakutin, M. V., Zumbrunn, T., & Leifeld, J. (2007). Organic carbon and microbial biomass in two soil development chronosequences following glacial retreat. *European Journal of Soil Science*, *58*(3), 758-762.

Cong, J., Yang, Y., Liu, X., Lu, H., Liu, X., Zhou, J., ... & Zhang, Y. (2015). Analyses of soil microbial community compositions and functional genes reveal potential consequences of natural forest succession. *Scientific Reports,* 5.

Conway, T., Wazny, J., Bromage, A., Zobel, J., & Beresford-Smith, B. (2012). Gossamer—a resource-efficient de novo assembler. *Bioinformatics,* 28(14), 1937-1938.

Copeland, A., Zeytun, A., Yassawong, M., Nolan, M., Lucas, S., Hammon, N., ... & Goodwin, L. A. (2012). Complete genome sequence of the orange-red pigmented, radioresistant *Deinococcus proteolyticus* type strain (MRP T). *Standards in Genomic Sciences,* 6(2), 240.

Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., & Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. *Trends in Biotechnology*, 23(6), 321-329.

Cowton, T., Nienow, P., Bartholomew, I., Sole, A., & Mair, D. (2012). Rapid erosion beneath the Greenland ice sheet. *Geology,* 40(4), 343-346.

D'Amico, M. E., Freppaz, M., Filippa, G., & Zanini, E. (2014). Vegetation influence on soil formation rate in a proglacial chronosequence (Lys Glacier, NW Italian Alps). *Catena,* 113, 122-137.

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., ... & Kirkegaard, R. H. (2015). Complete nitrification by *Nitrospira* bacteria. *Nature,* 528(7583), 504.

Daniel, R. (2005). The metagenomics of soil. *Nature Reviews Microbiology*, 3(6), 470-478.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods,* 9(8), 772.

DasSarma, S., & DasSarma, P. (2006). Halophiles. *Engineering in Life Sciences*, 1-13.

Davidson, E. A., & Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature*, 440(7081), 165-173.

Davila, P.M., D. Figueroa, & E. Muller. (2002). Freshwater input into the coastal ocean and its relation with the salinity distribution off austral Chile. *Continental Shelf Research,* 22(3): p. 521-534.

Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., ... & Vogel, T. M. (2011). Accessing the soil metagenome for studies of microbial diversity. *Applied and Environmental Microbiology,* 77(4), 1315-1324.

Deng, X., Naccache, S. N., Ng, T., Federman, S., Li, L., Chiu, C. Y., & Delwart, E. L. (2015). An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. *Nucleic Acids Research,* 43(7), e46-e46.

Denison, R. F., & Kiers, E. T. (2004). Why are most rhizobia beneficial to their plant hosts, rather than parasitic?. *Microbes and Infection*, 6(13), 1235-1239.

Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., & Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. *Bioinformatics,* 24(8), 1035-1040.

Deslippe, J. R., and Egger, K. N. (2006). Molecular diversity of nifH genes from bacteria associated with high arctic dwarf shrubs. *Microbial Ecology*, 51(4), 516-525.

Dethier, M. N., & Schoch, G. C. (2005). The consequences of scale: assessing the distribution of benthic populations in a complex estuarine fjord. *Estuarine, Coastal and Shelf Science*, 62(1), 253-270.

Dixon, R. and Kahn, D. (2004). Genetic regulation of biological nitrogen fixation. Nature reviews *Microbiology*, 2.8.

Dreyfus, B., Garcia, J. L., and Gillis, M. (1988). Characterization of Azorhizobium caulinodans gen. nov., sp. nov., a stem-nodulating nitrogen-fixing bacterium isolated from Sesbania rostrata. *International Journal of Systematic and Evolutionary Microbiology*, 38(1), 89-98.

Duc, L., Noll, M., Meier, B. E., Bürgmann, H., and Zeyer, J. (2009). High diversity of diazotrophs in the forefield of a receding alpine glacier. *Microbial Ecology,* 57(1), 179-190.

Dutton, A., Carlson, A. E., Long, A. J., Milne, G. A., Clark, P. U., DeConto, R., ... & Raymo, M. E. (2015). Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science*, 349(6244).

Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., ... & Nguyen, N. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research,* 21(12), 2224-2241.

Edwards, A., & Cook, S. (2015). Microbial dynamics in glacier forefield soils show succession is not just skin deep. *Molecular Ecology,* 24(5), 963-966.

Egli, M., Filip, D., Mavris, C., Fischer, B., Götze, J., Raimondi, S., & Seibert, J. (2012). Rapid transformation of inorganic to organic and plant-available phosphorous in soils of a glacier forefield. *Geoderma*, *189*, 215-226.

Eisen, J. A., Nelson, K. E., Paulsen, I. T., Heidelberg, J. F., Wu, M., Dodson, R. J., ... & Hickey, E. K. (2002). The complete genome sequence of Chlorobium tepidum TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proceedings of the National Academy of Sciences*, 99(14), 9509-9514.

Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., & Kyrpides, N. C. (2016a). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology,* 1(4), 15032.

Eloe-Fadrosh, E. A., Paez-Espino, D., Jarett, J., Dunfield, P. F., Hedlund, B. P., Dekas, A. E., ... & Li, W. J. (2016b). Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nature communications,* 7, 10476.

Erlich, H. A. (1989). *PCR technology.* Stockton press: UK.

Fagerli, I. L., & Svenning, M. M. (2005). Arctic and subarctic soil populations of Rhizobium leguminosarum biovar trifolii nodulating three different clover species: characterisation by diversity at chromosomal and symbiosis loci. *Plant and soil,* 275(1), 371-381.

Fastie, C. L. (1995). Causes and ecosystem consequences of multiple pathways of primary succession at Glacier Bay, Alaska. *Ecology*, 76(6), 1899-1916.

Ferrer, M., Martínez-Abarca, F., & Golyshin, P. N. (2005). Mining genomes and 'metagenomes' for novel catalysts. *Current Opinion in Biotechnology,* 16(6), 588-593.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., ... and Knight, R. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology,* 73(21), 7059-7066.

Fierer, N., Jackson, J. A., Vilgalys, R., & Jackson, R. B. (2005). Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. *Applied and Environmental Microbiology,* 71(7), 4117-4120.

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., ... & Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences,* 109(52), 21390-21395.

Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., ... & Toppo, S. (2011). Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Briefings in Bioinformatics*, bbr063.

Flemming, H. C., & Wingender, J. (2010). The biofilm matrix. *Nature Reviews Microbiology*, 8(9), 623.

Franche, C., Lindström, K., & Elmerich, C. (2009). Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. *Plant and Soil,* 321(1-2), 35-59.

Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E., & Oakley, B. B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences*, 102(41), 14683-14688.

Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., & Huttenhower, C. (2015). Sequencing and beyond: integrating molecular'omics' for microbial community profiling. *Nature Reviews Microbiology,* 13(6), 360.

Freeman, K,R., Pescador, M,Y., Reed, S,C., Costello, E,K., Robeson, M,S., Schmidt, S,K. (2009) Soil CO2 flux and photoautotrophic community composition in high-elevation, 'barren' soil. *Environmental Microbiology,* 11, 674–686.

Frenot, Y., Gloaguen, J. C., Cannavacciuolo, M., & Bellido, A. (1998). Primary succession on glacier forelands in the subantarctic Kerguelen Islands. *Journal of Vegetation Science,* 9(1), 75-84.

Frey, B., Bühler, L., Schmutz, S., Zumsteg, A., & Furrer, G. (2013). Molecular characterization of phototrophic microorganisms in the forefield of a receding glacier in the Swiss Alps. *Environmental Research Letters,* 8(1), 015033.

Frey, B., Rieder, S. R., Brunner, I., Plötze, M., Koetzsch, S., Lapanje, A., ... & Furrer, G. (2010). Weathering-associated bacteria from the Damma glacier forefield: physiological capabilities and impact on granite dissolution. *Applied and Environmental Microbiology,* 76(14), 4788-4796.

Gaby, J. C., & Buckley, D. H. (2012). A comprehensive evaluation of PCR primers to amplify the nifH gene of nitrogenase. *PloS one,* 7(7), e42149.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., & Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols,* (1).

González, H. E., Castro, L. R., Daneri, G., Iriarte, J. L., Silva, N., Tapia, F., ... & Vargas, C. A. (2013). Land–ocean gradient in haline stratification and its effects on plankton dynamics and trophic carbon fluxes in Chilean Patagonian fjords (47–50 S). *Progress in Oceanography*, 119, 32-47.

Göransson, H., Venterink, H. O., & Bååth, E. (2011). Soil bacterial growth and nutrient limitation along a chronosequence from a glacier forefield. *Soil Biology and Biochemistry*, 43(6), 1333-1340.

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *Journal of Investigative Dermatology,* 133(8), e11.

Gruber, N. (2008). The marine nitrogen cycle: overview and challenges. *Nitrogen in the Marine Environment*, 2, 1-50.

Guelland, K., Hagedorn, F., Smittenberg, R. H., Göransson, H., Bernasconi, S. M., Hajdas, I., & Kretzschmar, R. (2013). Evolution of carbon fluxes during initial soil formation along the forefield of Damma glacier, Switzerland. *Biogeochemistry,* 113(1-3), 545-561.

Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology,* 52, 696-704.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29(8), 1072-1075.

Gutiérrez, M. H., Galand, P. E., Moffat, C., & Pantoja, S. (2015). Melting glacier impacts community structure of Bacteria, Archaea and Fungi in a Chilean Patagonia fjord. *Environmental Microbiology,* 17(10), 3882-3897.

Hahn, A. S., & Quideau, S. A. (2013). Shifts in soil microbial community biomass and resource utilization along a Canadian glacier chronosequence*. Canadian Journal of Soil Science*, 93(3), 305-318.

Hajirasouliha, I., Hormozdiari, F., Alkan, C., Kidd, J. M., Birol, I., Eichler, E. E., & Sahinalp, S. C. (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics,* 26(10), 1277-1283.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews,* 68(4), 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10).

Hardy, R., Burns, R. C., & Holsten, R. D. (1973). Applications of the acetylene-ethylene assay for measurement of nitrogen fixation. *Soil Biology and Biochemistry*, 5(1), 47-81.

Hawkings, J. R., Wadham, J. L., Tranter, M., Lawson, E., Sole, A., Cowton, T., ... & Telling, J. (2015). The effect of warming climate on nutrient and solute export from the Greenland Ice Sheet. Geochemical Perspective Letters, 1, 94-104.

Hawkings, J. R., Wadham, J. L., Tranter, M., Raiswell, R., Benning, L. G., Statham, P. J., ... & Telling, J. (2014). Ice sheets as a significant source of highly reactive nanoparticulate iron to the oceans. *Nature Communications,* 5.

He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H., and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic acids research*, 44(W1).

Hedges, J. I., and Stern, J. H. (1984). Carbon and nitrogen determinations of carbonate-containing solids. *Limnology and Oceanography*, 29(3), 657-663.

Hernandez, D., François, P., Farinelli, L., Østerås, M., & Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research,* 18(5), 802-809.

Hodkinson, I. D., Coulson, S. J., & Webb, N. R. (2003). Community assembly along proglacial chronosequences in the high Arctic: vegetation and soil development in north-west Svalbard. *Journal of Ecology,* 91(4), 651-663.

Hodkinson, I. D., Webb, N. R., & Coulson, S. J. (2002). Primary community assembly on land–the missing stages: why are the heterotrophic organisms always there first? *Journal of Ecology*, 90(3), 569-577.

Hodson, A. (2006). Biogeochemistry of snowmelt in an Antarctic glacial ecosystem*. Water Resources Research*, 42.

Hodson, A. J., Mumford, P. N., Kohler, J., & Wynn, P. M. (2005). The High Arctic glacial ecosystem: new insights from nutrient budgets. *Biogeochemistry,* 72(2), 233-256.

Hodson, A., Anesio, A. M., Tranter, M., Fountain, A., Osborn, M., Priscu, J., ... & Sattler, B. (2008). Glacial ecosystems. *Ecological Monographs,* 78(1), 41-67.

Holmer, M., & Storkholm, P. (2001). Sulfate reduction and sulfur cycling in lake sediments: a review. *Freshwater Biology*, 46(4), 431-451.

Hood, E., & Scott, D. (2008). Riverine organic matter and nutrients in southeast Alaska affected by glacial coverage. *Nature Geoscience,* 1(9), 583-587.

Hood, E., Fellman, J., Spencer, R. G., Hernes, P. J., Edwards, R., D'Amore, D., & Scott, D. (2009). Glaciers as a source of ancient and labile organic matter to the marine environment. *Nature,* 462(7276), 1044-1047.

Hopwood, M,J., Statham, P,J., Tranter, M & Wadham, J,L. (2014) Glacial flours as a potential source of Fe(II) and Fe(III) to polar waters. *Biogeochemistry,* 443-452.

Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D. O., Liuni, S., Sammeth, M., ... & Pesole, G. (2009). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, bbp046.

Howard, J.B., and Rees, D.C. (1996) Structural basis of biological nitrogen fixation. *Chem Rev,* 96: 2965–2982.

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences,* 111(13), 4904-4909.

Howe, A., & Chain, P. S. (2015). Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Frontiers in Microbiology*, 6, 678.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, 9(9), 868-877.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., ... & Jensen, L. J. (2015). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research,* 44(D1), D286-D293.

Hugenholtz, P., & Tyson, G. W. (2008). Microbiology: metagenomics. Nature, 455(7212), 481.

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., & Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology,* 16(1), 279.

Hurek, T., and Reinhold-Hurek, B. (1995). Identification of grass-associated and toluene-degrading diazotrophs, Azoarcus spp., by analyses of partial 16S ribosomal DNA sequences. *Applied and Environmental Microbiology,* 61(6), 2257-2261.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research,* 17(3).

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., ... and Tappu, R. (2016). MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology,* 12(6), e1004957.

Huson, D. H., Halpern, A. L., Lai, Z., Myers, E. W., Reinert, K., & Sutton, G. G. (2001). Comparing assemblies using fragments and mate-pairs. *Algorithms in Bioinformatics,* 294-306.

Illumina (2014). Estimating sequencing coverage [Online]. Available from: https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf [Accessed 13-2-2018].

Imhoff, J. F. (2005). 'Chromatiales ord. nov'. In Holt. J., Sneath. P., *Bergey's Manual of systematic bacteriology*. Springer: Boston, 1-59.

Imhoff, J. F., and Pfennig, N. (2001). Thioflavicoccus mobilis gen. nov., sp. nov., a novel purple sulfur bacterium with bacteriochlorophyll b. *International Journal of Systematic and Evolutionary Microbiology*, 51(1), 105-110.

İnceoğlu, Ö., Hoogwout, E. F., Hill, P., & van Elsas, J. D. (2010). Effect of DNA extraction method on the apparent microbial diversity of soil. *Applied and Environmental Microbiology,* 76(10), 3378-3382.

IPCC, (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern- mental Panel on Climate Change*. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535.

Irgens, R. L., Gosink, J. J., and Staley, J. T. (1996). Polaromonas vacuolata gen. nov., sp. nov., a psychrophilic, marine, gas vacuolate bacterium from Antarctica. *International Journal of Systematic and Evolutionary Microbiology,* 46(3), 822-826.

Iriarte, J. L., González, H. E., & Nahuelhual, L. (2010). Patagonian fjord ecosystems in southern Chile as a highly vulnerable region: problems and needs. *AMBIO: A Journal of the Human Environment,* 39(7), 463-466.

Iriarte, J. L., González, H. E., Liu, K. K., Rivas, C., & Valenzuela, C. (2007). Spatial and temporal variability of chlorophyll and primary productivity in surface waters of southern Chile (41.5–43 S). *Estuarine, Coastal and Shelf Science,* 74(3), 471-480.

Iriarte, J. L., Pantoja, S., & Daneri, G. (2014). Oceanographic processes in Chilean Fjords of Patagonia: from small to large-scale studies. *Progress in Oceanography,* 129, 1-7.

Ishii, S. I., Suzuki, S., Tenney, A., Norden-Krichmar, T. M., Nealson, K. H., & Bretschger, O. (2015). Microbial metabolic networks in a complex electrogenic biofilm recovered from a stimulus-induced metatranscriptomics approach. *Scientific Reports,* 5.

Janssen, P. H., Yates, P. S., Grinton, B. E., Taylor, P. M., & Sait, M. (2002). Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Applied and Environmental Microbiology,* 68(5), 2391-2396.

Jetten, M. S. (2008). The microbial nitrogen cycle. *Environmental Microbiology*, 10(11), 2903-2909.

Joan, L., Slonczewski, & John W. F. (2011), *Microbiology: An Evolving Science (2nd Edition),* Norton: London.

Jumpponen, A. (2003). Soil fungal community assembly in a primary successional glacier forefront ecosystem as inferred from rDNA sequence analyses. *New Phytologist,* 158(3), 569-578.

Kandeler, E., Deiglmayr, K., Tscherko, D., Bru, D., & Philippot, L. (2006). Abundance of narG, nirS, nirK, and nosZ genes of denitrifying bacteria during primary successions of a glacier foreland. *Applied and Environmental Microbiology,* 72(9), 5957-5962.

Kaštovská, K., Elster, J., Stibal, M., & Šantrůčková, H. (2005). Microbial assemblages in soil microbial succession after glacial retreat in Svalbard (High Arctic). *Microbial Ecology*, 50(3), 396.

Keck, A., Wiktor, J., Hapter, R., & Nilsen, R. (2000). Phytoplankton assemblages related to physical gradients in an arctic, glacier-fed fjord in summer. *ICES Journal of Marine Science*, 56, 203-214.

Kirk, J. L., Beaudette, L. A., Hart, M., Moutoglis, P., Klironomos, J. N., Lee, H., & Trevors, J. T. (2004). Methods of studying soil microbial diversity. *Journal of Microbiological Methods*, 58(2), 169-188.

Kleindienst, S., Herbst, F. A., Stagars, M., Von Netzer, F., Von Bergen, M., Seifert, J., ... & Knittel, K. (2014). Diverse sulfate-reducing bacteria of the Desulfosarcina/Desulfococcus clade are the key alkane degraders at marine seeps. *The ISME journal,* 8(10), 2029.

Knelman, J. E., Legg, T. M., O'Neill, S. P., Washenberger, C. L., González, A., Cleveland, C. C., & Nemergut, D. R. (2012). Bacterial community structure and function change in association with colonizer plants during early primary succession in a glacier forefield. *Soil Biology and Biochemistry,* 46, 172-180.

Knight, P.G (1999). *Galciers.* Stanley Thornes: Cheltenham UK.

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences,* 102(7), 2567-2572.

Kuever, J. (2014). 'The family Desulfobacteraceae'. In: Dworkin, M., Falkow, S., Schleifer K.H., Stackebrandt, E. (Eds) *The Prokaryotes.* Springer: Berlin, Heidelberg, 45-73.

Kumar, S., & Blaxter, M. L. (2010). Comparing de novo assemblers for 454 transcriptome data. *BMC genomics*, 11(1), 571.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. (2008) A bioinformatician's guide to metagenomics. *Molecular Biology Reviews,* 72, 557–578.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), R12.

Lai, B., Ding, R., Li, Y., Duan, L., & Zhu, H. (2012). A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics,* 28(11), 1455-1462.

Landaeta, M. F., López, G., Suárez-Donoso, N., Bustos, C. A., & Balbontín, F. (2012). Larval fish distribution, growth and feeding in Patagonian fjords: potential effects of freshwater discharge. *Environmental Biology of Fishes*, 93(1), 73-87.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357.

Lawson, E. C., Wadham, J. L., Tranter, M., Stibal, M., Lis, G. P., Butler, C. E., ... & Dewsbury, P. (2014). Greenland Ice Sheet exports labile organic carbon to the Arctic oceans. *Biogeosciences,* 11(14), 4015-4028.

Lee, B., Han, Y., Huh, Y., Lundstrom, C., Siame, L. L., Lee, J. I., ... & ASTER Team. (2013). Chemical and physical weathering in south Patagonian rivers: A combined Sr–U–Be isotope approach. *Geochimica et Cosmochimica Acta,* 101, 173-190.

Lenaerts, J. T., Van Den Broeke, M. R., van Wessem, J. M., van de Berg, W. J., van Meijgaard, E., van Ulft, L. H., & Schaefer, M. (2014). Extreme precipitation and climate gradients in Patagonia revealed by high-resolution regional atmospheric climate modeling. *Journal of Climate,* 27(12), 4607-4621.

Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences*, 112(7), 2076-2081.

Li, H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics,* 25, 1754-60.

Li, Y., Hu, Y., Bolund, L., & Wang, J. (2010). State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human Genomics*, 4(4), 1.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., & Deng, H. W. (2011). Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics*, *27*(15), 2031-2037.

Liu, G. X., Hu, P., Zhang, W., Wu, X., Yang, X., Chen, T., ... and Li, S. W. (2012). Variations in soil culturable bacteria communities and biochemical characteristics in the Dongkemadi glacier forefield along a chronosequence*. Folia Microbiologica*, 57(6), 485-494.

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., and Linder, C. R. (2011). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, 61(1), 90-106.

Liu, Y., Yao, T., Jiao, N., Kang, S., Zeng, Y., & Huang, S. (2006). Microbial community structure in moraine lakes and glacial meltwaters, Mount Everest. *FEMS microbiology letters*, 265(1), 98-105.

Lovley, D. R., Giovannoni, S. J., White, D. C., Champine, J. E., Phillips, E. J. P., Gorby, Y. A., & Goodwin, S. (1993). Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. *Archives of Microbiology,* 159(4), 336-344.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology,* 13(5), e1005457.

Mackelprang, R., Waldrop, M. P., DeAngelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., ... & Jansson, J. K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature,* 480(7377), 368-371.

Maddison, W. P. and D.R. Maddison. (2017). Mesquite: a modular system for evolutionary analysis. Version 3.2 [Online]. Available at: http://mesquiteproject.org [Accessed 08-10-17].

Malard, L. A., & Pearce, D. A. (2018). Microbial diversity and biogeography in Arctic soils. *Environmental Microbiology Reports*, *10*(6), 611-625.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics,* 24(3), 133-141.

Margesin, R., Gander, S., Zacke, G., Gounot, A. M., & Schinner, F. (2003). Hydrocarbon degradation and enzyme activities of cold-adapted bacteria and yeasts. *Extremophiles,* 7(6), 451-458.

Marshall, W. A. & Chalmers, M. O. (1997) Airborne dispersal of Antarctic algae and cyanobacteria. *Ecography,* 20, 585–594

Marx, V. (2017). Microbiology: the return of culture. *Nature Methods,* 14(1), 37-41.

Matsunaga, Okamura, Fukuda, Wahyudi, Murase, and Takeyama (2005). Complete Genome Sequence of the Facultative Anaerobic Magnetotactic Bacterium Magnetospirillum sp. strain AMB-1. *DNA Research*, 12,157–166.

Matsuura, N., Tourlousse, D. M., Sun, L., Toyonaga, M., Kuroda, K., Ohashi, A., ... & Sekiguchi, Y. (2015). Draft genome sequence of Anaerolineae strain TC1, a novel isolate from a methanogenic wastewater treatment system*. Genome Announcements,* 3(5), e01104-15.

Matthews, J. A., & Vater, A. E. (2015). Pioneer zone geo-ecological change: Observations from a chronosequence on the Storbreen glacier foreland, Jotunheimen, southern Norway. *CATENA,* 135, 219-230.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., ... & Lapidus, A. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6), 495-500.

Mayer, L. M., Keil, R. G., Macko, S. A., Joye, S. B., Ruttenberg, K. C., & Aller, R. C. (1998). Importance of suspended participates in riverine delivery of bioavailable nitrogen to coastal zones. *Global Biogeochemical Cycles*, *12*(4), 573-579.

Meerhoff, E., Castro, L. & Tapia, F. (2013). Influence of freshwater discharges and tides on the abundance and distribution of larval and juvenile Munida gregaria in the Baker river estuary, Chilean Patagonia. *Continental Shelf Research*, 61–62(0), 1-11.

Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., ... & Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data*. PloS one*, 7(2), e31386.

Menge, D. N., & Hedin, L. O. (2009). Nitrogen fixation in different biogeochemical niches along a 120 000-year chronosequence in New Zealand. *Ecology,* 90(8), 2190-2201.

Menzel, P., Ng K., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications,* 7, 11257.

Meola, M., Lazzaro, A., & Zeyer, J. (2014). Diversity, resistance and resilience of the bacterial communities at two alpine glacier forefields after a reciprocal soil transplantation. *Environmental Microbiology,* 16(6), 1918-1934.

Methe, B. A., Nelson, K. E., Eisen, J. A., Paulsen, I. T., Nelson, W., Heidelberg, J. F., ... & Dodson, R. J. (2003). Genome of Geobacter sulfurreducens: metal reduction in subsurface environments. *Science,* 302(5652), 1967-1969.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics,* 11(1), 31-46.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., ... & Wilkening, J. (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics,* 9(1), 386.

Miller, J., Koren, S., Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics,* 95: 315–327

Mindl, B., Anesio, A. M., Meirer, K., Hodson, A. J., Laybourn-Parry, J., Sommaruga, R., & Sattler, B. (2007). Factors influencing bacterial dynamics along a transect from supraglacial runoff to proglacial lakes of a high Arctic glacieri. *FEMS Microbiology Ecology,* 59(2), 307-317.

Miniaci, C., Bunge, M., Duc, L., Edwards, I., Bürgmann, H., and Zeyer, J. (2007). Effects of pioneering plants on microbial structures and functions in a glacier forefield. *Biology and Fertility of Soils,* 44(2), 289-297.

Montero, P., Pérez-Santos, I., Daneri, G., Gutiérrez, M. H., Igor, G., Seguel, R., ... & Crawford, D. W. (2017). A winter dinoflagellate bloom drives high rates of primary production in a Patagonian fjord ecosystem. *Estuarine, Coastal and Shelf Science,* 199, 105-116.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L. K., ... & Smith, C. B. (2013). Sizing up metatranscriptomics. *The ISME journal,* 7(2), 237.

Moreau, J. W., Zierenberg, R. A., & Banfield, J. F. (2010). Diversity of dissimilatory sulfite reductase genes (dsrAB) in a salt marsh impacted by long-term acid mine drainage. *Applied and Environmental Microbiology,* 76(14), 4819-4828.

Moreau, M., Mercier, D., Laffly, D., & Roussel, E. (2008). Impacts of recent paraglacial dynamics on plant colonization: a case study on Midtre Lovénbreen foreland, Spitsbergen (79 N). *Geomorphology*, 95(1), 48-60.

Mueller, D. R., Vincent, W. F., Pollard, W. H., & Fritsen, C. H. (2001). Glacial cryoconite ecosystems: a bipolar comparison of algal communities and habitats. *Nova Hedwigia Beiheft,* 123, 173-198.

Müller, A. L., Kjeldsen, K. U., Rattei, T., Pester, M., & Loy, A. (2015). Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi) sulfite reductases. *The ISME journal,* 9(5), 1152.

Mullikin, J. C., & Ning, Z. (2003). The phusion assembler. *Genome Research,* 13(1), 81-90.

Murdock, S. A., & Juniper, S. K. (2017). Capturing compositional variation in denitrifying communities: A multiple primer approach that includes autotrophic denitrifying Epsilonproteobacteria. *Applied and Environmental Microbiology,* AEM-02753.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., ... & Anson, E. L. (2000). A whole-genome assembly of Drosophila. *Science*, *287*(5461), 2196-2204.

Mylona, P., Pawlowski, K., & Bisseling, T. (1995). Symbiotic nitrogen fixation. *The Plant Cell,* 7(7), 869.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics,* 14(3), 157-167.

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., ... & Allen, E. E. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME journal,* 6(1), 81.

Narzisi, G and Mishra, B. (2011) Comparing De Novo genome assembly: the long and short of it. *PLoS ONE*, 6.

Nash, M. V., Anesio, A. M., Barker, G., Tranter, M., Varliero, G., Eloe-Fadrosh, E. A., ... & Sánchez-Baracaldo, P. (2018). Metagenomic insights into diazotrophic communities across Arctic glacier forefields. *FEMS Microbiology Ecology,* 94(9), fiy114.

Nemergut, D. R., Anderson, S. P., Cleveland, C. C., Martin, A. P., Miller, A. E., Seimon, A., & Schmidt, S. K. (2007). Microbial community succession in an un-vegetated, recently de glaciated soil. *Microbial Ecology,* 53(1), 110-122.

Newton, C. R., Graham, A., & Ellison, J. S. (1997). *PCR*. BIOS Scientific Publishers: Chicago

Nicholson, W. L., Munakata, N., Horneck, G., Melosh, H. J., & Setlow, P. (2000). Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments. *Microbiology and Molecular Biology Reviews,* 64(3), 548-572.

Nicol, G. W., Tscherko, D., Embley, T. M., and Prosser, J. I. (2005). Primary succession of soil Crenarchaeota across a receding glacier foreland. *Environmental Microbiology,* 7(3), 337-347.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., ... & Pelletier, E. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology,* 32(8).

Nübel, U., Garcia-Pichel, F., & Muyzer, G. (1997). PCR primers to amplify 16S rRNA genes from cyanobacteria. *Applied and Environmental Microbiology,* 63(8), 3327-3332.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., ... & Stepanauskas, R. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology,* 20(10), 714-737.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research,* 27(5), 824-834.

Oda, Y., Samanta, S. K., Rey, F. E., Wu, L., Liu, X., Yan, T., ... and Harwood, C. S. (2005). Functional genomic analysis of three nitrogenase isozymes in the photosynthetic bacterium Rhodopseudomonas palustris. *Journal of Bacteriology,* 187(22), 7784-7794.

Ohtonen, R., Fritze, H., Pennanen, T., Jumpponen, A., & Trappe, J. (1999). Ecosystem properties and microbial community changes in primary succession on a glacier forefront. *Oecologia,* 119(2), 239-246.

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2015). 'Krona: Interactive Metagenomic Visualization in a Web Browser'. In Nelson, K.E. (Eds). *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools,* 339-346.

Oschmann, W. (2001). 'Oxygen in the ocean'. In D. Briggs., P. Crowther. (Eds.) *Palaeobiology II.* Wiley-Blackwell: Oxford, 470-472.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., ... & Vonstein, V. (2013). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research,* 42(D1), D206-D214.

Ozsolak, F. (2012). Third-generation sequencing techniques and applications to drug discovery. *Expert Opinion on Drug Discovery,* 7(3), 231-243.

Palma, S., & Silva, N. (2004). Distribution of siphonophores, chaetognaths, euphausiids and oceanographic conditions in the fjords and channels of southern Chile. *Deep Sea Research Part II: Topical Studies in Oceanography,* 51(6), 513-535.

Park, S. J., Kim, J. G., Jung, M. Y., Kim, S. J., Cha, I. T., Ghai, R., ... & Rhee, S. K. (2012). Draft genome sequence of an ammonia-oxidizing archaeon,"Candidatus Nitrosopumilus sediminis" AR2, from Svalbard in the Arctic Circle. *Journal of Bacteriology*, 194(24), 6948-6949.

Pattanaik, B., Schumann, R., and Karsten, U. (2007). Effects of ultraviolet radiation on cyanobacteria and their protective mechanisms. *Algae and Cyanobacteria in Extreme Environments,* 29-45.

Pearce, D. A., Newsham, K. K., Thorne, M. A., Calvo-Bado, L., Krsek, M., Laskaris, P., ... & Wellington, E. M. (2012). Metagenomic analysis of a southern maritime Antarctic soil. *Frontiers in Microbiology*, 3.

Peduzzi, S., Welsh, A., Demarta, A., Decristophoris, P., Peduzzi, R., Hahn, D., & Tonolla, M. (2011). Thiocystis chemoclinalis sp. nov. and Thiocystis cadagnonensis sp. nov., motile purple sulfur bacteria isolated from the chemocline of a meromictic lake. *International Journal of Systematic and Evolutionary Microbiology,* 61(7), 1682-1687.

Peng, Y., Leung, H. C., Yiu, S. M., & Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28(11), 1420-1428.

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17), 9748-9753.

Pickard, G. L. (1971). Some physical oceanographic features of inlets of Chile. *Journal of the Fisheries Board of Canada,* 28(8), 1077-1106.

Piganeau, G., & Moreau, H. (2007). Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta). *Gene,* 406(1).

Pignatelli, M., & Moya, A. (2011). Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One,* 6(5), e19984.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics,* 10(4), 354-366.

Prado-Fiedler, R. (2009). Winter and summer distribution of dissolved oxygen, pH and nutrients at the heads of fjords in Chilean Patagonia with possible phosphorus limitation. *Revista de Biología Marina y Oceanografía,* 44(3), 783-789.

Prakash, T., & Taylor, T. D. (2012). Functional assignment of metagenomic data: challenges and applications. *Briefings in Bioinformatics,* 13(6), 711-727

Prietzel, J., Wu, Y., Dümig, A., Zhou, J., & Klysubun, W. (2013). Soil sulfur speciation in two glacier forefield soil chronosequences assessed by SK-edge XANES spectroscopy. *European Journal of Soil Science*, 64(2), 260-272.

Qaigen Bioinformatics (2016) CLC Assembly Cell. [Online]. Available from: http://resources.qiagenbioinformatics.com/manuals/clcassemblycell/current/index.php?manual=Introduction.html [Accessed 25-01-2018].

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., ... & Mende, D. R. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature,* 464(7285), 59.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., ... & Peng, Y. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature,* 490(7418), 55-60.

Qin, W., Heal, K. R., Ramdasi, R., Kobelt, J. N., Martens-Habbena, W., Bertagnolli, A. D., ... & Devol, A. H. (2017). Nitrosopumilus maritimus gen. nov., sp. nov., Nitrosopumilus cobalaminigenes sp. nov., Nitrosopumilus oxyclinae sp. nov., and Nitrosopumilus ureiphilus sp. nov., four marine ammonia-oxidizing archaea of the phylum Thaumarchaeota. *International Journal of Systematic and Evolutionary Microbiology,* 67(12), 5067-5079.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1), 341.

Quaiser, A., Ochsenreiter, T., Klenk, H. P., Kletzin, A., Treusch, A. H., Meurer, G., ... & Schleper, C. (2002). First insight into the genome of an uncultivated crenarchaeote from soil. *Environmental Microbiology,* 4(10), 603-611.

Raiswell, R. (1984). Chemical models of solute acquisition in glacial meltwaters. *Journal of Glaciology,* 30(104), 49-57.

Raymond, J., Siefert, J. L., Staples, C. R., and Blankenship, R. E. (2004). The natural history of nitrogen fixation. *Molecular Biology and Evolution,* 21(3), 541-554.

Reddy, P. V. V., Rao, S. S. S. N., Pratibha, M. S., Sailaja, B., Kavya, B., Manorama, R. R., ... & Shivaji, S. (2009). Bacterial diversity and bioprospecting for cold-active enzymes from culturable bacteria associated with sediment from a melt water stream of Midtre Lovénbreen glacier, an Arctic glacier. *Research in Microbiology,* 160(8), 538-546.

Reed, D. C., Algar, C. K., Huber, J. A., & Dick, G. J. (2014). Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proceedings of the National Academy of Sciences*, 111(5), 1879-1884.

Renner, M., Arimitsu, M. L., & Piatt, J. F. (2012). Structure of marine predator and prey communities along environmental gradients in a glaciated fjord. *Canadian Journal of Fisheries and Aquatic Sciences*, 69(12), 2029-2045.

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annual Reviews Genetics,* 38, 525-552.

Rime, T., Hartmann, M., Brunner, I.(2015) Vertical distribution of the soil microbiota along a successional gradient in a glacier forefield. *Molecular Ecology,* 24, 1091–1108.

Rinnan, R., Rousk, J., Yergeau, E., Kowalchuk, G. A., & Bååth, E. (2009). Temperature adaptation of soil bacterial communities along an Antarctic climate gradient: predicting responses to climate warming. *Global Change Biology,* 15(11), 2615-2625.

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology,* 14(6), 405.

Rodriguez, L. M., & Konstantinidis, K. T. (2014). Estimating coverage in metagenomic data sets and why it matters. *The ISME journal,* 8(11), 2349.

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

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., ... & Tiong, C. L. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology,* 66(6), 2541-2547.

Rüger, H. J., Fritze, D., & Spröer, C. (2000). New psychrophilic and psychrotolerant Bacillus marinus strains from tropical and polar deep-sea sediments and emended description of the species. *International Journal of Systematic and Evolutionary Microbiology,* 50(3), 1305-1313.

Sakata Bekku, Y., Nakatsubo, T., Kume, A., & Koizumi, H. (2004). Soil microbial biomass, respiration rate, and temperature dependence on a successional glacier foreland in Ny-Ålesund, Svalbard. *Arctic, Antarctic, and Alpine Research,* 36(4), 395-399.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., ... & Marçais, G. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3), 557-567.

Sanford, R. A., Cole, J. R., and Tiedje, J. M. (2002). Characterization and description of Anaeromyxobacter dehalogenans gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium. *Applied and Environmental Microbiology,* 68(2), 893-900.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., ... & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687.

Sattin, S. R., Cleveland, C. C., Hood, E., Reed, S. C., King, A. J., Schmidt, S. K., ... & Nemergut, D. R. (2009). Functional shifts in unvegetated, perhumid, recently-deglaciated soils do not correlate with shifts in soil bacterial community composition. *The Journal of Microbiology,* 47(6), 673-681.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics,* 19(R2), R227-R240.

Schink, B. (2006). 'The genus Pelobacter'. In: Dworkin, M., Falkow, S., Schleifer K.H., Stackebrandt, E. (Eds). *The Prokaryotes.* Springer: New York, 5-11.

Schink, B., and Stieb, M. (1983). Fermentative degradation of polyethylene glycol by a strictly anaerobic, gram-negative, nonsporeforming bacterium, Pelobacter venetianus sp. nov. *Applied and Environmental Microbiology,* 45(6), 1905-1913.

Schmalenberger, A., & Noll, M. (2009). Shifts in desulfonating bacterial communities along a soil chronosequence in the forefield of a receding glacier. *FEMS Microbiology Ecology,* 71(2), 208-217.

Schmidt, S. K., Reed, S. C., Nemergut, D. R., Grandy, A. S., Cleveland, C. C., Weintraub, M. N., ... & Martin, A. M. (2008). The earliest stages of ecosystem succession in high-elevation (5000 metres above sea level), recently deglaciated soils. *Proceedings of the Royal Society of London B: Biological Sciences,* 275(1653), 2793-2802.

Schmieder, R., & Edwards, R. (2011). Quality control and pre-processing of metagenomic datasets. *Bioinformatics,* 27(6), 863-864

Schulz, S., Brankatschk, R., Dümig, A., Kögel-Knabner, I., Schloter, M., and Zeyer, J. (2013). The role of microorganisms at different stages of ecosystem development for soil formation. *Biogeosciences*, 10(6), 3983-3996.

Schütte, U. M., Abdo, Z., Bent, S. J., Williams, C. J., Schneider, G. M., Solheim, B., & Forney, L. J. (2009). Bacterial succession in a glacier foreland of the High Arctic. *The ISME journal,* 3(11), 1258-1268.

Schütte, U. M., Abdo, Z., Foster, J., Ravel, J., Bunge, J., Solheim, B., and Forney, L. J. (2010). Bacterial diversity in a glacier foreland of the high Arctic. *Molecular Ecology,* 19(s1), 54-66.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., ... & Bremges, A. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods,* 14(11), 1063.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8), 811-814.

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics,* 14(9), 618.

Sharon, I., & Banfield, J. F. (2013). Genomes from metagenomics. *Science,* 342(6162).

Shelobolina, E. S., Vrionis, H. A., Findlay, R. H., and Lovley, D. R. (2008). Geobacter uraniireducens sp. nov., isolated from subsurface sediment undergoing uranium bioremediation. *International Journal of Systematic and Evolutionary Microbiology*, 58(5), 1075-1078.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology,* 26(10), 1135-1145.

Sigler, W. V., & Zeyer, J. (2002). Microbial diversity and activity along the forefields of two receding glaciers. *Microbial Ecology,* 43(4), 397-407.

Silva, N. (2008). 'Physical and chemical characteristics of the surface sediments in the austral Chilean channels and fjords'. In Silva, N., Palma, S.(Eds.), *Progress in the oceanographic knowledge of Chilean interior waters, from Puerto Montt to Cape Horn. Comité Oceanográfico Nacional-Pontificia Universidad Católica de Valparaíso*, Valparaíso, 69-75.

Silva, N., & Vargas, C. A. (2014). Hypoxia in Chilean patagonian fjords. *Progress in Oceanography*, 129, 62-74.

Simon, C., Wiezer, A., Strittmatter, A. W., & Daniel, R. (2009). Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Applied and Environmental Microbiology,* 75(23), 7519-7526.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S. J., and Birol, I. (2006) AbySS: a parallel assembler for short read sequence data. *Genome Research,*19(6), 1117-1123.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics,* 15(2), 121-132.

Skidmore, M. L., Foght, J. M., & Sharp, M. J. (2000). Microbial life beneath a high Arctic glacier. *Applied and Environmental Microbiology,* 66(8), 3214-3220.

Sorokin, D. Y., Lysenko, A. M., Mityushina, L. L., Tourova, T. P., Jones, B. E., Rainey, F. A., ... & Kuenen, G. J. (2001). Thioalkalimicrobium aerophilum gen. nov., sp. nov. and Thioalkalimicrobium sibericum sp. nov., and Thioalkalivibrio versutus gen. nov., sp. nov., Thioalkalivibrio nitratis sp. nov., novel and Thioalkalivibrio denitrificancs sp. nov., novel obligately alkaliphilic and obligately chemolithoautotrophic sulfur-oxidizing bacteria from soda lakes. *International Journal of Systematic and Evolutionary Microbiology,* 51(2), 565-580.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30(9), 1312-1313.

Statham, P. J., Skidmore, M., & Tranter, M. (2008). Inputs of glacially derived dissolved and colloidal iron to the coastal ocean and implications for primary productivity. *Global Biogeochemical Cycles*, 22(3).

Stibal, M., Hasan, F., Wadham, J. L., Sharp, M. J., & Anesio, A. M. (2012). Prokaryotic diversity in sediments beneath two polar glaciers with contrasting organic carbon substrates. *Extremophiles,* 16(2), 255-265.

Stibal, M., Šabacká, M., & Kaštovská, K. (2006). Microbial communities on glacier surfaces in Svalbard: impact of physical and chemical properties on abundance and structure of cyanobacteria and algae. *Microbial Ecology,* 52(4), 644-654.

Stibal, M., Tranter, M., Benning, L. G., & Řehák, J. (2008). Microbial primary production on an Arctic glacier is insignificant in comparison with allochthonous organic carbon input. *Environmental Microbiology*, 10(8), 2172-2178.

Strauss, S. L., Garcia-Pichel, F., & Day, T. A. (2012). Soil microbial carbon and nitrogen transformations at a glacial foreland on Anvers Island, Antarctic Peninsula. *Polar Biology,* 35(10), 1459-1471.

Strauss, S. L., Ruhland, C. T., & Day, T. A. (2009). Trends in soil characteristics along a recently deglaciated foreland on Anvers Island, Antarctic Peninsula. *Polar Biology,* 32(12), 1779-1788.

Streit, W. R., & Schmitz, R. A. (2004). Metagenomics–the key to the uncultured microbes. *Current Opinion in Microbiology,* 7(5), 492-498.

Sutton, G. G., White, O., Adams, M. D., & Kerlavage, A. R. (1995). TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology,* 1(1), 9-19.

Suzuki, K. I., Sasaki, J., Uramoto, M., Nakase, T., & Komagata, K. (1997). Cryobacterium psychrophilum gen. nov., sp. nov., nom. rev., comb. nov., an obligately psychrophilic actinomycete to accommodate "Curtobacterium psychrophilum" Inoue and Komagata 1976. *International Journal of Systematic and Evolutionary Microbiology*, 47(2), 474-478.

Swerdlow, H., Wu, S.L., Harke, H. & Dovichi, N.J. (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography*, 516, 61–67.

Szynkiewicz, A., Modelska, M., Buczyński, S., Borrok, D. M., & Merrison, J. P. (2013). The polar sulfur cycle in the Werenskioldbreen, Spitsbergen: Possible implications for understanding the deposition of sulfate minerals in the North Polar Region of Mars. *Geochimica et Cosmochimica Acta,* 106, 326-343.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA*. Molecular Ecology,* 21(8), 1789-1793.

Takai, K., Campbell, B. J., Cary, S. C., Suzuki, M., Oida, H., Nunoura, T., ... & Horikoshi, K. (2005). Enzymatic and genetic characterization of carbon and energy metabolisms by deep-sea hydrothermal chemolithoautotrophic isolates of Epsilonproteobacteria. *Applied and Environmental Microbiology*, 71(11), 7310-7320.

Tamaru, Y., Takani, Y., Yoshida, T., & Sakamoto, T. (2005). Crucial role of extracellular polysaccharides in desiccation and freezing tolerance in the terrestrial cyanobacterium Nostoc commune. *Applied and Environmental Microbiology*, 71(11), 7327-7333.

Taylor, J., & Parkes, R. J. (1983). The cellular fatty acids of the sulfate-reducing bacteria, Desulfobacter sp., Desulfobulbus sp. and Desulfovibrio desulfuricans. *Microbiology,* 129(11), 3303-3309.

Teeling, H., & Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Briefings in Bioinformatics,* bbs039.

Teixeira, R. L. F., Von Der Weid, I., Seldin, L., and Rosado, A. S. (2008). Differential expression of nifH and anfH genes in Paenibacillus durus analysed by reverse transcriptase-PCR and denaturing gradient gel electrophoresis. *Letters in Applied Microbiology,* 46(3), 344-349.

Telling, J., Anesio, A. M., Hawkings, J., Tranter, M., Wadham, J. L., Hodson, A. J., ... & Yallop, M. L. (2010). Measuring rates of gross photosynthesis and net community production in cryoconite holes: a comparison of field methods. *Annals of Glaciology,* 51(56), 153-162.

Telling, J., Anesio, A. M., Tranter, M., Irvine-Fynn, T., Hodson, A., Butler, C., and Wadham, J. (2011). Nitrogen fixation on Arctic glaciers, Svalbard. Journal of Geophysical Research: *Biogeosciences,* 116(G3).

Telling, J., Stibal, M., Anesio, A. M., Tranter, M., Nias, I., Cook, J., ... & Nienow, P. (2012). Microbial nitrogen cycling on the Greenland Ice Sheet. *Biogeosciences*, 9(7), 2431-2442.

Teske, A., Alm, E., Regan, J. M., Toze, S., Rittmann, B. E., & Stahl, D. A. (1994). Evolutionary relationships among ammonia-and nitrite-oxidizing bacteria. *Journal of Bacteriology,* 176(21), 6623-6630.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics-a guide from sampling to data analysis. *Microbial Informatics and Experimentation,* 2(3), 1-12.

Torsvik, V., & Øvreås, L. (2002). Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology,* 5(3), 240-245.

Tranter, M., Brown, G. H., Hodson, A., Gurnell, A. M., & Sharp, M. J. (1994). Variations in the nitrate concentration of glacial runoff in alpine and sub-polar environments*. Snow and Ice covers: Interactions with the Atmosphere and Ecosystems*, 299-311.

Tranter, M., Sharp, M. J., Lamb, H. R., Brown, G. H., Hubbard, B. P., & Willis, I. C. (2002). Geochemical weathering at the bed of Haut Glacier d'Arolla, Switzerland—a new model. *Hydrological Processes,* 16(5), 959-993.

Tréguer, P.J. and Rocha C.L. (2013). The World Ocean Silica Cycle. *Annual Review of Marine Science,* 5(1), 477-501.

Treseder, K. K. (2008). Nitrogen additions and microbial biomass: A meta-analysis of ecosystem studies. *Ecology Letters,* 11(10), 1111-1120.

Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics,* 6(11), 805-814.

Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., ... & Bork, P. (2005). Comparative metagenomics of microbial communities. *Science,* 308(5721).

Tscherko, D., Hammesfahr, U., Zeltner, G., Kandeler, E., & Böcker, R. (2005). Plant succession and rhizosphere microbial communities in a recently deglaciated alpine terrain. *Basic and Applied Ecology,* 6(4), 367-383.

Tscherko, D., Rustemeier, J., Richter, A., Wanek, W., & Kandeler, E. (2003). Functional diversity of the soil microflora in primary succession across two glacier forelands in the Central Alps. *European Journal of Soil Science,* 54(4), 685-696.

Turpin-Jelfs, T., Michaelides, K., Blacker, J. J., Benning, L. G., Williams, J. M., & Anesio, A. M. (2019). Distribution of soil nitrogen and nitrogenase activity in the forefield of a High Arctic receding glacier. *Annals of Glaciology*, 1-8.

Van Der Heijden, M. G., Bardgett, R. D., & Van Straalen, N. M. (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology Letters,* 11(3), 296-310.

Van der Walt, A. J., Van Goethem, M. W., Ramond, J. B., Makhalanyane, T. P., Reva, O., & Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *bioRxiv,* 120154

Vázquez-Castellanos, J. F., García-López, R., Pérez-Brocal, V., Pignatelli, M., & Moya, A. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics,* 15(1), 37.

Vezzi, F., Cattonaro, F., & Policriti, A. (2011). e-RGA: enhanced reference guided assembly of complex genomes. *EMBnet. journal,* 17(1), pp-46.

Vincent, J., Wilke, A., Bischof, J., Desai, N., D'Souza, M., Glass, E., ... & Meyer, F. (2013). Analysis of Shotgun Metagenomes with MG-RAST. *Journal of Biomolecular Techniques*, 24, S20.

Vincent, W. F. , J. A. E. Gibson , R. Pienitz , and V. Villeneuve (2000). Ice shelf microbial ecosystems in the high Arctic and implications for life on Snowball Earth. *Naturwissenschaften* 87: 137–141.

Vishnivetskaya, T. A., Layton, A. C., Lau, M. C., Chauhan, A., Cheng, K. R., Meyers, A. J., ... & Pfiffner, S. M. (2014). Commercial DNA extraction kits impact observed microbial community composition in permafrost samples. *FEMS Microbiology Ecology*, 87(1), 217-230.

Vollmers, J., Wiegand, S., & Kaster, A. K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist’s perspective-Not only size matters!. *PloS one,* 12(1).

Vos, M., Quince, C., Pijl, A. S., de Hollander, M., & Kowalchuk, G. A. (2012). A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. *PLoS One,* 7(2), e30600.

Wadham, J. L., Bottrell, S., Tranter, M., & Raiswell, R. (2004). Stable isotope evidence for microbial sulfate reduction at the bed of a polythermal high Arctic glacier. *Earth and Planetary Science Letters*, *219*(3-4), 341-355.

Wadham, J. L., Tranter, M., Tulaczyk, S., & Sharp, M. (2008). Subglacial methanogenesis: a potential climatic amplifier?. *Global Biogeochemical Cycles*, *22*(2).

Wadham, J. L., Arndt, S., Tulaczyk, S., Stibal, M., Tranter, M., Telling, J., ... & Sharp, M. J. (2012). *Potential methane reservoirs beneath Antarctica. Nature*, 488(7413), 633-637.

Wadham, J. L., De'ath, R., Monteiro, F. M., Tranter, M., Ridgwell, A., Raiswell, R., & Tulaczyk, S. (2013). The potential role of the Antarctic Ice Sheet in global biogeochemical cycles. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh,* 104(01), 55-67.

Wadham, J. L., Tranter, M., Skidmore, M., Hodson, A. J., Priscu, J., Lyons, W. B., ... & Jackson, M. (2010). Biogeochemical weathering under ice: size matters. *Global Biogeochemical Cycles,* 24(3).

Wagner, M., Loy, A., Klein, M., Lee, N., Ramsing, N. B., Stahl, D. A., & Friedrich, M. W. (2005). Functional marker genes for identification of sulfate-reducing prokaryotes. *Methods in Enzymology,* 397, 469-489.

Wainwright, M. (1978). Microbial sulfur oxidation in soil. *Science Progress*, 459-475.

Wall, L. G. (2000). The actinorhizal symbiosis. *Journal of plant growth regulation*, 19(2), 167-182.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics,* 10(1), 57.

Ward, D. M., Weller, R., & Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature,* 345(6270), 63.

Ward, N. L., Challacombe, J. F., Janssen, P. H., Henrissat, B., Coutinho, P. M., Wu, M., ... & Barabote, R. D. (2009). Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. *Applied and Environmental Microbiology,* 75(7), 2046-2056.

Warren, R.L., Sutton, G.G., Jones, S.J.M., Holt, R.A. (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics,* (23)500.

Wash, S., & Image, C. (2008). DNA sequencing: generation next–next. *Nature Methods,* 5(3), 267.

Weiner, R., Langille ,S., and Quintero, E. (1995). Structure, function and immunochemistry of bacterial exopolysaccharides. *Joural of Industrial Microbiology,* 15, 339–346.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology,* 173(2), 697-703.

White, D. C., Sutton, S. D., & Ringelberg, D. B. (1996). The genus Sphingomonas: physiology and ecology. *Current Opinion in Biotechnology,* 7(3), 301-306.

Widdel, F and Hansen, T.A. (1992) 'The dissimilatory sulfate- and sulfur-reducing bacteria'. In: Balows A., Trüper H.G., Dworkin M. Harder W., Schleifer K.H. (Eds) *The Prokaryotes. A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications.* Springer: New York, 583-624.

Wojcik, R., Donhauser, J., Frey, B., Holm, S., Holland, A., Anesio, A. M., ... & Benning, L. G. (2019). Linkages between geochemistry and microbiology in a proglacial terrain in the High Arctic. *Annals of Glaciology,* 1-16.

Wolicka, D., Zdanowski, M. K., Żmuda-Baranowska, M. J., Poszytek, A., & Grzesiak, J. (2014). Sulfate reducing activity detected in soil samples from Antarctica, Ecology Glacier forefield, King George Island. *Polish Journal of Microbiology,* 63(4), 443-450.

Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: read length matters. *Applied and Environmental Microbiology, 74*(5), 1453-1463.

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2).

Wu, Y. W., Simmons, B. A., & Singer, S. W. (2015). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4), 605-607.

Xu, J. (2006). Invited review: microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Molecular Ecology,* 15(7), 1713-1731.

Yang, L., Hu, Y., Liu, Y., Zhang, J., Ulstrup, J., & Molin, S. (2011). Distinct roles of extracellular polymeric substances in Pseudomonas aeruginosa biofilm development. *Environmental Microbiology*, 13(7), 1705-1717.

Yde, J. C., Bárcena, T. G., & Finster, K. W. (2011). Subglacial and proglacial ecosystem responses to climate change. In *Climate Change-Geophysical Foundations and Ecological Effects*. IntechOpen.

Yilmaz, S., Allgaier, M., & Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods,* 7(12), 943.

Zehr, J. P., Jenkins, B. D., Short, S. M., and Steward, G. F. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environmental Microbiology,* 5(7), 539-554.

Zeikus, J. G. (1977). The biology of methanogenic bacteria. *Bacteriological Reviews,* 41(2), 514.

Zemp, M., Frey, H., Gärtner-Roer, I., Nussbaumer, S. U., Hoelzle, M., Paul, F., ... & Bajracharya, S. (2015). Historically unprecedented global glacier decline in the early 21st century. *Journal of Glaciology,* 61(228), 745-762.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18(5), 821-829.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., & Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one,* 6(3), e17915.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology,* 7(1-2), 203-214.

Zumsteg, A., Bååth, E., Stierli, B., Zeyer, J., and Frey, B. (2013). Bacterial and fungal community responses to reciprocal soil transfer along a temperature and soil moisture gradient in a glacier forefield. *Soil Biology and Biochemistry,* 61, 121-132.

# Appendix 1

***A1 Table 1:*** *Full list of assemblies carried out in the assembler evaluation. The assembler, metagenome (A-D) and parameter set are identified.*

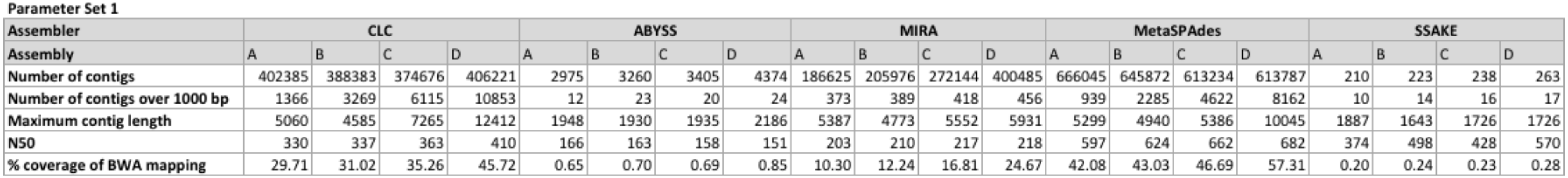
| Assembly number | Assembler | Metagenome | Parameter set |
|---|---|---|---|
| 1 | MIRA | A | Set 1 |
| 2 | MIRA | A | Set 2 |
| 3 | MIRA | A | Set 3 |
| 4 | MIRA | A | Set 4 |
| 5 | MIRA | A | Set 5 |
| 6 | MIRA | B | Set 1 |
| 7 | MIRA | B | Set 2 |
| 8 | MIRA | B | Set 3 |
| 9 | MIRA | B | Set 4 |
| 10 | MIRA | B | Set 5 |
| 11 | MIRA | C | Set 1 |
| 12 | MIRA | C | Set 2 |
| 13 | MIRA | C | Set 3 |
| 14 | MIRA | C | Set 4 |
| 15 | MIRA | C | Set 5 |
| 16 | MIRA | D | Set 1 |
| 17 | MIRA | D | Set 2 |
| 18 | MIRA | D | Set 3 |
| 19 | MIRA | D | Set 4 |
| 20 | MIRA | D | Set 5 |
| 21 | SSAKE | A | Set 1 |
| 22 | SSAKE | A | Set 2 |
| 23 | SSAKE | A | Set 3 |
| 24 | SSAKE | A | Set 4 |
| 25 | SSAKE | A | Set 5 |
| 26 | SSAKE | B | Set 1 |
| 27 | SSAKE | B | Set 2 |
| 28 | SSAKE | B | Set 3 |
| 29 | SSAKE | B | Set 4 |
| 30 | SSAKE | B | Set 5 |
| 31 | SSAKE | C | Set 1 |
| 32 | SSAKE | C | Set 2 |
| 33 | SSAKE | C | Set 3 |
| 34 | SSAKE | C | Set 4 |
| 35 | SSAKE | C | Set 5 |
| 36 | SSAKE | D | Set 1 |
| 37 | SSAKE | D | Set 2 |
| 38 | SSAKE | D | Set 3 |
| 39 | SSAKE | D | Set 4 |
| 40 | SSAKE | D | Set 5 |
| 41 | ABYSS | A | Set 1 |
| 42 | ABYSS | A | Set 2 |
| 43 | ABYSS | A | Set 3 |
| 44 | ABYSS | A | Set 4 |
| 45 | ABYSS | A | Set 5 |
| 46 | ABYSS | B | Set 1 |
| 47 | ABYSS | B | Set 2 |
| 48 | ABYSS | B | Set 3 |
| 49 | ABYSS | B | Set 4 |
| 50 | ABYSS | B | Set 5 |
| 51 | ABYSS | C | Set 1 |
| 52 | ABYSS | C | Set 2 |
| 53 | ABYSS | C | Set 3 |
| 54 | ABYSS | C | Set 4 |
| 55 | ABYSS | C | Set 5 |
| 56 | ABYSS | D | Set 1 |

| | | | |
|---|---|---|---|
| 57 | ABYSS | D | Set 2 |
| 58 | ABYSS | D | Set 3 |
| 59 | ABYSS | D | Set 4 |
| 60 | ABYSS | D | Set 5 |
| 61 | MetaSPAdes | A | Set 1 |
| 62 | MetaSPAdes | A | Set 2 |
| 63 | MetaSPAdes | A | Set 3 |
| 64 | MetaSPAdes | A | Set 4 |
| 65 | MetaSPAdes | A | Set 5 |
| 66 | MetaSPAdes | B | Set 1 |
| 67 | MetaSPAdes | B | Set 2 |
| 68 | MetaSPAdes | B | Set 3 |
| 69 | MetaSPAdes | B | Set 4 |
| 70 | MetaSPAdes | B | Set 5 |
| 71 | MetaSPAdes | C | Set 1 |
| 72 | MetaSPAdes | C | Set 2 |
| 73 | MetaSPAdes | C | Set 3 |
| 74 | MetaSPAdes | C | Set 4 |
| 75 | MetaSPAdes | C | Set 5 |
| 76 | MetaSPAdes | D | Set 1 |
| 77 | MetaSPAdes | D | Set 2 |
| 78 | MetaSPAdes | D | Set 3 |
| 79 | MetaSPAdes | D | Set 4 |
| 80 | MetaSPAdes | D | Set 5 |
| 81 | CLC | A | Set 1 |
| 82 | CLC | A | Set 2 |
| 83 | CLC | A | Set 3 |
| 84 | CLC | A | Set 4 |
| 85 | CLC | A | Set 5 |
| 86 | CLC | B | Set 1 |
| 87 | CLC | B | Set 2 |
| 88 | CLC | B | Set 3 |
| 89 | CLC | B | Set 4 |
| 90 | CLC | B | Set 5 |
| 91 | CLC | C | Set 1 |
| 92 | CLC | C | Set 2 |
| 93 | CLC | C | Set 3 |
| 94 | CLC | C | Set 4 |
| 95 | CLC | C | Set 5 |
| 96 | CLC | D | Set 1 |
| 97 | CLC | D | Set 2 |
| 98 | CLC | D | Set 3 |
| 99 | CLC | D | Set 4 |
| 100 | CLC | D | Set 5 |

***A1 Table 2:*** *Parameter settings modified during assemblies, with description of the parameter function. All other settings (unlisted) were kept at the default values, which can be found in the individual assembler manuals and publications (references listed).*

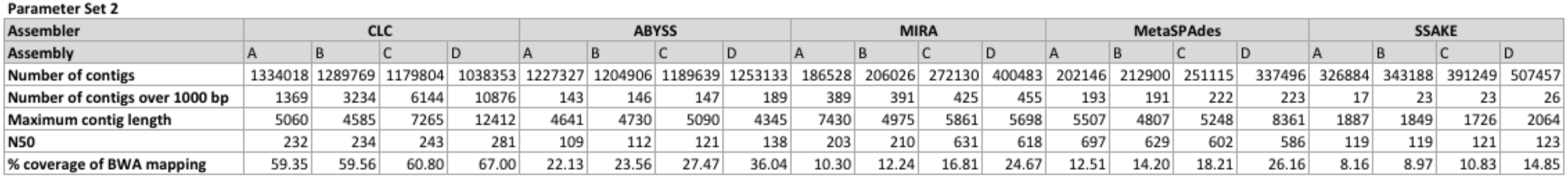
| **Assembler** | **Parameter details** | **Set 1** | **Set 2** | **Set 3** | **Set 4** | **Set 5** |
|---|---|---|---|---|---|---|
| MetaSPAdes 3.7.0<br>(Nurk *et al.,* 2013) | -k : kmer length | -k : 21,33,55 | -k : 41 | -k : 61 | -k : 71 | -k : 51 |
| SSAKE 3.8.4<br>(Warren *et al.,* 2007) | -w : minimum contig coverage depth | -w : 5 | -w : 1 | -w : 3 | -w : 1 | -w : 1 |
| | -m Minimum number of overlapping bases required during contig assembly | -m : default (20) | -m : default (20) | -m : default (20) | -m : 16 | -m : default (20) |
| | -t Number of contig bases to trim when other extension possibilities are depleted | -t : 0 | -t : 0 | -t : 0 | -t : 0 | -t : 1 |
| ABYSS 1.9.0<br>(Simpson *et al.,* 2009) | -k : kmer length | -k : 64 | -k : 40 | -k : 64 | -k : 70 | -k : 64 |
| | -c : remove contigs below N coverage threshold | -c : N/A | -c : N/A | -c : 2 | -c : N/A | -c : N/A |
| | -b : pop bubbles less than N base pairs | -b : N/A | -b : N/A | -b : N/A | -b : N/A | -b : 192 |
| MIRA 4.0<br>(Chavreux, B., 2014) | SOLEXA_SETTINGS : flag used to indicate illumine specific settings | SOLEXA_SETTINGS : off | SOLEXA_SETTINGS : on | SOLEXA_SETTINGS : on | SOLEXA_SETTINGS : on | SOLEXA_SETTINGS : on |
| | -AS:mrl : minimum read length | -AS:mrl= N/A | -AS:mrl= N/A | -AS:mrl= N/A | -AS:mrl= N/A | -AS:mrl= 50 |
| | -AS:ardml : minimum contig length | -AS:ardml = N/A | -AS:ardml = N/A | -AS:ardml = 100 | -AS:ardml = 150 | -AS:ardml = 200 |
| CLC 4.4.1<br>(Qaigen Bioinformatics, 2016) | -m : minimum output contig length | -m : default (200) | -m: 100 | -m : default (200) | -m : default (200) | -m : default (200) |
| | -w : wordsize for the de Bruijn graph (12 – 64). Default sets the value based on the input dataset size. | -w: default | -w: default | -w: 20 | -w: 60 | -w: default |
| | -b : maximum bubble size for the de Bruijn graph | -b: default (50) | -b: default (50) | -b: default (50) | -b: default (50) | -b: 40 |

***A1 Table 3:*** *Summary results table for metagenome assemblies, covering assembly size, contiguity and completeness. The assembler results are split by parameter sets (1-5), the assembler used, and the metagenome assembled (A-D). Assembly contiguity is identified by maximum contig length, contig N50 and the number of contigs over 1000bp. The completeness of the assembly is shown through the percentage coverage (of raw reads mapped to the input metagenome). Finally the number of contigs details the assembly size.*



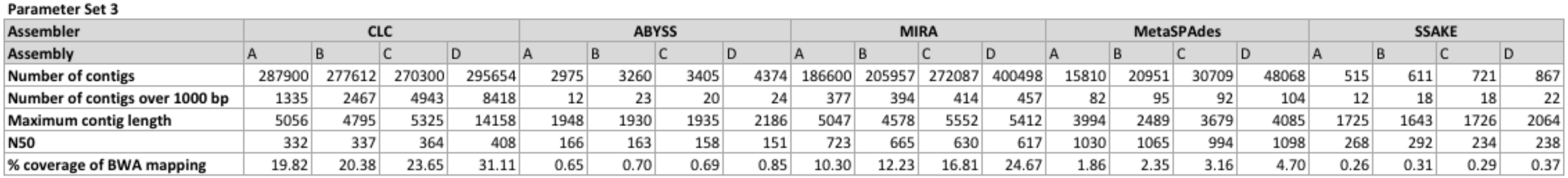
Parameter Set 1

| Assembler | CLC | | | | ABYSS | | | | MIRA | | | | MetaSPAdes | | | | SSAKE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assembly | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D |
| Number of contigs | 402385 | 388383 | 374676 | 406221 | 2975 | 3260 | 3405 | 4374 | 186625 | 205976 | 272144 | 400485 | 666045 | 645872 | 613234 | 613787 | 210 | 223 | 238 | 263 |
| Number of contigs over 1000 bp | 1366 | 3269 | 6115 | 10853 | 12 | 23 | 20 | 24 | 373 | 389 | 418 | 456 | 939 | 2285 | 4622 | 8162 | 10 | 14 | 16 | 17 |
| Maximum contig length | 5060 | 4585 | 7265 | 12412 | 1948 | 1930 | 1935 | 2186 | 5387 | 4773 | 5552 | 5931 | 5299 | 4940 | 5386 | 10045 | 1887 | 1643 | 1726 | 1726 |
| N50 | 330 | 337 | 363 | 410 | 166 | 163 | 158 | 151 | 203 | 210 | 217 | 218 | 597 | 624 | 662 | 682 | 374 | 498 | 428 | 570 |
| % coverage of BWA mapping | 29.71 | 31.02 | 35.26 | 45.72 | 0.65 | 0.70 | 0.69 | 0.85 | 10.30 | 12.24 | 16.81 | 24.67 | 42.08 | 43.03 | 46.69 | 57.31 | 0.20 | 0.24 | 0.23 | 0.28 |



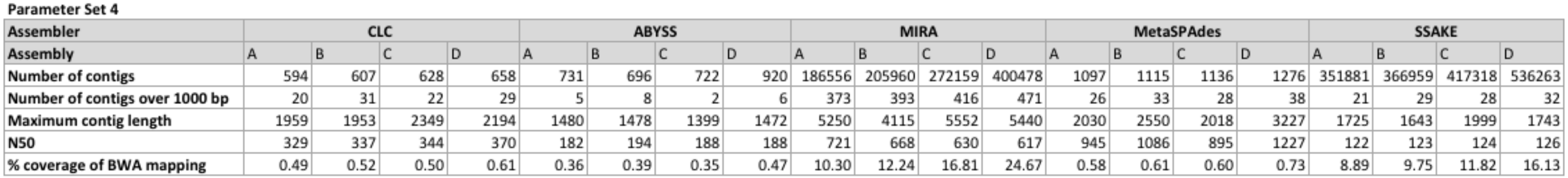
Parameter Set 2

| Assembler | CLC | | | | ABYSS | | | | MIRA | | | | MetaSPAdes | | | | SSAKE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assembly | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D |
| Number of contigs | 1334018 | 1289769 | 1179804 | 1038353 | 1227327 | 1204906 | 1189639 | 1253133 | 186528 | 206026 | 272130 | 400483 | 202146 | 212900 | 251115 | 337496 | 326884 | 343188 | 391249 | 507457 |
| Number of contigs over 1000 bp | 1369 | 3234 | 6144 | 10876 | 143 | 146 | 147 | 189 | 389 | 391 | 425 | 455 | 193 | 191 | 222 | 223 | 17 | 23 | 23 | 26 |
| Maximum contig length | 5060 | 4585 | 7265 | 12412 | 4641 | 4730 | 5090 | 4345 | 7430 | 4975 | 5861 | 5698 | 5507 | 4807 | 5248 | 8361 | 1887 | 1849 | 1726 | 2064 |
| N50 | 232 | 234 | 243 | 281 | 109 | 112 | 121 | 138 | 203 | 210 | 631 | 618 | 697 | 629 | 602 | 586 | 119 | 119 | 121 | 123 |
| % coverage of BWA mapping | 59.35 | 59.56 | 60.80 | 67.00 | 22.13 | 23.56 | 27.47 | 36.04 | 10.30 | 12.24 | 16.81 | 24.67 | 12.51 | 14.20 | 18.21 | 26.16 | 8.16 | 8.97 | 10.83 | 14.85 |



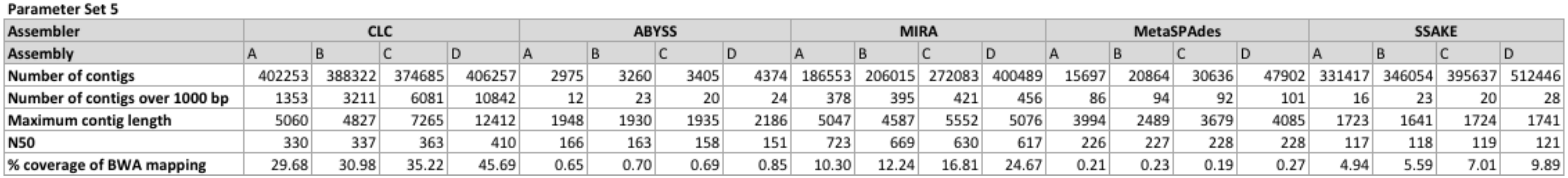
Parameter Set 3

| Assembler | CLC | | | | ABYSS | | | | MIRA | | | | MetaSPAdes | | | | SSAKE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assembly | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D |
| Number of contigs | 287900 | 277612 | 270300 | 295654 | 2975 | 3260 | 3405 | 4374 | 186600 | 205957 | 272087 | 400498 | 15810 | 20951 | 30709 | 48068 | 515 | 611 | 721 | 867 |
| Number of contigs over 1000 bp | 1335 | 2467 | 4943 | 8418 | 12 | 23 | 20 | 24 | 377 | 394 | 414 | 457 | 82 | 95 | 92 | 104 | 12 | 18 | 18 | 22 |
| Maximum contig length | 5056 | 4795 | 5325 | 14158 | 1948 | 1930 | 1935 | 2186 | 5047 | 4578 | 5552 | 5412 | 3994 | 2489 | 3679 | 4085 | 1725 | 1643 | 1726 | 2064 |
| N50 | 332 | 337 | 364 | 408 | 166 | 163 | 158 | 151 | 723 | 665 | 630 | 617 | 1030 | 1065 | 994 | 1098 | 268 | 292 | 234 | 238 |
| % coverage of BWA mapping | 19.82 | 20.38 | 23.65 | 31.11 | 0.65 | 0.70 | 0.69 | 0.85 | 10.30 | 12.23 | 16.81 | 24.67 | 1.86 | 2.35 | 3.16 | 4.70 | 0.26 | 0.31 | 0.29 | 0.37 |



Parameter Set 4

| Assembler | CLC | | | | ABYSS | | | | MIRA | | | | MetaSPAdes | | | | SSAKE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assembly | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D |
| Number of contigs | 594 | 607 | 628 | 658 | 731 | 696 | 722 | 920 | 186556 | 205960 | 272159 | 400478 | 1097 | 1115 | 1136 | 1276 | 351881 | 366959 | 417318 | 536263 |
| Number of contigs over 1000 bp | 20 | 31 | 22 | 29 | 5 | 8 | 2 | 6 | 373 | 393 | 416 | 471 | 26 | 33 | 28 | 38 | 21 | 29 | 28 | 32 |
| Maximum contig length | 1959 | 1953 | 2349 | 2194 | 1480 | 1478 | 1399 | 1472 | 5250 | 4115 | 5552 | 5440 | 2030 | 2550 | 2018 | 3227 | 1725 | 1643 | 1999 | 1743 |
| N50 | 329 | 337 | 344 | 370 | 182 | 194 | 188 | 188 | 721 | 668 | 630 | 617 | 945 | 1086 | 895 | 1227 | 122 | 123 | 124 | 126 |
| % coverage of BWA mapping | 0.49 | 0.52 | 0.50 | 0.61 | 0.36 | 0.39 | 0.35 | 0.47 | 10.30 | 12.24 | 16.81 | 24.67 | 0.58 | 0.61 | 0.60 | 0.73 | 8.89 | 9.75 | 11.82 | 16.13 |



Parameter Set 5

| Assembler | CLC | | | | ABYSS | | | | MIRA | | | | MetaSPAdes | | | | SSAKE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Assembly | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D | A | B | C | D |
| Number of contigs | 402253 | 388322 | 374685 | 406257 | 2975 | 3260 | 3405 | 4374 | 186553 | 206015 | 272083 | 400489 | 15697 | 20864 | 30636 | 47902 | 331417 | 346054 | 395637 | 512446 |
| Number of contigs over 1000 bp | 1353 | 3211 | 6081 | 10842 | 12 | 23 | 20 | 24 | 378 | 395 | 421 | 456 | 86 | 94 | 92 | 101 | 16 | 23 | 20 | 28 |
| Maximum contig length | 5060 | 4827 | 7265 | 12412 | 1948 | 1930 | 1935 | 2186 | 5047 | 4587 | 5552 | 5076 | 3994 | 2489 | 3679 | 4085 | 1723 | 1641 | 1724 | 1741 |
| N50 | 330 | 337 | 363 | 410 | 166 | 163 | 158 | 151 | 723 | 669 | 630 | 617 | 226 | 227 | 228 | 228 | 117 | 118 | 119 | 121 |
| % coverage of BWA mapping | 29.68 | 30.98 | 35.22 | 45.69 | 0.65 | 0.70 | 0.69 | 0.85 | 10.30 | 12.24 | 16.81 | 24.67 | 0.21 | 0.23 | 0.19 | 0.27 | 4.94 | 5.59 | 7.01 | 9.89 |

***A1 Table 4:*** *One way analysis of variance (ANOVA) results for best performing parameter sets. F value and P value significance for dependent variables, percentage coverage, number of contigs and maximum contig length are shown. For each dependent variable, the ANOVA was carried out between the five tested assemblers for the best performing parameter sets.*

| | % Coverage | Number of contigs | Maximum contig length |
|---|---|---|---|
| **F value** | 59.32 | 92.67 | 4.72 |
| **P value** | <0.0001 | <0.0001 | 0.112 |

***A1 Table 5:*** *One way analysis of variance (ANOVA) results for each assembler, comparing the contiguity, completeness and assembly size across the four metagenomes tested. For each metagenome, the summary statistics used included: number of contigs; contigs over 1000bp; maximum contig length and % coverage. These were compared across the assembly results for metagenomes A-D, to identify significant differences stemming from metagenome complexity, rather than the assembler.*

| | CLC | ABYSS | MIRA | MetaSPAdes | SSAKE |
|---|---|---|---|---|---|
| **F value** | 0.86 | 1.10 | 1.00 | 0.66 | 1.00 |
| **P value** | 0.49 | 0.41 | 0.42 | 0.59 | 0.43 |

# Appendix 2

***A2 Table 1:*** *Samples used in metagenomic sequencing across the four Arctic forefields, and the corresponding metadata. Samples were obtained in a transect across each forefield, moving away from the glacier terminus. For the Midtre Lovénbreen (Ml), Russell (Rl) and Storglaciären (St) sites, three parallel transects were conducted to obtain field replicates. This was not possible for the Rabots (Rb) site.*

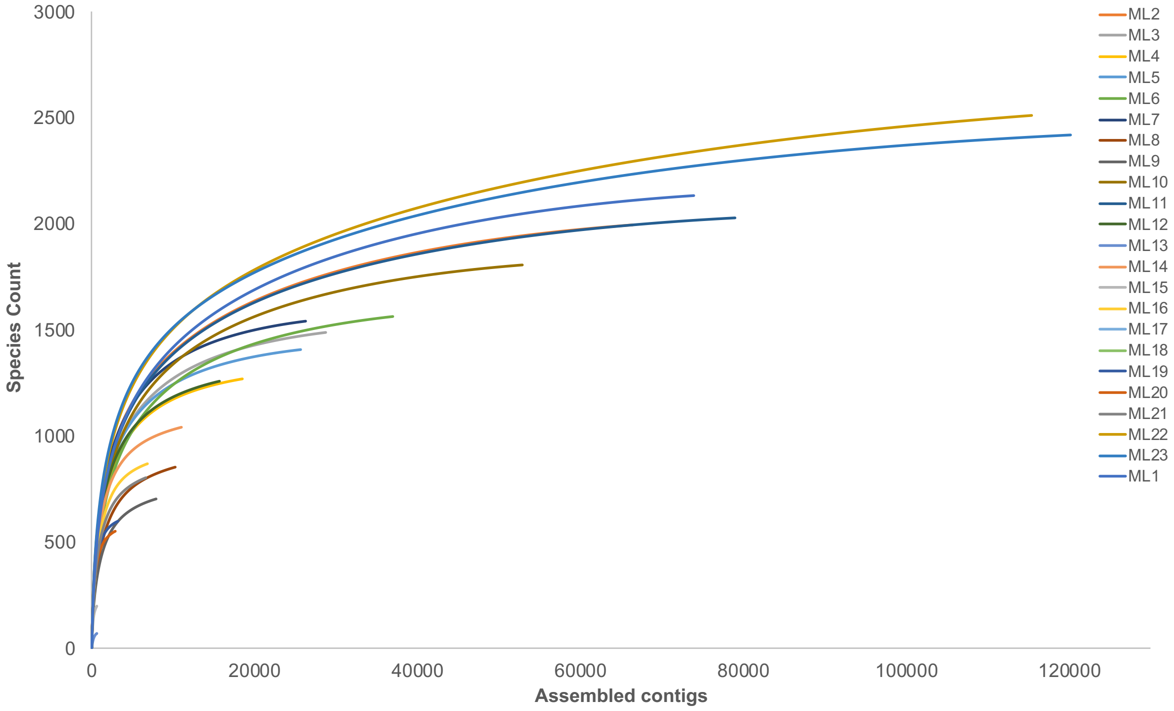
| Site ID | Forefield | Latitude | Longitude | Altitude (m) | Date obtained |
|---|---|---|---|---|---|
| Ml 1 | Ml | 79.100555 | 12.156111 | 54 | 29/07/2013 |
| Ml 2 | Ml | 79.112223 | 12.175555 | 44 | 29/07/2013 |
| Ml 3 | Ml | 79.112222 | 12.258333 | 44 | 29/07/2013 |
| Ml 4 | Ml | 79.118333 | 12.093611 | 54 | 29/07/2013 |
| Ml 5 | Ml | 79.113611 | 12.195833 | 52 | 29/07/2013 |
| Ml 6 | Ml | 79.104444 | 12.278888 | 52 | 29/07/2013 |
| Ml 7 | Ml | 79.152555 | 12.215555 | 50 | 29/07/2013 |
| Ml 8 | Ml | 79.151388 | 12.253611 | 43 | 29/07/2013 |
| Ml 9 | Ml | 79.140833 | 12.092222 | 43 | 29/07/2013 |
| Ml 10 | Ml | 78.927777 | 12.254166 | 35 | 29/07/2013 |
| Ml 11 | Ml | 78.921666 | 12.076666 | 40 | 29/07/2013 |
| Ml 12 | Ml | 78.907777 | 12.164444 | 48 | 29/07/2013 |
| Ml 13 | Ml | 78.900555 | 12.076111 | 30 | 29/07/2013 |
| Ml 14 | Ml | 78.900555 | 12.076111 | 40 | 29/07/2013 |
| Ml 15 | Ml | 78.900555 | 12.076111 | 105 | 29/07/2013 |
| Ml 16 | Ml | 78.900555 | 12.082777 | 29 | 29/07/2013 |
| Ml 17 | Ml | 78.991666 | 12.233333 | 30 | 29/07/2013 |
| Ml 18 | Ml | 78.978888 | 12.332222 | 30 | 29/07/2013 |
| Ml 19 | Ml | 79.768333 | 12.143611 | 19 | 29/07/2013 |
| Ml 20 | Ml | 79.768333 | 12.143611 | 19 | 29/07/2013 |
| Ml 21 | Ml | 79.768333 | 12.143611 | 19 | 29/07/2013 |
| Ml 22 | Ml | 79.484166 | 12.092222 | 105 | 29/07/2013 |
| Ml 23 | Ml | 79.484166 | 12.092222 | 105 | 29/07/2013 |
| | | | | | |
| Rl 1 | Rl | 67.15650902 | -50.06398397 | 439 | 24/07/2014 |
| Rl 2 | Rl | 67.15651598 | -50.06386997 | 440 | 24/07/2014 |
| Rl 3 | Rl | 67.15655998 | -50.06389101 | 439 | 24/07/2014 |
| Rl 4 | Rl | 67.16295303 | -50.01826898 | 589 | 25/07/2014 |
| Rl 5 | Rl | 67.16301103 | -50.01844500 | 589 | 25/07/2014 |
| Rl 6 | Rl | 67.16306903 | -50.01828399 | 589 | 25/07/2014 |
| Rl 7 | Rl | 67.15211598 | -50.04869697 | 515 | 26/07/2014 |
| Rl 8 | Rl | 67.15208103 | -50.04859303 | 516 | 26/07/2014 |
| Rl 9 | Rl | 67.15210701 | -50.04851701 | 516 | 26/07/2014 |
| Rl 10 | Rl | 67.15685402 | -50.08261903 | 404 | 26/07/2014 |
| Rl 11 | Rl | 67.15680499 | -50.08249900 | 403 | 26/07/2014 |
| Rl 12 | Rl | 67.15684304 | -50.08236104 | 404 | 26/07/2014 |
| Rl 13 | Rl | 67.15642001 | -50.08365101 | 403 | 26/07/2014 |

| | | | | | |
|---|---|---|---|---|---|
| RI 14 | RI | 67.15646099 | -50.08366501 | 403 | 26/07/2014 |
| RI 15 | RI | 67.15638698 | -50.08376702 | 403 | 26/07/2014 |
| RI 16 | RI | 67.15559204 | -50.08486102 | 411 | 26/07/2014 |
| RI 17 | RI | 67.15558601 | -50.08499002 | 411 | 26/07/2014 |
| RI 18 | RI | 67.15567200 | -50.08488097 | 411 | 26/07/2014 |
| RI 19 | RI | 67.08225802 | -50.32251497 | 251 | 27/07/2014 |
| RI 20 | RI | 67.08222499 | -50.32243500 | 238 | 27/07/2014 |
| RI 21 | RI | 67.08218602 | -50.32229704 | 237 | 27/07/2014 |
| RI 22 | RI | 67.05702002 | -50.45979604 | 147 | 27/07/2014 |
| RI 23 | RI | 67.05700200 | -50.45969403 | 148 | 27/07/2014 |
| RI 24 | RI | 67.05692296 | -50.45960804 | 148 | 27/07/2014 |
| | | | | | |
| Rb 1 | Rb | 67.910855 | 18.470863 | 1250 | 07/02/2014 |
| Rb 2 | Rb | 67.907119 | 18.447522 | 1105 | 07/02/2014 |
| Rb 3 | Rb | 67.907119 | 18.447522 | 1105 | 07/02/2014 |
| Rb 4 | Rb | 67.906846 | 18.445550 | 1110 | 07/02/2014 |
| Rb 5 | Rb | 67.872223 | 16.713705 | 1054 | 07/02/2014 |
| | | | | | |
| St 1 | St | 67.904568 | 18.607115 | 1131 | 07/01/2014 |
| St 2 | St | 67.904687 | 18.610965 | 1103 | 07/01/2014 |
| St 3 | St | 67.904687 | 18.610965 | 1103 | 07/01/2014 |
| St 4 | St | 67.904687 | 18.610965 | 1103 | 07/02/2014 |
| St 5 | St | 67.899243 | 18.344347 | 1147 | 07/01/2014 |
| St 6 | St | 67.899244 | 18.344371 | 1147 | 07/01/2014 |
| St 7 | St | 67.900853 | 18.441750 | 1146 | 07/01/2014 |
| St 8 | St | 67.900853 | 18.441750 | 1146 | 07/01/2014 |
| St 9 | St | 67.900853 | 18.441750 | 1146 | 07/01/2014 |
| St 10 | St | 67.900879 | 18.434740 | 1130 | 07/01/2014 |
| St 11 | St | 67.900879 | 18.434740 | 1130 | 07/01/2014 |
| St 12 | St | 67.900879 | 18.434740 | 1130 | 07/01/2014 |
| St 13 | St | 67.901082 | 18.428257 | 1113 | 07/01/2014 |
| St 14 | St | 67.901082 | 18.428257 | 1113 | 07/01/2014 |
| St 15 | St | 67.865505 | 16.714941 | 1103 | 07/01/2014 |
| St 16 | St | 67.865505 | 16.714941 | 1103 | 07/01/2014 |
| St 17 | St | 67.865505 | 16.714941 | 1103 | 07/01/2014 |
| St 18 | St | 67.903128 | 18.604355 | 1182 | 07/01/2014 |

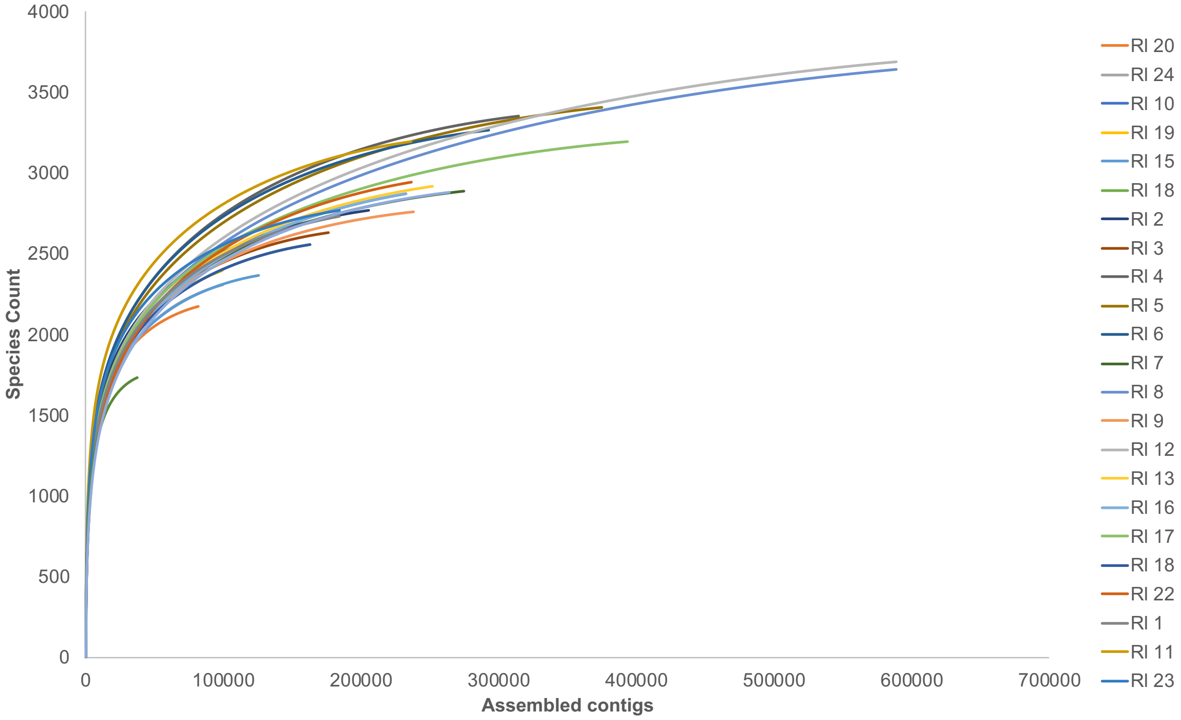
***A2 Table 2:*** *Output statistics for metagenome sequencing and assembly, for each site. The number of raw reads returned from sequencing is given, alongside the subsequent assembly sizes, in both sequences and bases.*

| | **Sequencing reads** | **Assembly size (sequences)** | **Assembly size (bases)** |
|---|---|---|---|
| MI 1 | 17465080 | 73894 | 33376241 |
| MI 2 | 28186722 | 66450 | 32695412 |
| MI 3 | 16801828 | 28727 | 17783002 |
| MI 4 | 25805412 | 18484 | 6405001 |
| MI 5 | 20630402 | 25643 | 10899365 |
| MI 6 | 21421268 | 36952 | 16375348 |
| MI 7 | 22655970 | 26238 | 13376228 |
| MI 8 | 20003790 | 10239 | 3947663 |
| MI 9 | 22650764 | 7883 | 3544627 |
| MI 10 | 58387594 | 52836 | 25394805 |
| MI 11 | 56518916 | 78995 | 29749304 |
| MI 12 | 37925220 | 15664 | 4751816 |
| MI 13 | 3952496 | 612 | 272071 |
| MI 14 | 22179120 | 10999 | 8951989 |
| MI 15 | 6433798 | 608 | 241660 |
| MI 16 | 22231192 | 6823 | 2827176 |
| MI 17 | 38684672 | 18250 | 16687388 |
| MI 18 | 68392102 | 78312 | 52131008 |
| MI 19 | 23405212 | 3192 | 998864 |
| MI 20 | 20152212 | 2871 | 859271 |
| MI 21 | 28777734 | 6620 | 2019562 |
| MI 22 | 22182332 | 115393 | 61826702 |
| MI 23 | 28234260 | 120239 | 81857371 |
| | | | |
| RI 1 | 67748804 | 204281 | 181720264 |
| RI 2 | 77474456 | 205426 | 192219830 |
| RI 3 | 61199512 | 176255 | 115011779 |
| RI 4 | 74992750 | 314429 | 325818814 |
| RI 5 | 74220220 | 374929 | 341656927 |
| RI 6 | 68161082 | 292741 | 302935645 |
| RI 7 | 75975046 | 274719 | 186606224 |
| RI 8 | 91844214 | 588635 | 395690349 |
| RI 9 | 77023868 | 238198 | 157627061 |
| RI 10 | 139381778 | 344195 | 180851046 |
| RI 11 | 91962344 | 203511 | 112190691 |
| RI 12 | 81681122 | 86956 | 82284058 |
| RI 13 | 105431412 | 251779 | 123396624 |
| RI 14 | 63240678 | 37335 | 22184640 |

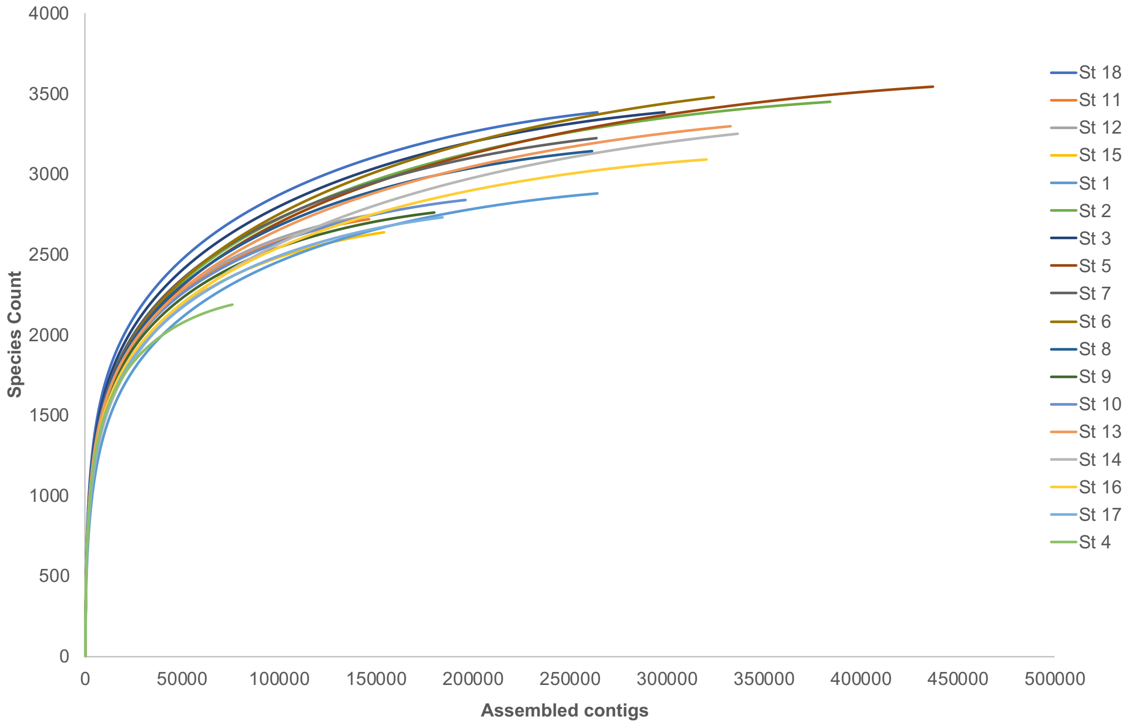
|  |  |  |  |
|---|---|---|---|
| Rl 15 | 72243790 | 125592 | 88672076 |
| Rl 16 | 82137112 | 232607 | 118301737 |
| Rl 17 | 107570706 | 393735 | 228570150 |
| Rl 18 | 78274268 | 162862 | 98312835 |
| Rl 19 | 75392190 | 99149 | 52942968 |
| Rl 20 | 65582044 | 81726 | 58507862 |
| Rl 21 | 91522052 | 264245 | 149668891 |
| Rl 22 | 96384726 | 236649 | 128282042 |
| Rl 23 | 96207238 | 406505 | 410018759 |
| Rl 24 | 80981820 | 99519 | 60773858 |
|  |  |  |  |
| Rb 1 | 76498938 | 350951 | 429543524 |
| Rb 2 | 72691820 | 275858 | 181694382 |
| Rb 3 | 70126632 | 334545 | 249845086 |
| Rb 4 | 61828928 | 303591 | 231705319 |
| Rb 5 | 74667258 | 375682 | 213265070 |
|  |  |  |  |
| St 1 | 68111048 | 336424 | 219250082 |
| St 2 | 71828498 | 154126 | 83781202 |
| St 3 | 85214054 | 320463 | 188269776 |
| St 4 | 71411294 | 184331 | 103140952 |
| St 5 | 66910678 | 323872 | 236040965 |
| St 6 | 72683122 | 384288 | 290006175 |
| St 7 | 60148730 | 298633 | 216231874 |
| St 8 | 67844804 | 75868 | 48601571 |
| St 9 | 63953088 | 437238 | 275297528 |
| St 10 | 83010234 | 324248 | 214964286 |
| St 11 | 74901072 | 263784 | 180223385 |
| St 12 | 83405572 | 261266 | 175258662 |
| St 13 | 64225756 | 179953 | 121520140 |
| St 14 | 64764076 | 196033 | 140418783 |
| St 15 | 66675200 | 146437 | 83396680 |
| St 16 | 66848090 | 149134 | 96714442 |
| St 17 | 72051286 | 332892 | 268415228 |
| St 18 | 78397352 | 478826 | 377436650 |



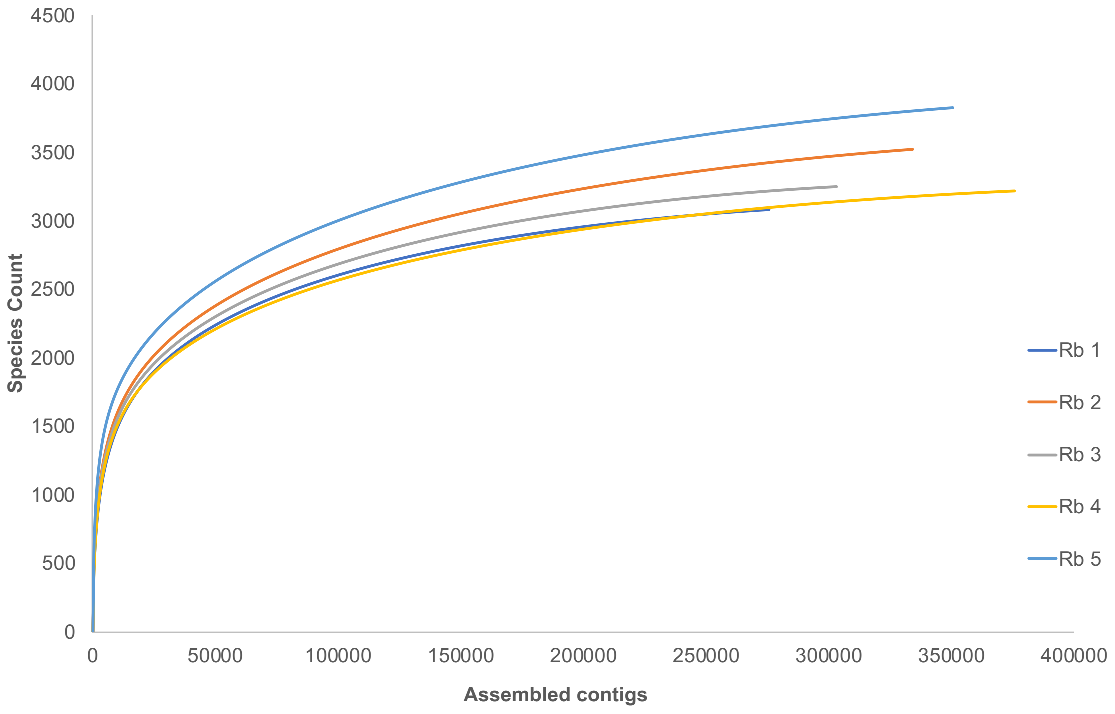
***A2 Figure 1:** Rarefaction curves for metagenomes sampled from Midtre Lovénbreen, Svalbard. The total assembled contigs in each metagenome is shown, against the total species count obtained from these contigs.*



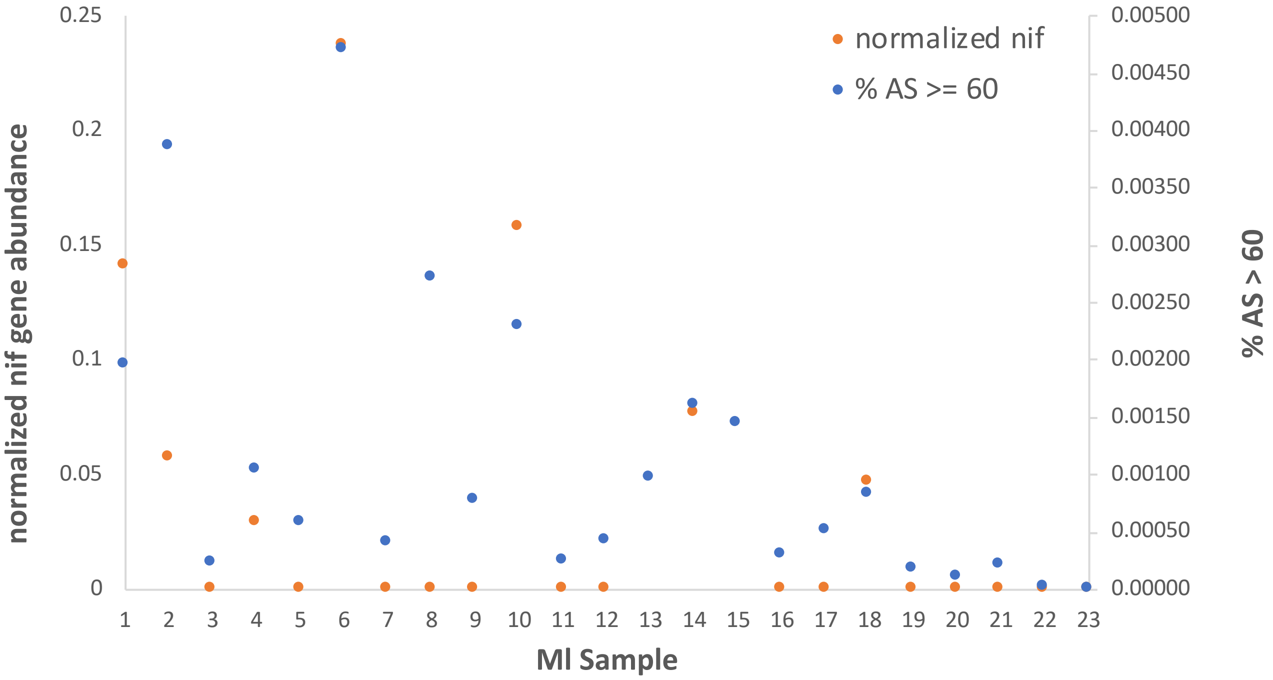
***A2 Figure 2:** Rarefaction curves for metagenomes sampled from Russell Glacier, Greenland. The total assembled contigs in each metagenome is shown, against the total species count obtained from these contigs.*



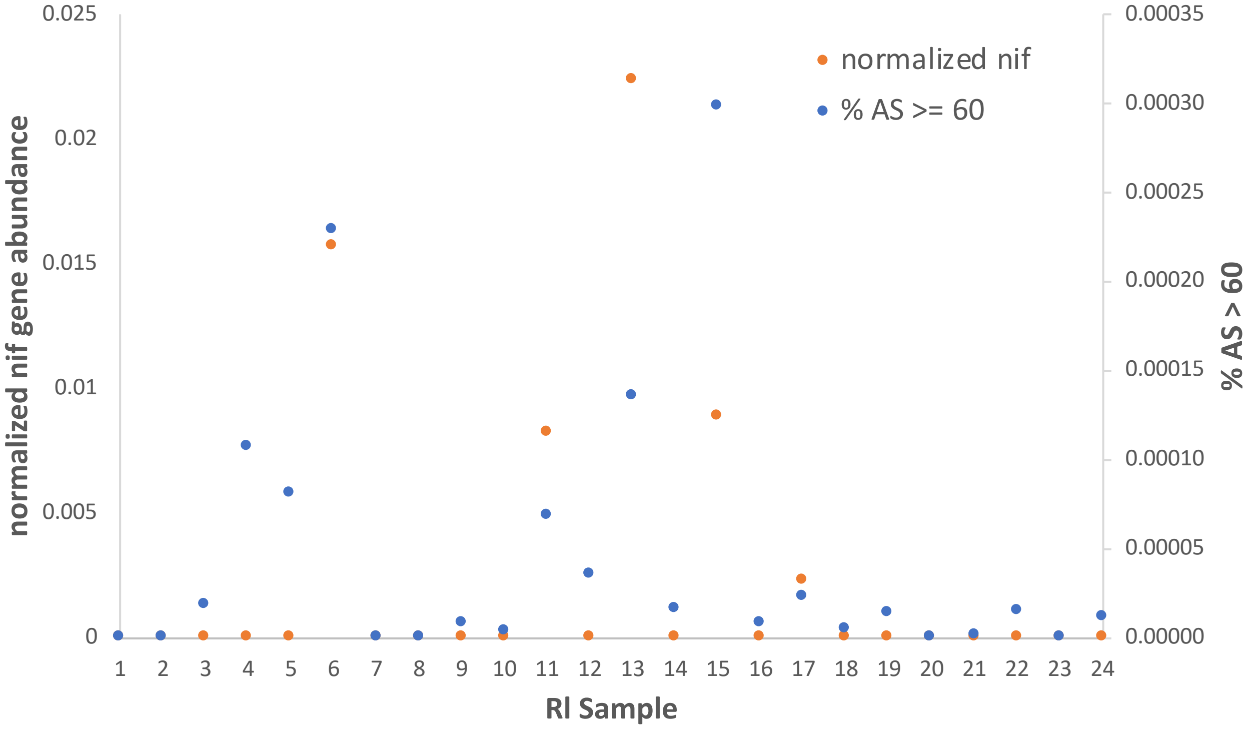
***A2 Figure 3:** Rarefaction curves for metagenomes sampled from Storglaciären, N-Sweden. The total assembled contigs in each metagenome is shown, against the total species count obtained from these contigs.*



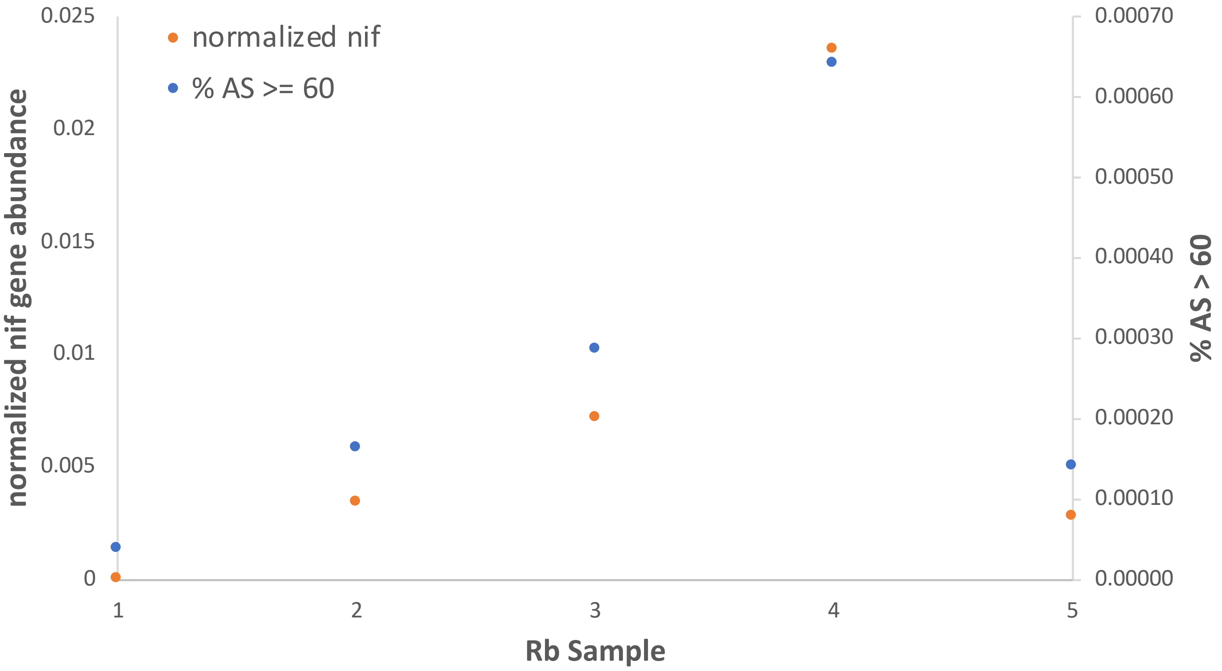
***A2 Figure 4:** Rarefaction curves for metagenomes sampled from Rabots glacier, N-Sweden. The total assembled contigs in each metagenome is shown, against the total species count obtained.*



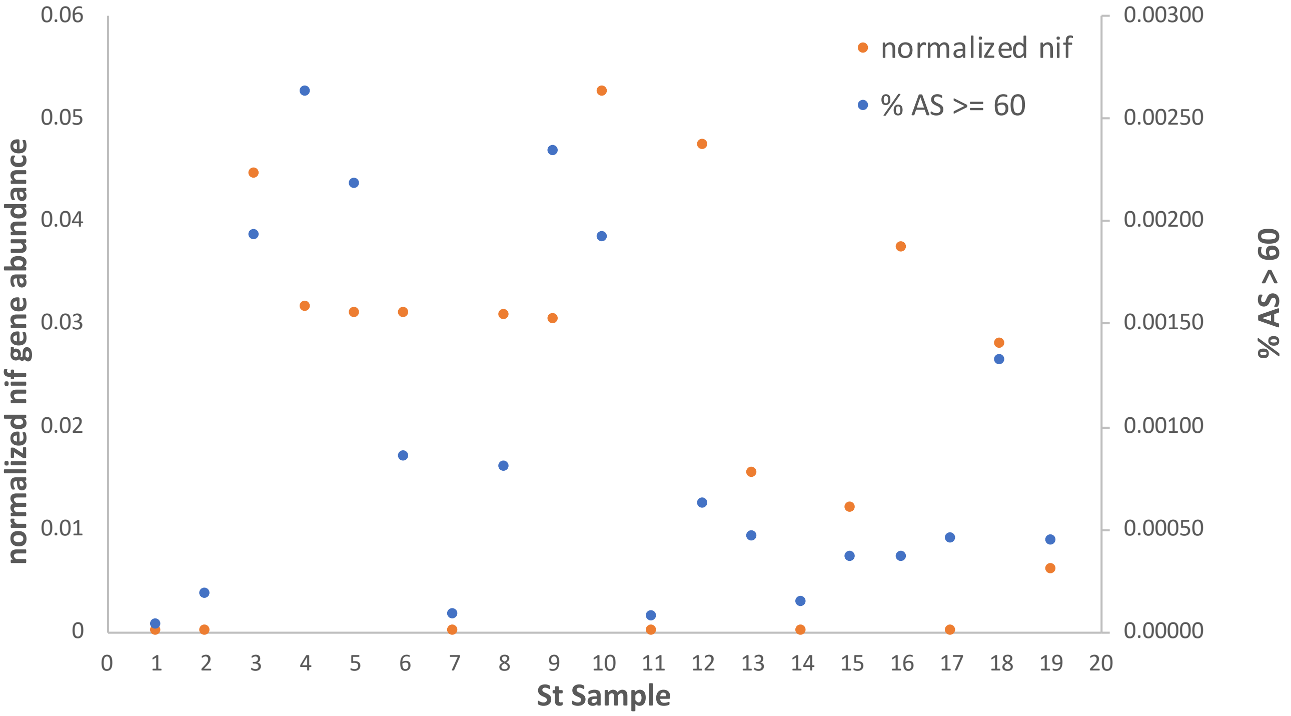
***A2 Figure 5:*** *rpoB normalized nif gene abundance and percentage of reads with an Alignment Score (AS) over 60, for samples obtained from Midtre Lovénbreen (Ml), Svalbard. The Alignment Score ranges between 0 and the maximum length of the reads (0-100) and indicates the quality of the alignment between reads and contigs.*



***A2 Figure 6:*** *rpoB normalized nif gene abundance and percentage of reads with an Alignment Score (AS) over 60, for samples obtained from Russell Glacier (Rl), Greenland. The Alignment Score ranges between 0 and the maximum length of the reads (0-150) and indicates the quality of the alignment between reads and contigs.*



***A2 Figure 7:*** *rpoB normalized nif gene abundance and percentage of reads with an Alignment Score (AS) over 60, for samples obtained from Rabots Glacier (Rb), N-Sweden. The Alignment Score ranges between 0 and the maximum length of the reads (0-150) and indicates the quality of the alignment between reads and contigs.*



***A2 Figure 8:*** *rpoB normalized nif gene abundance and percentage of reads with an Alignment Score (AS) over 60, for samples obtained from Storglaciären Glacier (St), N-Sweden. The Alignment Score ranges between 0 and the maximum length of the reads (0-150) and indicates the quality of the alignment between reads and contigs.*

***A2 Table 3:*** *Mapping alignments between nif genes and raw sequencing reads. The number of read alignments in each metagenome is shown, grouped by the alignment score (AS) >=30, >=60, >=90, >=120 and >=140. A total for each forefield is also provided. The alignment score represents the quality of the alignment and ranges between 0-100 for the MI dataset and 0-150 for the Rb, St and RI datasets. The number of sequencing reads is provided, shown as the total number of forward and reverse reads for each sample. The percentage of alignments with an AS equal or higher than 60 is listed for each metagenome sample (% AS >= 60).*

| Sample | Number of reads | AS >= 30 | AS >= 60 | AS >= 90 | AS >= 120 | AS >= 140 | % AS >= 60 |
|---|---|---|---|---|---|---|---|
| **MI 1** | 17465080 | 407 | 342 | 280 | - | - | 0.0019582 |
| **MI 2** | 28186722 | 1175 | 1089 | 947 | - | - | 0.0038635 |
| **MI 3** | 16801828 | 62 | 39 | 20 | - | - | 0.0002321 |
| **MI 4** | 25805412 | 466 | 268 | 79 | - | - | 0.0010385 |
| **MI 5** | 20630402 | 242 | 118 | 16 | - | - | 0.0005720 |
| **MI 6** | 21421268 | 1653 | 1008 | 449 | - | - | 0.0047056 |
| **MI 7** | 22655970 | 172 | 91 | 8 | - | - | 0.0004017 |
| **MI 8** | 20003790 | 1040 | 542 | 69 | - | - | 0.0027095 |
| **MI 9** | 22650764 | 344 | 175 | 17 | - | - | 0.0007726 |
| **MI 10** | 58387594 | 2260 | 1336 | 230 | - | - | 0.0022882 |
| **MI 11** | 56518916 | 244 | 135 | 30 | - | - | 0.0002389 |
| **MI 12** | 37925220 | 314 | 161 | 22 | - | - | 0.0004245 |
| **MI 13** | 3952496 | 72 | 38 | 3 | - | - | 0.0009614 |
| **MI 14** | 22179120 | 660 | 355 | 46 | - | - | 0.0016006 |
| **MI 15** | 6433798 | 164 | 93 | 11 | - | - | 0.0014455 |
| **MI 16** | 22231192 | 125 | 68 | 8 | - | - | 0.0003059 |
| **MI 17** | 38684672 | 333 | 197 | 31 | - | - | 0.0005092 |
| **MI 18** | 68392102 | 932 | 560 | 93 | - | - | 0.0008188 |
| **MI 19** | 23405212 | 79 | 40 | 6 | - | - | 0.0001709 |
| **MI 20** | 20152212 | 37 | 20 | 8 | - | - | 0.0000992 |
| **MI 21** | 28777734 | 98 | 61 | 12 | - | - | 0.0002120 |
| **MI 22** | 22182332 | 24 | 4 | 0 | - | - | 0.0000180 |
| **MI 23** | 28234260 | 6 | 0 | 0 | - | - | 0.0000000 |
| **MI forefield** | 633078096 | 10909 | 6740 | 2385 | 0 | 0 | 0.0010646 |
| **RI 1** | 67748804 | 28 | 0 | 0 | 0 | 0 | 0.0000000 |
| **RI 2** | 77474456 | 20 | 0 | 0 | 0 | 0 | 0.0000000 |
| **RI 3** | 61199512 | 39 | 11 | 7 | 0 | 4 | 0.0000180 |
| **RI 4** | 74992750 | 130 | 80 | 64 | 23 | 49 | 0.0001067 |
| **RI 5** | 74220220 | 109 | 60 | 43 | 16 | 34 | 0.0000808 |
| **RI 6** | 68161082 | 240 | 156 | 121 | 46 | 90 | 0.0002289 |
| **RI 7** | 75975046 | 28 | 0 | 0 | 0 | 0 | 0.0000000 |
| **RI 8** | 91844214 | 47 | 0 | 0 | 0 | 0 | 0.0000000 |
| **RI 9** | 77023868 | 34 | 6 | 4 | 2 | 3 | 0.0000078 |
| **RI 10** | 139381778 | 32 | 5 | 1 | 0 | 0 | 0.0000036 |
| **RI 11** | 91962344 | 131 | 63 | 32 | 16 | 27 | 0.0000685 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Rl 12** | 81681122 | 61 | 29 | 18 | 1 | 9 | 0.0000355 |
| **Rl 13** | 105431412 | 190 | 143 | 92 | 18 | 53 | 0.0001356 |
| **Rl 14** | 63240678 | 26 | 10 | 4 | 0 | 3 | 0.0000158 |
| **Rl 15** | 72243790 | 428 | 215 | 76 | 5 | 29 | 0.0002976 |
| **Rl 16** | 82137112 | 23 | 6 | 1 | 0 | 0 | 0.0000073 |
| **Rl 17** | 107570706 | 74 | 24 | 10 | 4 | 6 | 0.0000223 |
| **Rl 18** | 78274268 | 33 | 3 | 1 | 0 | 1 | 0.0000038 |
| **Rl 19** | 75392190 | 30 | 10 | 2 | 0 | 0 | 0.0000133 |
| **Rl 20** | 65582044 | 16 | 0 | 0 | 0 | 0 | 0.0000000 |
| **Rl 21** | 91522052 | 24 | 1 | 0 | 0 | 0 | 0.0000011 |
| **Rl 22** | 96384726 | 43 | 14 | 3 | 0 | 2 | 0.0000145 |
| **Rl 23** | 96207238 | 5 | 0 | 0 | 0 | 0 | 0.0000000 |
| **Rl 24** | 80981820 | 30 | 9 | 3 | 0 | 1 | 0.0000111 |
| **Rl forefield** | 1996633232 | 1821 | 845 | 482 | 131 | 311 | 0.0000423 |
| **Rb 1** | 76498938 | 95 | 29 | 23 | 2 | 11 | 0.0000379 |
| **Rb 2** | 72691820 | 211 | 117 | 76 | 21 | 49 | 0.0001610 |
| **Rb 3** | 70126632 | 379 | 199 | 132 | 55 | 104 | 0.0002838 |
| **Rb 4** | 61828928 | 584 | 396 | 307 | 99 | 226 | 0.0006405 |
| **Rb 5** | 74667258 | 215 | 105 | 52 | 10 | 32 | 0.0001406 |
| **Rb Forefield** | 355813576 | 1484 | 846 | 590 | 187 | 422 | 0.0002378 |
| **St 1** | 68111048 | 247 | 120 | 67 | 8 | 37 | 0.0001762 |
| **St 2** | 71828498 | 1817 | 1385 | 1031 | 480 | 749 | 0.0019282 |
| **St 3** | 85214054 | 2871 | 2235 | 1571 | 734 | 1143 | 0.0026228 |
| **St 4** | 71411294 | 1967 | 1552 | 1150 | 484 | 804 | 0.0021733 |
| **St 5** | 66910678 | 726 | 563 | 457 | 285 | 389 | 0.0008414 |
| **St 6** | 72683122 | 142 | 59 | 39 | 17 | 28 | 0.0000812 |
| **St 7** | 60148730 | 645 | 481 | 349 | 190 | 274 | 0.0007997 |
| **St 8** | 67844804 | 1956 | 1586 | 1213 | 784 | 1026 | 0.0023377 |
| **St 9** | 63953088 | 1447 | 1222 | 1003 | 616 | 851 | 0.0019108 |
| **St 10** | 83010234 | 102 | 60 | 44 | 25 | 38 | 0.0000723 |
| **St 11** | 74901072 | 577 | 459 | 369 | 182 | 297 | 0.0006128 |
| **St 12** | 83405572 | 586 | 384 | 259 | 100 | 182 | 0.0004604 |
| **St 13** | 64225756 | 119 | 86 | 71 | 19 | 44 | 0.0001339 |
| **St 14** | 64764076 | 295 | 229 | 182 | 83 | 140 | 0.0003536 |
| **St 15** | 66675200 | 350 | 238 | 185 | 101 | 150 | 0.0003570 |
| **St 16** | 66848090 | 363 | 302 | 250 | 127 | 203 | 0.0004518 |
| **St 17** | 72051286 | 1064 | 947 | 760 | 447 | 611 | 0.0013143 |
| **St 18** | 78397352 | 485 | 346 | 260 | 118 | 209 | 0.0004413 |
| **St Forefield** | 1282383954 | 15759 | 12254 | 9260 | 4800 | 7175 | 0.0009556 |

# Appendix 3

***A3 Table 1:** COG domains used in Tree construction using KBASE Species Tree Builder for binned genomes, sourced from https://kbase.us*

| COG | Gene | Description |
|---|---|---|
| COG0013 | AlaS | Alanyl-tRNA synthetase |
| COG0016 | PheS | Phenylalanyl-tRNA synthetase alpha subunit |
| COG0018 | ArgS | Arginyl-tRNA synthetase |
| COG0030 | KsgA | Dimethyladenosine transferase (rRNA methylation) |
| COG0041 | PurE | Phosphoribosylcarboxyaminoimidazole (NCAIR) mutase |
| COG0046 | PurL | Phosphoribosylformylglycinamidine (FGAM) synthase, synthetase domain |
| COG0048 | RpsL | Ribosomal protein S12 |
| COG0049 | RpsG | Ribosomal protein S7 |
| COG0051 | RpsJ | Ribosomal protein S10 |
| COG0052 | RpsB | Ribosomal protein S2 |
| COG0072 | PheT | Phenylalanyl-tRNA synthetase beta subunit |
| COG0080 | RplK | Ribosomal protein L11 |
| COG0081 | RplA | Ribosomal protein L1 |
| COG0082 | AroC | Chorismate synthase |
| COG0086 | RpoC | DNA-directed RNA polymerase, beta' subunit/160 kD subunit |
| COG0087 | RplC | Ribosomal protein L3 |
| COG0088 | RplD | Ribosomal protein L4 |
| COG0089 | RplW | Ribosomal protein L23 |
| COG0090 | RplB | Ribosomal protein L2 |
| COG0091 | RplV | Ribosomal protein L22 |
| COG0092 | RpsC | Ribosomal protein S3 |
| COG0093 | RplN | Ribosomal protein L14 |
| COG0094 | RplE | Ribosomal protein L5 |
| COG0096 | RpsH | Ribosomal protein S8 |
| COG0097 | RplF | Ribosomal protein L6P/L9E |
| COG0098 | RpsE | Ribosomal protein S5 |
| COG0099 | RpsM | Ribosomal protein S13 |
| COG0100 | RpsK | Ribosomal protein S11 |
| COG0102 | RplM | Ribosomal protein L13 |
| COG0103 | RpsI | Ribosomal protein S9 |
| COG0105 | Ndk | Nucleoside diphosphate kinase |
| COG0126 | Pgk | 3-phosphoglycerate kinase |
| COG0127 | COG0127 | Xanthosine triphosphate pyrophosphatase |

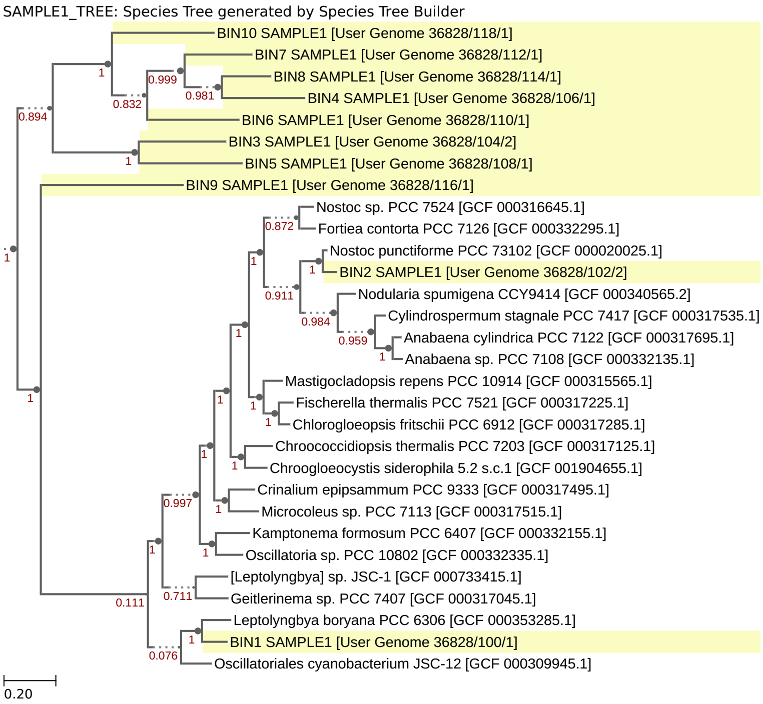
|  |  |  |
|---|---|---|
| COG0130 | TruB | Pseudouridine synthase |
| COG0150 | PurM | Phosphoribosylaminoimidazole (AIR) synthetase |
| COG0151 | PurD | Phosphoribosylamine-glycine ligase |
| COG0164 | RnhB | Ribonuclease HII |
| COG0172 | SerS | Seryl-tRNA synthetase |
| COG0185 | RpsS | Ribosomal protein S19 |
| COG0186 | RpsQ | Ribosomal protein S17 |
| COG0215 | CysS | Cysteinyl-tRNA synthetase |
| COG0244 | RplJ | Ribosomal protein L10 |
| COG0256 | RplR | Ribosomal protein L18 |
| COG0343 | Tgt | Queuine/archaeosine tRNA-ribosyltransferase |
| COG0504 | PyrG | CTP synthase (UTP-ammonia lyase) |
| COG0519 | GuaA | GMP synthase, PP-ATPase domain/subunit |
| COG0532 | InfB | Translation initiation factor 2 (IF-2; GTPase) |
| COG0533 | QRI7 | Metal-dependent proteases with possible chaperone activity |

***A3 Table 2:*** *Read based community composition, expressed as the percentage of annotated reads for Midtre Lovénbreen sample sites at the genus level. Replicate samples for each soil age have been combined. Green indicates the recovery of the genus from each sample.*

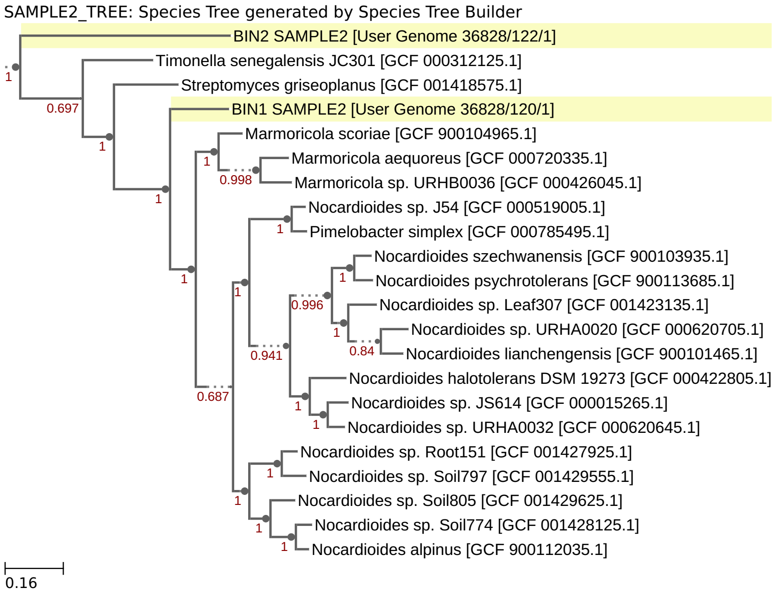
| Genus | 0 | 3 | 5 | 29 | 50 | 50 - 113 | 113 | 2000 | Cryoconite | Basal Ice |
|---|---|---|---|---|---|---|---|---|---|---|
| *Acidobacteriales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.04 | 0.00 |
| *Actinoplanes* | 0.00 | 0.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 0.00 |
| *Arthrobacter* | 0.00 | 0.29 | 0.21 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Bacillales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.55 | 0.00 |
| *Bacillus* | 0.00 | 0.00 | 0.46 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Bacteroidales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.61 | 0.00 |
| *Bradyrhizobium* | 0.00 | 1.42 | 1.03 | 1.48 | 1.23 | 0.75 | 1.18 | 2.78 | 0.00 | 0.67 |
| *Brevundimonas* | 0.87 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Burkholderia* | 0.00 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 0.00 |
| *Burkholderiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 13.96 | 0.00 |
| *Cand. Nitrosocosmicus* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 |
| *Candidatus Solibacter* | 0.00 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.17 | 0.00 | 0.00 |
| *Caulobacterales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.62 | 0.00 |
| *Cellulomonas* | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Chitinophagales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.70 | 0.00 |
| *Chromatiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.65 | 0.00 |
| *Chroococcales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.86 | 0.00 |
| *Chthoniobacter* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.76 | 0.52 | 0.71 | 0.00 | 0.00 |
| *Clostridiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.57 | 0.00 |
| *Collinsella* | 0.00 | 0.00 | 0.00 | 0.42 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Conexibacter* | 0.00 | 0.89 | 0.00 | 0.23 | 0.00 | 0.00 | 1.00 | 0.41 | 0.00 | 0.00 |
| *Corynebacteriales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.74 | 0.00 |
| *Cryobacterium* | 0.51 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Cytophagales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.83 | 0.00 |
| *Desulfuromonadales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.66 | 0.00 |
| *Devosia* | 0.00 | 0.31 | 0.68 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Enterobacterales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.60 | 0.00 |
| *Enterococcus* | 3.42 | 2.72 | 0.67 | 0.00 | 0.00 | 1.25 | 0.00 | 0.77 | 0.00 | 1.88 |
| *Flavisolibacter* | 0.00 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| *Flavobacteriales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.65 | 0.00 |
| *Flavobacterium* | 1.45 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Frankia* | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 0.00 |
| *Frankiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.57 | 0.00 |
| *Gemmata* | 0.00 | 0.45 | 0.19 | 0.83 | 0.72 | 1.00 | 1.43 | 0.00 | 0.00 | 0.00 |
| *Gemmatimonadales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.77 | 0.00 |
| *Gemmatimonas* | 0.00 | 0.34 | 0.00 | 0.00 | 0.23 | 0.00 | 0.19 | 0.00 | 0.00 | 0.00 |
| *Geobacter* | 0.00 | 0.00 | 0.18 | 0.00 | 0.00 | 0.66 | 0.00 | 0.00 | 0.00 | 0.00 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Hymenobacter* | 1.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 0.00 |
| *Hyphomicrobium* | 0.00 | 0.00 | 0.19 | 0.35 | 0.57 | 0.00 | 0.51 | 0.00 | 0.00 | 0.00 |
| *Ilumatobacter* | 0.00 | 0.55 | 0.00 | 0.18 | 0.00 | 0.00 | 0.32 | 0.00 | 0.00 | 0.00 |
| *Intrasporangium* | 0.00 | 0.34 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Kribbella* | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Luteipulveratus* | 0.00 | 0.27 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Marmoricola* | 0.00 | 0.00 | 0.29 | 0.77 | 0.71 | 0.00 | 0.13 | 0.00 | 0.00 | 0.00 |
| *Massilia* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.55 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Mesorhizobium* | 0.00 | 0.35 | 0.42 | 0.73 | 0.78 | 0.00 | 0.70 | 0.66 | 0.00 | 0.00 |
| *Methylibium* | 0.66 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Methylobacterium* | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 0.00 |
| *Methylophilales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.61 | 0.00 |
| *Micrococcales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 8.30 | 0.00 |
| *Micromonosporales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.65 | 0.00 |
| *Mycobacterium* | 0.00 | 0.79 | 0.66 | 1.23 | 1.32 | 1.26 | 1.73 | 1.35 | 0.00 | 0.00 |
| *Myxococcales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.17 | 0.00 |
| *Neisseria* | 0.00 | 0.00 | 0.00 | 0.39 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Niabella* | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Niastella* | 0.00 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 0.00 |
| *Nitrospira* | 0.00 | 0.29 | 0.16 | 0.16 | 0.52 | 0.50 | 0.29 | 0.16 | 0.00 | 0.00 |
| *Nocardioides* | 0.00 | 2.45 | 1.97 | 3.15 | 2.28 | 1.10 | 1.42 | 0.00 | 0.00 | 0.96 |
| *Nostoc* | 1.93 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Nostocales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.86 | 0.00 |
| *Novosphingobium* | 0.54 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.04 |
| *Opitutus* | 0.00 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Oscillatoriales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.29 | 0.00 |
| *Pedobacter* | 1.15 | 0.41 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Pedosphaera* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.39 | 0.00 | 0.00 |
| *Phormidesmis* | 6.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Phycicoccus* | 0.00 | 0.00 | 0.92 | 0.83 | 0.35 | 0.52 | 0.00 | 0.00 | 0.00 | 0.60 |
| *Pimelobacter* | 0.00 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Pirellula* | 0.00 | 0.24 | 0.22 | 0.00 | 0.00 | 0.04 | 0.21 | 0.00 | 0.00 | 0.00 |
| *Planctomyces* | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.00 | 0.00 |
| *Planctomycetales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.62 | 0.00 |
| *Polaromonas* | 3.12 | 1.02 | 0.94 | 1.02 | 0.58 | 0.82 | 0.00 | 0.25 | 0.00 | 0.52 |
| *Propionibacteriales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.12 | 0.00 |
| *Pseudomonadales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.12 | 0.00 |
| *Pseudomonas* | 0.00 | 2.51 | 0.39 | 0.17 | 0.39 | 1.15 | 0.21 | 0.20 | 0.00 | 0.63 |
| *Pseudonocardia* | 0.00 | 0.37 | 0.16 | 0.74 | 0.35 | 0.00 | 0.43 | 0.00 | 0.00 | 0.00 |
| *Pseudonocardiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.24 | 0.00 |
| *Purpureocillium* | 0.58 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Pyrinomonas* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.51 | 0.00 | 0.00 |
| *Rhizobacter* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

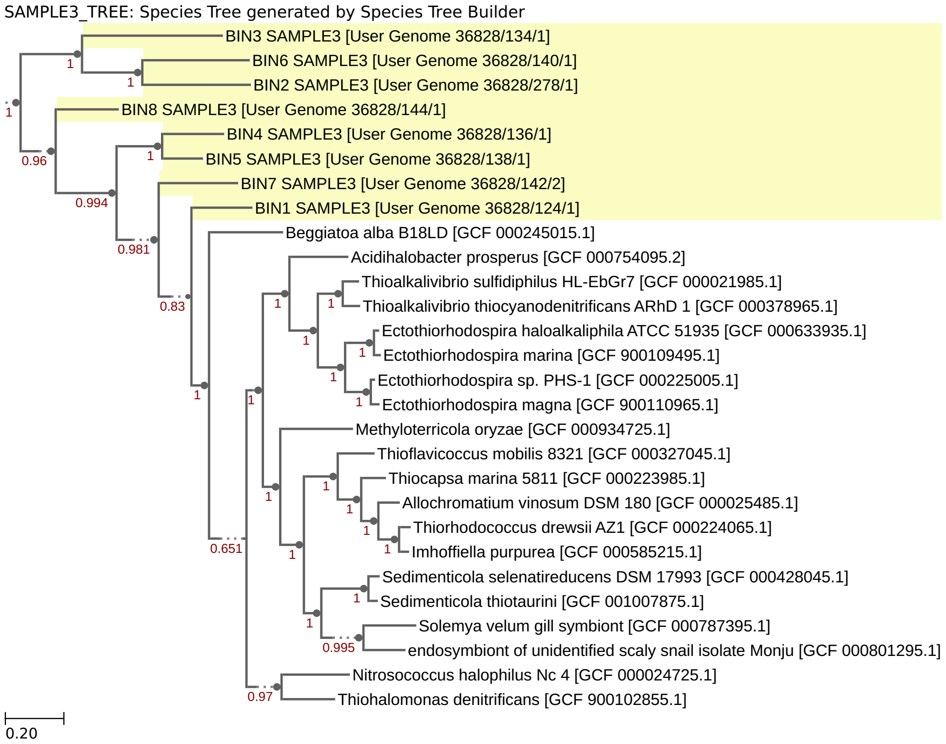
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Rhizobiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.59 | 0.00 |
| *Rhizobium* | 1.22 | 0.00 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Rhodobacterales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.82 | 0.00 |
| *Rhodococcus* | 0.00 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.00 | 0.00 |
| *Rhodocyclales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.73 | 0.00 |
| *Rhodoferax* | 0.68 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Rhodoplanes* | 0.00 | 0.27 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.68 | 0.00 | 0.00 |
| *Rhodopseudomonas* | 0.00 | 0.36 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Rhodospirillales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.66 | 0.00 |
| *Shewanella* | 0.00 | 0.00 | 5.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 15.11 |
| *Singulisphaera* | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.00 | 0.52 | 0.00 | 0.00 | 0.00 |
| *Solibacterales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.65 | 0.00 |
| *Solirubrobacter* | 0.00 | 0.00 | 0.00 | 0.48 | 0.69 | 0.00 | 0.93 | 1.23 | 0.00 | 0.00 |
| *Sorangium* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| *Sphingobacteriales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.53 | 0.00 |
| *Sphingobium* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 | 0.00 |
| *Sphingomonadales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.28 | 0.00 |
| *Sphingomonas* | 1.25 | 1.47 | 1.76 | 2.39 | 2.07 | 1.64 | 1.74 | 0.73 | 0.00 | 1.01 |
| *Sphingopyxis* | 0.00 | 0.27 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 | 0.00 |
| *Spirosoma* | 0.00 | 0.30 | 0.28 | 0.00 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 | 0.00 |
| *Streptomyces* | 0.52 | 1.67 | 1.42 | 1.99 | 1.81 | 1.24 | 2.09 | 1.32 | 0.00 | 0.83 |
| *Streptomycetales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.93 | 0.00 |
| *Streptosporangiales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.59 | 0.00 |
| *Synechococcales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.80 | 0.00 |
| *Thiobacillus* | 0.79 | 1.10 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.80 |
| *Variovorax* | 1.05 | 0.00 | 0.20 | 0.20 | 0.93 | 0.64 | 0.28 | 0.00 | 0.00 | 0.00 |
| *Xanthomonadales* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.47 | 0.00 |
| *Zavarzinella* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.29 | 0.00 | 0.00 | 0.00 |
| Viruses | 0.08 | 0.13 | 0.05 | 0.04 | 0.03 | 0.05 | 0.07 | 0.02 | 0.22 | 0.05 |
| cannot be assigned | 31.29 | 26.02 | 33.30 | 33.91 | 36.86 | 35.40 | 30.08 | 45.24 | 6.94 | 31.93 |
| genus > 0.5% of all reads | 41.29 | 48.47 | 46.66 | 47.25 | 47.58 | 49.64 | 50.84 | 42.42 | 15.44 | 43.96 |


SAMPLE1_TREE: Species Tree generated by Species Tree Builder
BIN10 SAMPLE1 [User Genome 36828/118/1]
BIN7 SAMPLE1 [User Genome 36828/112/1]
BIN8 SAMPLE1 [User Genome 36828/114/1]
BIN4 SAMPLE1 [User Genome 36828/106/1]
BIN6 SAMPLE1 [User Genome 36828/110/1]
BIN3 SAMPLE1 [User Genome 36828/104/2]
BIN5 SAMPLE1 [User Genome 36828/108/1]
BIN9 SAMPLE1 [User Genome 36828/116/1]
Nostoc sp. PCC 7524 [GCF 000316645.1]
Fortiea contorta PCC 7126 [GCF 000332295.1]
Nostoc punctiforme PCC 73102 [GCF 000020025.1]
BIN2 SAMPLE1 [User Genome 36828/102/2]
Nodularia spumigena CCY9414 [GCF 000340565.2]
Cylindrospermum stagnale PCC 7417 [GCF 000317535.1]
Anabaena cylindrica PCC 7122 [GCF 000317695.1]
Anabaena sp. PCC 7108 [GCF 000332135.1]
Mastigocladopsis repens PCC 10914 [GCF 000315565.1]
Fischerella thermalis PCC 7521 [GCF 000317225.1]
Chlorogloeopsis fritschii PCC 6912 [GCF 000317285.1]
Chroococcidiopsis thermalis PCC 7203 [GCF 000317125.1]
Chroogloeocystis siderophila 5.2 s.c.1 [GCF 001904655.1]
Crinalium epipsammum PCC 9333 [GCF 000317495.1]
Microcoleus sp. PCC 7113 [GCF 000317515.1]
Kamptonema formosum PCC 6407 [GCF 000332155.1]
Oscillatoria sp. PCC 10802 [GCF 000332335.1]
[Leptolyngbya] sp. JSC-1 [GCF 000733415.1]
Geitlerinema sp. PCC 7407 [GCF 000317045.1]
Leptolyngbya boryana PCC 6306 [GCF 000353285.1]
BIN1 SAMPLE1 [User Genome 36828/100/1]
Oscillatoriales cyanobacterium JSC-12 [GCF 000309945.1]
1
0.999
0.981
0.832
0.894
1
1
0.872
1
1
0.911
0.984
0.959
1
1
1
1
1
1
1
0.997
1
1
1
0.711
0.111
1
0.076
0.20

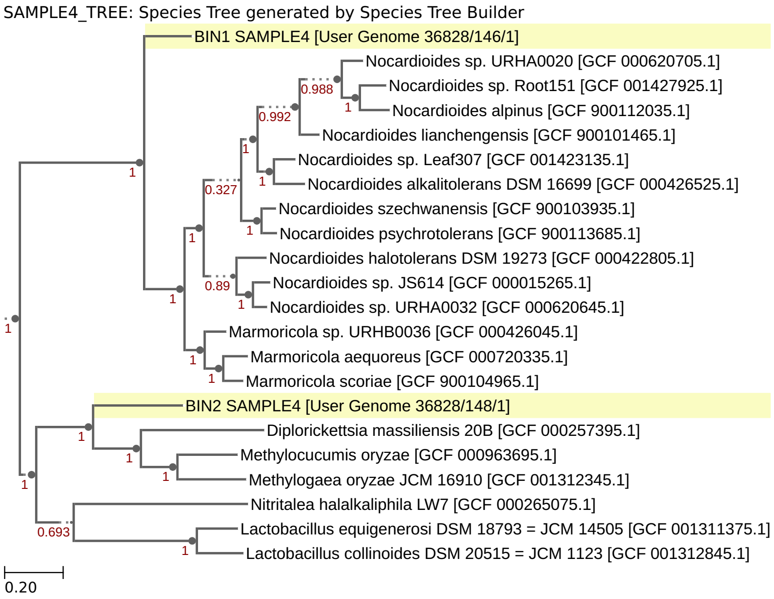
***A3 Figure 1:*** *Maximum likelihood Species tree for Sample 1, bins 1-10. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE2_TREE: Species Tree generated by Species Tree Builder
BIN2 SAMPLE2 [User Genome 36828/122/1]
Timonella senegalensis JC301 [GCF 000312125.1]
Streptomyces griseoplanus [GCF 001418575.1]
BIN1 SAMPLE2 [User Genome 36828/120/1]
Marmoricola scoriae [GCF 900104965.1]
Marmoricola aequoreus [GCF 000720335.1]
Marmoricola sp. URHB0036 [GCF 000426045.1]
Nocardioides sp. J54 [GCF 000519005.1]
Pimelobacter simplex [GCF 000785495.1]
Nocardioides szechwanensis [GCF 900103935.1]
Nocardioides psychrotolerans [GCF 900113685.1]
Nocardioides sp. Leaf307 [GCF 001423135.1]
Nocardioides sp. URHA0020 [GCF 000620705.1]
Nocardioides lianchengensis [GCF 900101465.1]
Nocardioides halotolerans DSM 19273 [GCF 000422805.1]
Nocardioides sp. JS614 [GCF 000015265.1]
Nocardioides sp. URHA0032 [GCF 000620645.1]
Nocardioides sp. Root151 [GCF 001427925.1]
Nocardioides sp. Soil797 [GCF 001429555.1]
Nocardioides sp. Soil805 [GCF 001429625.1]
Nocardioides sp. Soil774 [GCF 001428125.1]
Nocardioides alpinus [GCF 900112035.1]
1
0.697
1
1
1
0.998
1
1
1
0.996
1
0.84
0.941
1
1
0.687
1
1
1
1
0.16

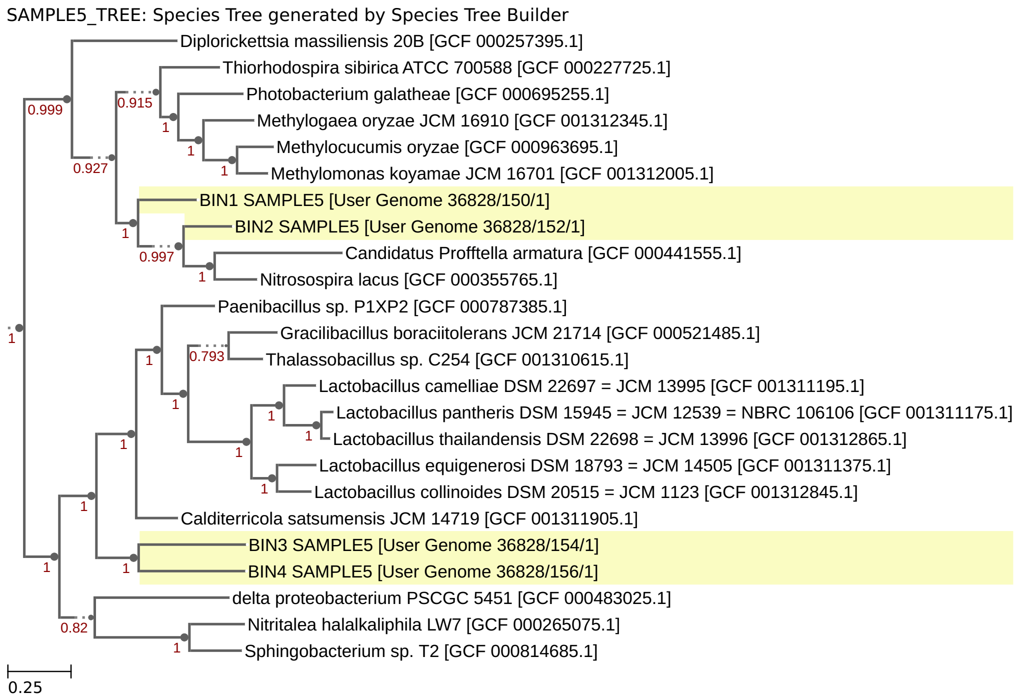
***A3 Figure 2:*** *Maximum likelihood Species tree for Sample 2, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE3_TREE: Species Tree generated by Species Tree Builder
BIN3 SAMPLE3 [User Genome 36828/134/1]
BIN6 SAMPLE3 [User Genome 36828/140/1]
BIN2 SAMPLE3 [User Genome 36828/278/1]
BIN8 SAMPLE3 [User Genome 36828/144/1]
BIN4 SAMPLE3 [User Genome 36828/136/1]
BIN5 SAMPLE3 [User Genome 36828/138/1]
BIN7 SAMPLE3 [User Genome 36828/142/2]
BIN1 SAMPLE3 [User Genome 36828/124/1]
Beggiatoa alba B18LD [GCF 000245015.1]
Acidihalobacter prosperus [GCF 000754095.2]
Thioalkalivibrio sulfidiphilus HL-EbGr7 [GCF 000021985.1]
Thioalkalivibrio thiocyanodenitrificans ARhD 1 [GCF 000378965.1]
Ectothiorhodospira haloalkaliphila ATCC 51935 [GCF 000633935.1]
Ectothiorhodospira marina [GCF 900109495.1]
Ectothiorhodospira sp. PHS-1 [GCF 000225005.1]
Ectothiorhodospira magna [GCF 900110965.1]
Methyloterricola oryzae [GCF 000934725.1]
Thioflavicoccus mobilis 8321 [GCF 000327045.1]
Thiocapsa marina 5811 [GCF 000223985.1]
Allochromatium vinosum DSM 180 [GCF 000025485.1]
Thiorhodococcus drewsii AZ1 [GCF 000224065.1]
Imhoffiella purpurea [GCF 000585215.1]
Sedimenticola selenatireducens DSM 17993 [GCF 000428045.1]
Sedimenticola thiotaurini [GCF 001007875.1]
Solemya velum gill symbiont [GCF 000787395.1]
endosymbiont of unidentified scaly snail isolate Monju [GCF 000801295.1]
Nitrosococcus halophilus Nc 4 [GCF 000024725.1]
Thiohalomonas denitrificans [GCF 900102855.1]
1
1
1
0.96
1
0.994
0.981
0.83
1
1
1
1
1
1
1
1
1
1
1
1
0.651
1
1
0.995
0.97
0.20

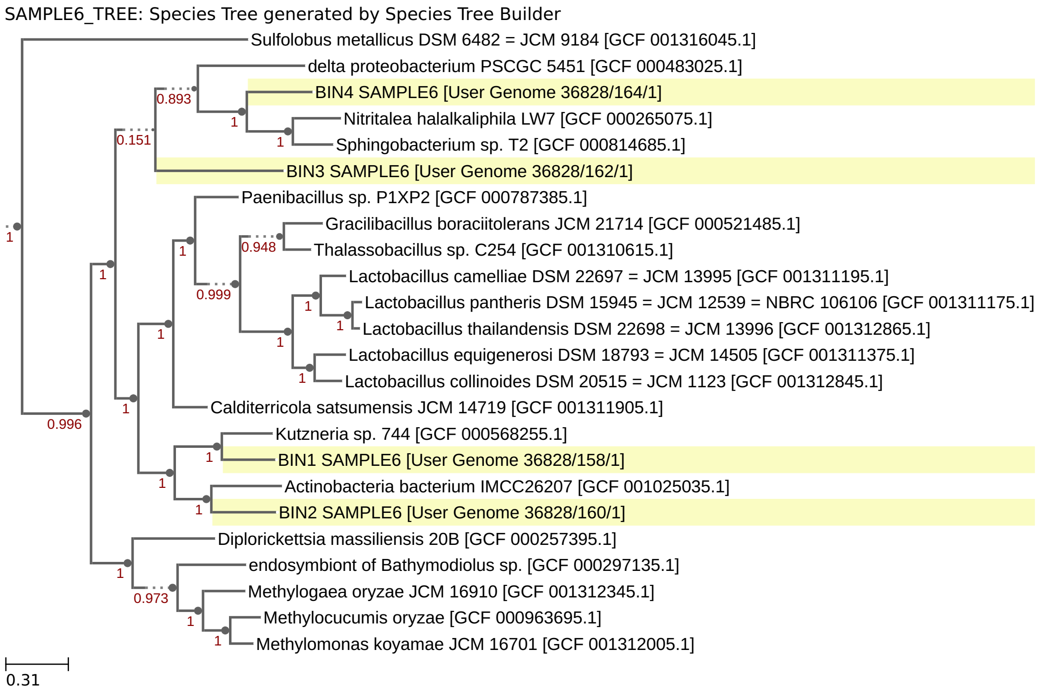
***A3 Figure 3:*** *Maximum likelihood Species tree for Sample 3, bins 1-8. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE4_TREE: Species Tree generated by Species Tree Builder
BIN1 SAMPLE4 [User Genome 36828/146/1]
Nocardioides sp. URHA0020 [GCF 000620705.1]
Nocardioides sp. Root151 [GCF 001427925.1]
Nocardioides alpinus [GCF 900112035.1]
Nocardioides lianchengensis [GCF 900101465.1]
Nocardioides sp. Leaf307 [GCF 001423135.1]
Nocardioides alkalitolerans DSM 16699 [GCF 000426525.1]
Nocardioides szechwanensis [GCF 900103935.1]
Nocardioides psychrotolerans [GCF 900113685.1]
Nocardioides halotolerans DSM 19273 [GCF 000422805.1]
Nocardioides sp. JS614 [GCF 000015265.1]
Nocardioides sp. URHA0032 [GCF 000620645.1]
Marmoricola sp. URHB0036 [GCF 000426045.1]
Marmoricola aequoreus [GCF 000720335.1]
Marmoricola scoriae [GCF 900104965.1]
BIN2 SAMPLE4 [User Genome 36828/148/1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
0.988
1
0.992
1
1
0.327
1
1
1
0.89
1
1
1
1
1
1
1
1
1
0.693
1
0.20

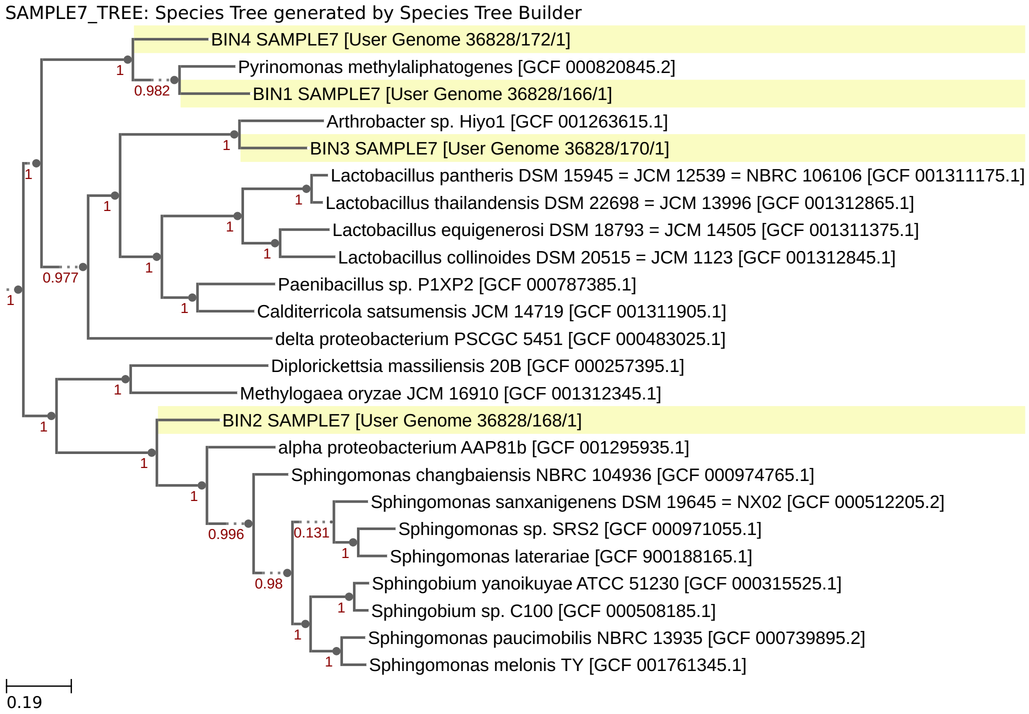
***A3 Figure 4:*** *Maximum likelihood Species tree for Sample 4, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE5_TREE: Species Tree generated by Species Tree Builder
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Thiorhodospira sibirica ATCC 700588 [GCF 000227725.1]
Photobacterium galatheae [GCF 000695255.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
BIN1 SAMPLE5 [User Genome 36828/150/1]
BIN2 SAMPLE5 [User Genome 36828/152/1]
Candidatus Profftella armatura [GCF 000441555.1]
Nitrosospira lacus [GCF 000355765.1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
BIN3 SAMPLE5 [User Genome 36828/154/1]
BIN4 SAMPLE5 [User Genome 36828/156/1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Sphingobacterium sp. T2 [GCF 000814685.1]
0.999
0.915
0.927
0.997
0.793
0.82
1
0.25

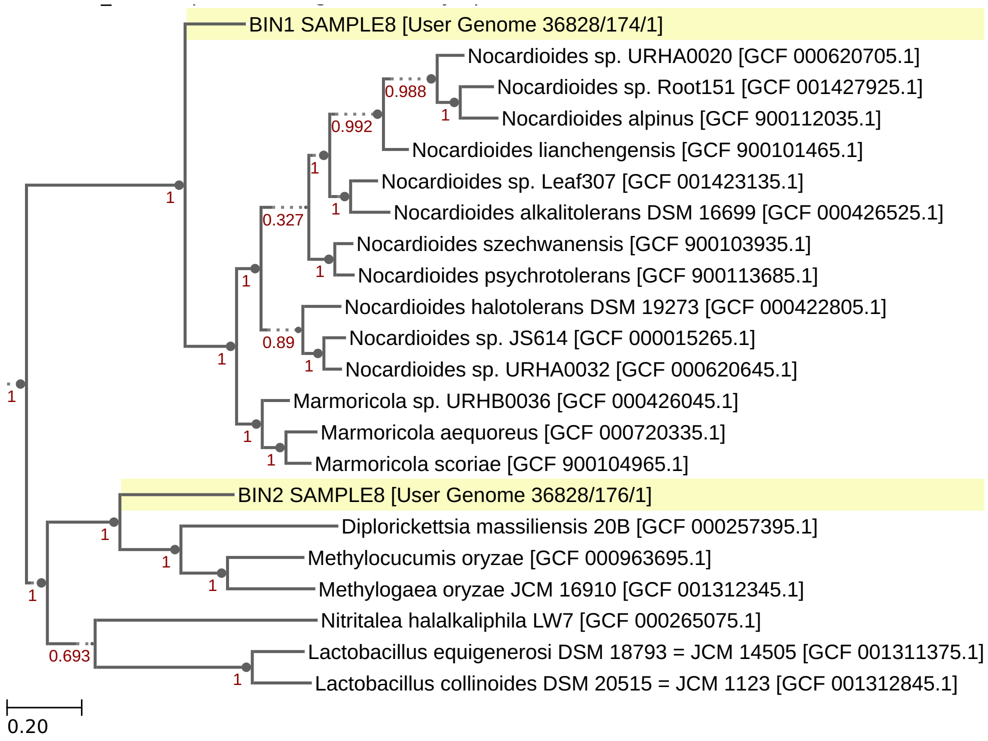
**A3 Figure 5:** *Maximum likelihood Species tree for Sample 1, bins 1-4. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE6_TREE: Species Tree generated by Species Tree Builder
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
BIN4 SAMPLE6 [User Genome 36828/164/1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Sphingobacterium sp. T2 [GCF 000814685.1]
BIN3 SAMPLE6 [User Genome 36828/162/1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Kutzneria sp. 744 [GCF 000568255.1]
BIN1 SAMPLE6 [User Genome 36828/158/1]
Actinobacteria bacterium IMCC26207 [GCF 001025035.1]
BIN2 SAMPLE6 [User Genome 36828/160/1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
endosymbiont of Bathymodiolus sp. [GCF 000297135.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
0.893
0.151
0.948
0.999
0.996
0.973
1
0.31

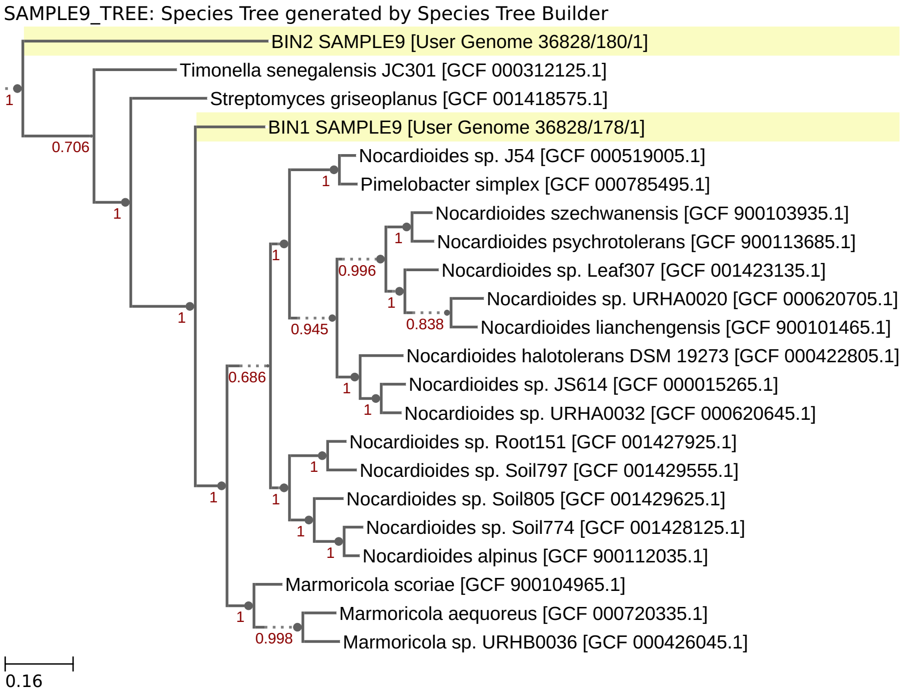
**A3 Figure 6:** *Maximum likelihood Species tree for Sample 6, bins 1 - 4. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE7_TREE: Species Tree generated by Species Tree Builder
BIN4 SAMPLE7 [User Genome 36828/172/1]
Pyrinomonas methylaliphatogenes [GCF 000820845.2]
BIN1 SAMPLE7 [User Genome 36828/166/1]
Arthrobacter sp. Hiyo1 [GCF 001263615.1]
BIN3 SAMPLE7 [User Genome 36828/170/1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
BIN2 SAMPLE7 [User Genome 36828/168/1]
alpha proteobacterium AAP81b [GCF 001295935.1]
Sphingomonas changbaiensis NBRC 104936 [GCF 000974765.1]
Sphingomonas sanxanigenens DSM 19645 = NX02 [GCF 000512205.2]
Sphingomonas sp. SRS2 [GCF 000971055.1]
Sphingomonas laterariae [GCF 900188165.1]
Sphingobium yanoikuyae ATCC 51230 [GCF 000315525.1]
Sphingobium sp. C100 [GCF 000508185.1]
Sphingomonas paucimobilis NBRC 13935 [GCF 000739895.2]
Sphingomonas melonis TY [GCF 001761345.1]
1
0.982
1
1
1
1
1
1
1
0.977
1
1
1
1
1
1
0.996
0.131
1
0.98
1
1
1
1
0.19

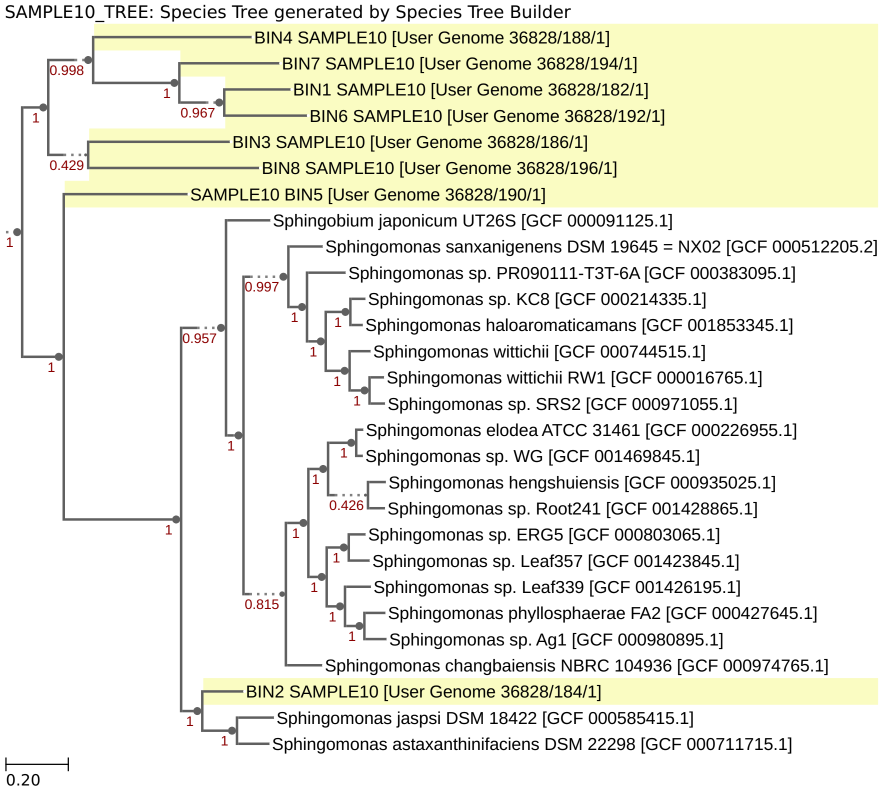
**A3 Figure 7:** *Maximum likelihood Species tree for Sample 1, bins 1-4. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


BIN1 SAMPLE8 [User Genome 36828/174/1]
Nocardioides sp. URHA0020 [GCF 000620705.1]
Nocardioides sp. Root151 [GCF 001427925.1]
Nocardioides alpinus [GCF 900112035.1]
Nocardioides lianchengensis [GCF 900101465.1]
Nocardioides sp. Leaf307 [GCF 001423135.1]
Nocardioides alkalitolerans DSM 16699 [GCF 000426525.1]
Nocardioides szechwanensis [GCF 900103935.1]
Nocardioides psychrotolerans [GCF 900113685.1]
Nocardioides halotolerans DSM 19273 [GCF 000422805.1]
Nocardioides sp. JS614 [GCF 000015265.1]
Nocardioides sp. URHA0032 [GCF 000620645.1]
Marmoricola sp. URHB0036 [GCF 000426045.1]
Marmoricola aequoreus [GCF 000720335.1]
Marmoricola scoriae [GCF 900104965.1]
BIN2 SAMPLE8 [User Genome 36828/176/1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
0.988
1
0.992
1
1
0.327
1
1
0.89
1
1
1
1
1
1
1
1
1
1
0.693
1
0.20

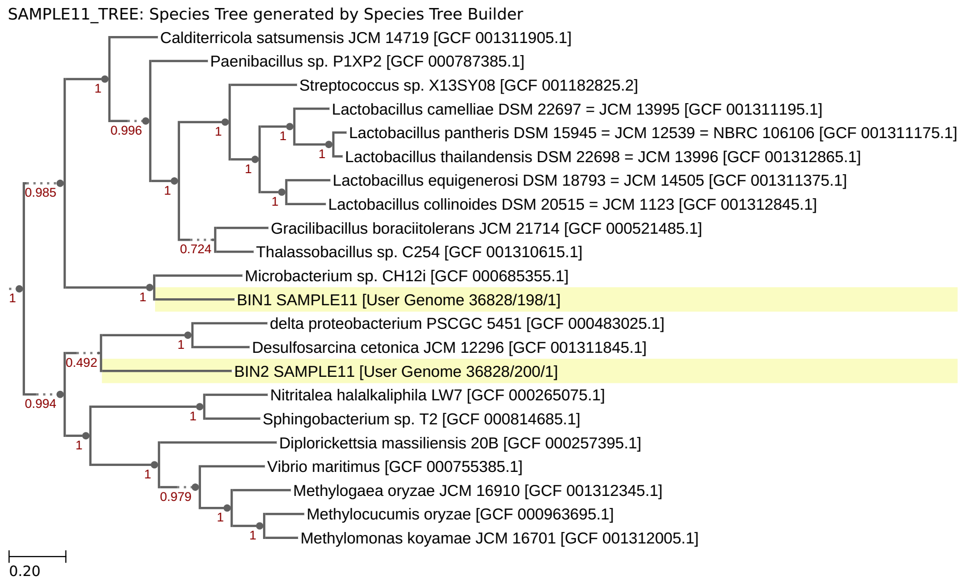
**A3 Figure 8:** *Maximum likelihood Species tree for Sample 8, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE9_TREE: Species Tree generated by Species Tree Builder
BIN2 SAMPLE9 [User Genome 36828/180/1]
Timonella senegalensis JC301 [GCF 000312125.1]
Streptomyces griseoplanus [GCF 001418575.1]
BIN1 SAMPLE9 [User Genome 36828/178/1]
Nocardioides sp. J54 [GCF 000519005.1]
Pimelobacter simplex [GCF 000785495.1]
Nocardioides szechwanensis [GCF 900103935.1]
Nocardioides psychrotolerans [GCF 900113685.1]
Nocardioides sp. Leaf307 [GCF 001423135.1]
Nocardioides sp. URHA0020 [GCF 000620705.1]
Nocardioides lianchengensis [GCF 900101465.1]
Nocardioides halotolerans DSM 19273 [GCF 000422805.1]
Nocardioides sp. JS614 [GCF 000015265.1]
Nocardioides sp. URHA0032 [GCF 000620645.1]
Nocardioides sp. Root151 [GCF 001427925.1]
Nocardioides sp. Soil797 [GCF 001429555.1]
Nocardioides sp. Soil805 [GCF 001429625.1]
Nocardioides sp. Soil774 [GCF 001428125.1]
Nocardioides alpinus [GCF 900112035.1]
Marmoricola scoriae [GCF 900104965.1]
Marmoricola aequoreus [GCF 000720335.1]
Marmoricola sp. URHB0036 [GCF 000426045.1]
1
0.706
1
1
1
1
0.996
1
0.838
0.945
0.686
1
1
1
1
1
1
1
1
0.998
0.16

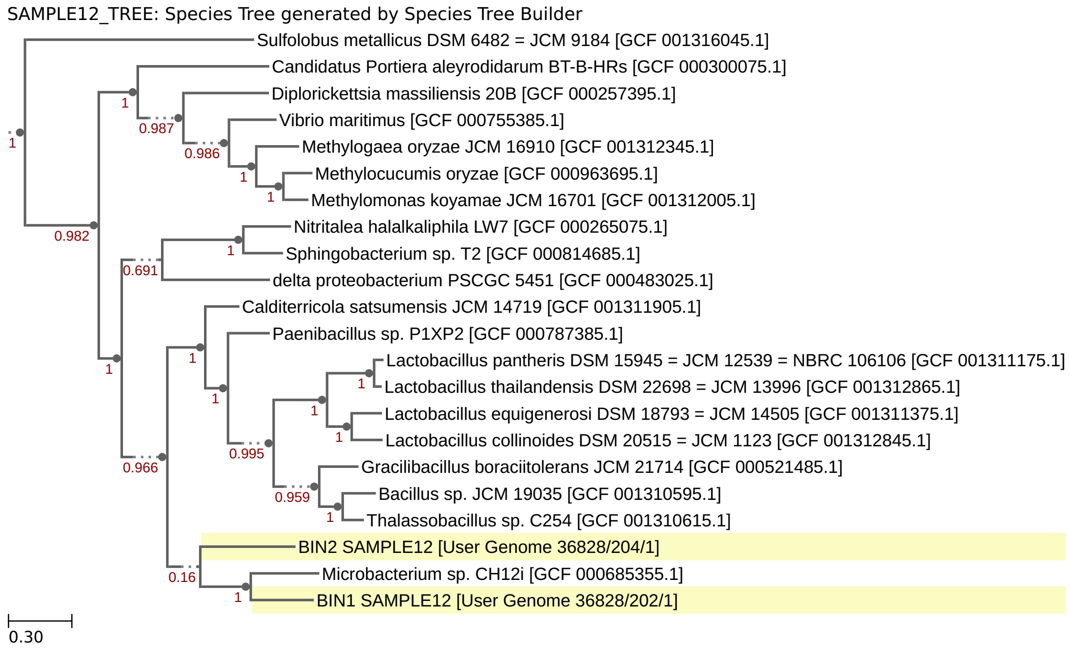
***A3 Figure 9:*** *Maximum likelihood Species tree for Sample 9, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE10_TREE: Species Tree generated by Species Tree Builder
BIN4 SAMPLE10 [User Genome 36828/188/1]
BIN7 SAMPLE10 [User Genome 36828/194/1]
BIN1 SAMPLE10 [User Genome 36828/182/1]
BIN6 SAMPLE10 [User Genome 36828/192/1]
BIN3 SAMPLE10 [User Genome 36828/186/1]
BIN8 SAMPLE10 [User Genome 36828/196/1]
SAMPLE10 BIN5 [User Genome 36828/190/1]
Sphingobium japonicum UT26S [GCF 000091125.1]
Sphingomonas sanxanigenens DSM 19645 = NX02 [GCF 000512205.2]
Sphingomonas sp. PR090111-T3T-6A [GCF 000383095.1]
Sphingomonas sp. KC8 [GCF 000214335.1]
Sphingomonas haloaromaticamans [GCF 001853345.1]
Sphingomonas wittichii [GCF 000744515.1]
Sphingomonas wittichii RW1 [GCF 000016765.1]
Sphingomonas sp. SRS2 [GCF 000971055.1]
Sphingomonas elodea ATCC 31461 [GCF 000226955.1]
Sphingomonas sp. WG [GCF 001469845.1]
Sphingomonas hengshuiensis [GCF 000935025.1]
Sphingomonas sp. Root241 [GCF 001428865.1]
Sphingomonas sp. ERG5 [GCF 000803065.1]
Sphingomonas sp. Leaf357 [GCF 001423845.1]
Sphingomonas sp. Leaf339 [GCF 001426195.1]
Sphingomonas phyllosphaerae FA2 [GCF 000427645.1]
Sphingomonas sp. Ag1 [GCF 000980895.1]
Sphingomonas changbaiensis NBRC 104936 [GCF 000974765.1]
BIN2 SAMPLE10 [User Genome 36828/184/1]
Sphingomonas jaspsi DSM 18422 [GCF 000585415.1]
Sphingomonas astaxanthinifaciens DSM 22298 [GCF 000711715.1]
0.998
1
0.967
1
0.429
1
1
0.997
1
1
1
1
1
0.957
1
1
1
0.426
1
1
1
1
1
0.815
1
1
1
0.20

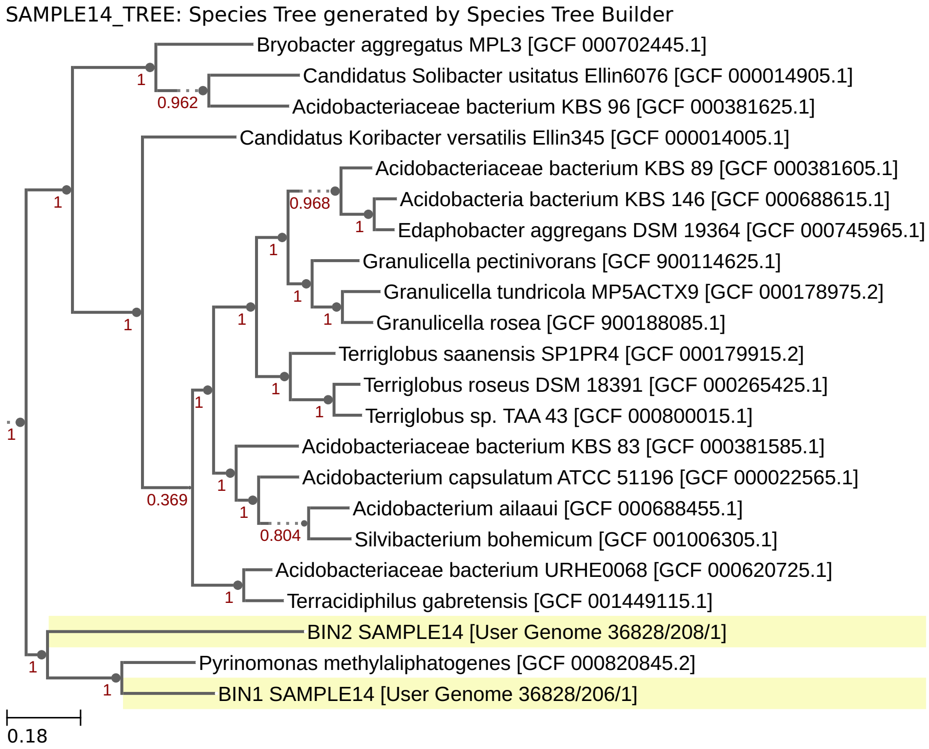
***A3 Figure 10:*** *Maximum likelihood Species tree for Sample 1, bins 1-8. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE11_TREE: Species Tree generated by Species Tree Builder
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Streptococcus sp. X13SY08 [GCF 001182825.2]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
Microbacterium sp. CH12i [GCF 000685355.1]
BIN1 SAMPLE11 [User Genome 36828/198/1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
Desulfosarcina cetonica JCM 12296 [GCF 001311845.1]
BIN2 SAMPLE11 [User Genome 36828/200/1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Sphingobacterium sp. T2 [GCF 000814685.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Vibrio maritimus [GCF 000755385.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
1
0.996
0.985
1
1
1
1
1
1
0.724
1
1
1
0.492
0.994
1
1
1
0.979
1
1
0.20

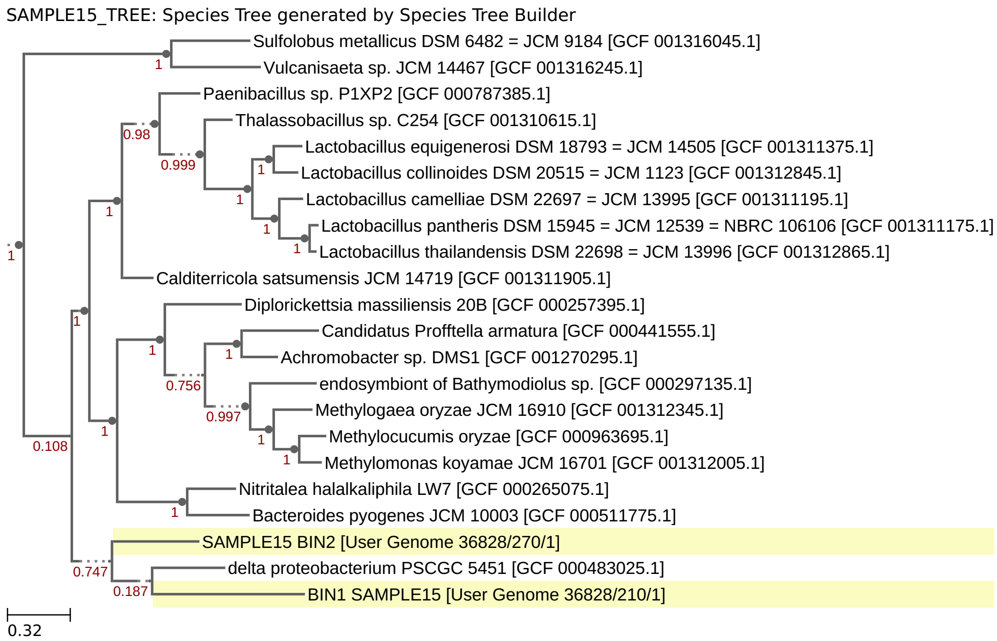
***A3 Figure 11:*** *Maximum likelihood Species tree for Sample 11, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE12_TREE: Species Tree generated by Species Tree Builder
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
Candidatus Portiera aleyrodidarum BT-B-HRs [GCF 000300075.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Vibrio maritimus [GCF 000755385.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Sphingobacterium sp. T2 [GCF 000814685.1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
Bacillus sp. JCM 19035 [GCF 001310595.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
BIN2 SAMPLE12 [User Genome 36828/204/1]
Microbacterium sp. CH12i [GCF 000685355.1]
BIN1 SAMPLE12 [User Genome 36828/202/1]
1
1
0.987
0.986
1
1
0.982
1
0.691
1
1
1
0.995
1
1
1
0.966
0.959
1
0.16
1
0.30

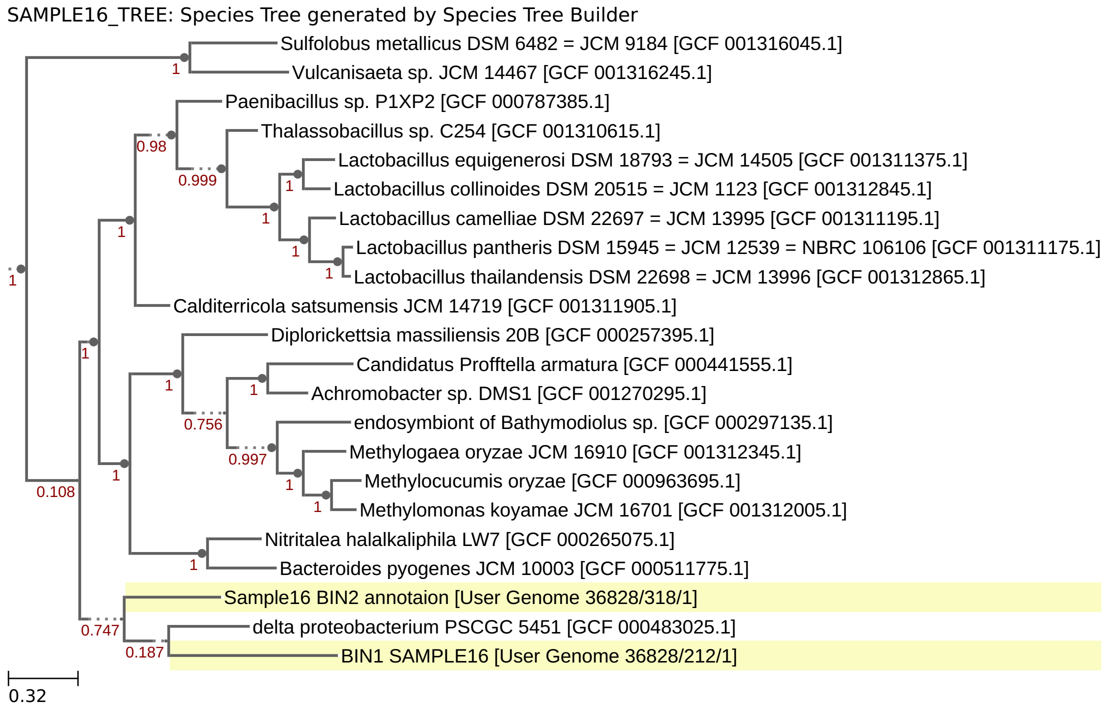
***A3 Figure 12:*** *Maximum likelihood Species tree for Sample 12, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE14_TREE: Species Tree generated by Species Tree Builder
Bryobacter aggregatus MPL3 [GCF 000702445.1]
Candidatus Solibacter usitatus Ellin6076 [GCF 000014905.1]
Acidobacteriaceae bacterium KBS 96 [GCF 000381625.1]
Candidatus Koribacter versatilis Ellin345 [GCF 000014005.1]
Acidobacteriaceae bacterium KBS 89 [GCF 000381605.1]
Acidobacteria bacterium KBS 146 [GCF 000688615.1]
Edaphobacter aggregans DSM 19364 [GCF 000745965.1]
Granulicella pectinivorans [GCF 900114625.1]
Granulicella tundricola MP5ACTX9 [GCF 000178975.2]
Granulicella rosea [GCF 900188085.1]
Terriglobus saanensis SP1PR4 [GCF 000179915.2]
Terriglobus roseus DSM 18391 [GCF 000265425.1]
Terriglobus sp. TAA 43 [GCF 000800015.1]
Acidobacteriaceae bacterium KBS 83 [GCF 000381585.1]
Acidobacterium capsulatum ATCC 51196 [GCF 000022565.1]
Acidobacterium ailaaui [GCF 000688455.1]
Silvibacterium bohemicum [GCF 001006305.1]
Acidobacteriaceae bacterium URHE0068 [GCF 000620725.1]
Terracidiphilus gabretensis [GCF 001449115.1]
BIN2 SAMPLE14 [User Genome 36828/208/1]
Pyrinomonas methylaliphatogenes [GCF 000820845.2]
BIN1 SAMPLE14 [User Genome 36828/206/1]
1
0.962
1
0.968
1
1
1
1
1
1
1
1
1
1
1
0.369
1
0.804
1
1
1
1
0.18

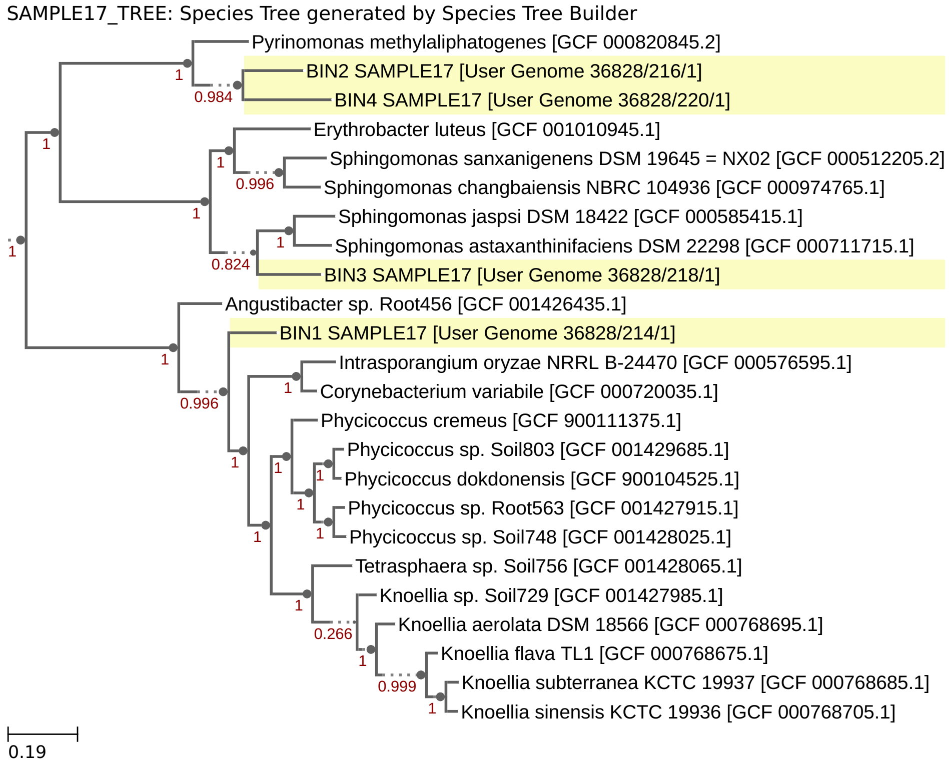
***A3 Figure 13:*** *Maximum likelihood Species tree for Sample 14, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE15_TREE: Species Tree generated by Species Tree Builder
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
Vulcanisaeta sp. JCM 14467 [GCF 001316245.1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Candidatus Profftella armatura [GCF 000441555.1]
Achromobacter sp. DMS1 [GCF 001270295.1]
endosymbiont of Bathymodiolus sp. [GCF 000297135.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Bacteroides pyogenes JCM 10003 [GCF 000511775.1]
SAMPLE15 BIN2 [User Genome 36828/270/1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
BIN1 SAMPLE15 [User Genome 36828/210/1]
1
0.98
0.999
1
1
1
1
1
1
1
0.756
1
0.997
1
1
1
0.108
1
0.747
0.187
0.32

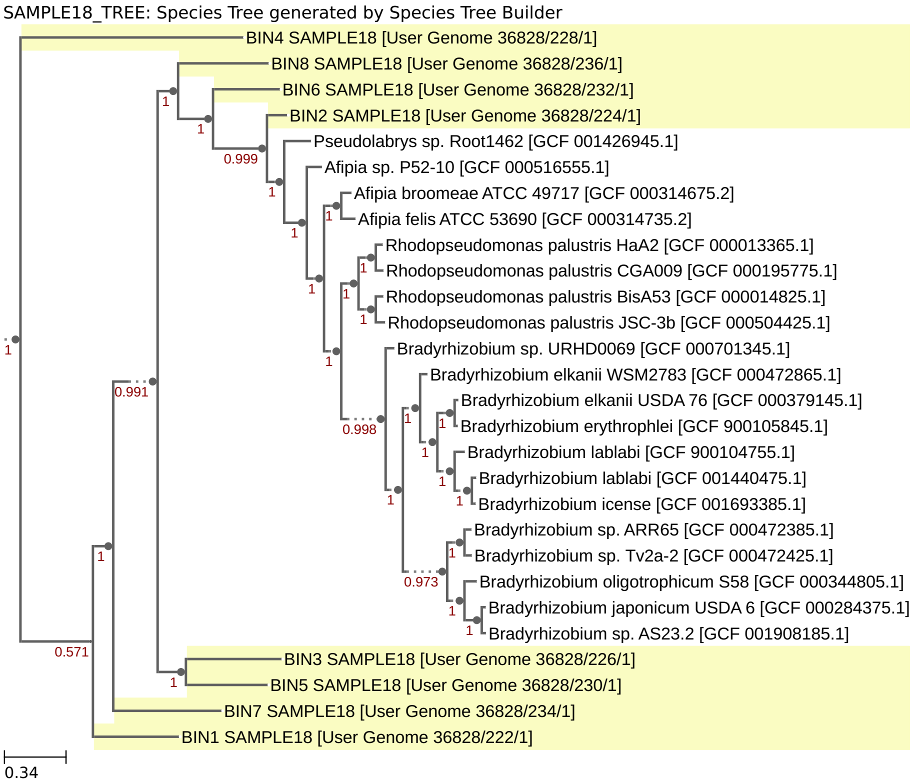
***A3 Figure 14:*** *Maximum likelihood Species tree for Sample 15, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE16_TREE: Species Tree generated by Species Tree Builder
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
Vulcanisaeta sp. JCM 14467 [GCF 001316245.1]
1
Paenibacillus sp. P1XP2 [GCF 000787385.1]
0.98
Thalassobacillus sp. C254 [GCF 001310615.1]
0.999
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
1
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
1
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
1
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
1
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
1
1
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
1
Candidatus Profftella armatura [GCF 000441555.1]
1
Achromobacter sp. DMS1 [GCF 001270295.1]
1
0.756
endosymbiont of Bathymodiolus sp. [GCF 000297135.1]
0.997
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
1
Methylocucumis oryzae [GCF 000963695.1]
1
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
1
0.108
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
1
Bacteroides pyogenes JCM 10003 [GCF 000511775.1]
Sample16 BIN2 annotaion [User Genome 36828/318/1]
0.747
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
0.187
BIN1 SAMPLE16 [User Genome 36828/212/1]
0.32

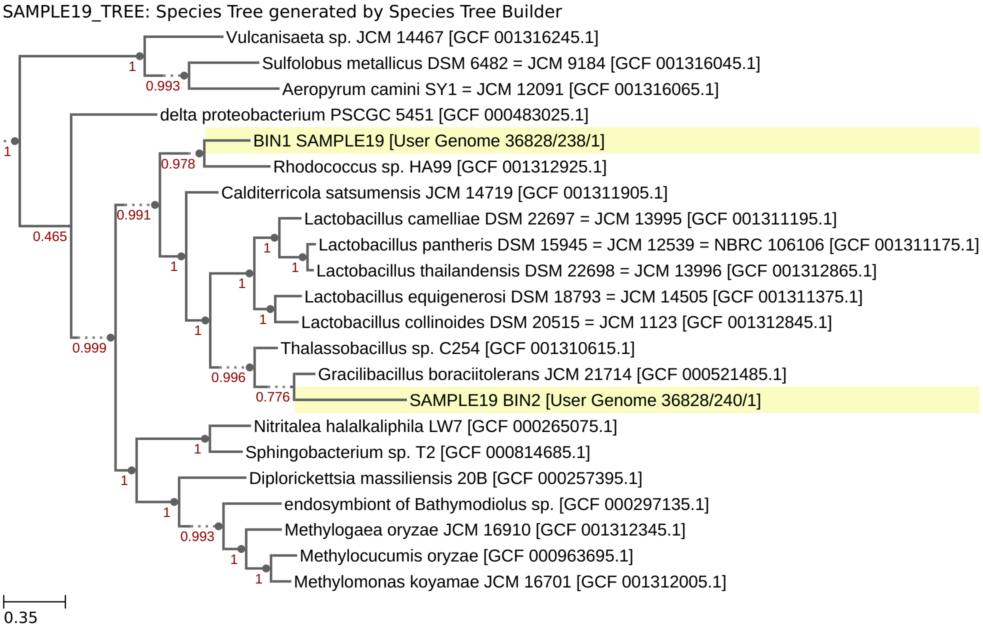
***A3 Figure 15:*** *Maximum likelihood Species tree for Sample 16, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE17_TREE: Species Tree generated by Species Tree Builder
Pyrinomonas methylaliphatogenes [GCF 000820845.2]
1
BIN2 SAMPLE17 [User Genome 36828/216/1]
0.984
BIN4 SAMPLE17 [User Genome 36828/220/1]
1
Erythrobacter luteus [GCF 001010945.1]
1
Sphingomonas sanxanigenens DSM 19645 = NX02 [GCF 000512205.2]
0.996
Sphingomonas changbaiensis NBRC 104936 [GCF 000974765.1]
1
Sphingomonas jaspsi DSM 18422 [GCF 000585415.1]
1
Sphingomonas astaxanthinifaciens DSM 22298 [GCF 000711715.1]
0.824
BIN3 SAMPLE17 [User Genome 36828/218/1]
1
Angustibacter sp. Root456 [GCF 001426435.1]
1
BIN1 SAMPLE17 [User Genome 36828/214/1]
Intrasporangium oryzae NRRL B-24470 [GCF 000576595.1]
1
Corynebacterium variabile [GCF 000720035.1]
0.996
Phycicoccus cremeus [GCF 900111375.1]
1
Phycicoccus sp. Soil803 [GCF 001429685.1]
1
Phycicoccus dokdonensis [GCF 900104525.1]
1
1
Phycicoccus sp. Root563 [GCF 001427915.1]
1
Phycicoccus sp. Soil748 [GCF 001428025.1]
1
Tetrasphaera sp. Soil756 [GCF 001428065.1]
1
Knoellia sp. Soil729 [GCF 001427985.1]
0.266
Knoellia aerolata DSM 18566 [GCF 000768695.1]
1
Knoellia flava TL1 [GCF 000768675.1]
0.999
Knoellia subterranea KCTC 19937 [GCF 000768685.1]
1
Knoellia sinensis KCTC 19936 [GCF 000768705.1]
0.19

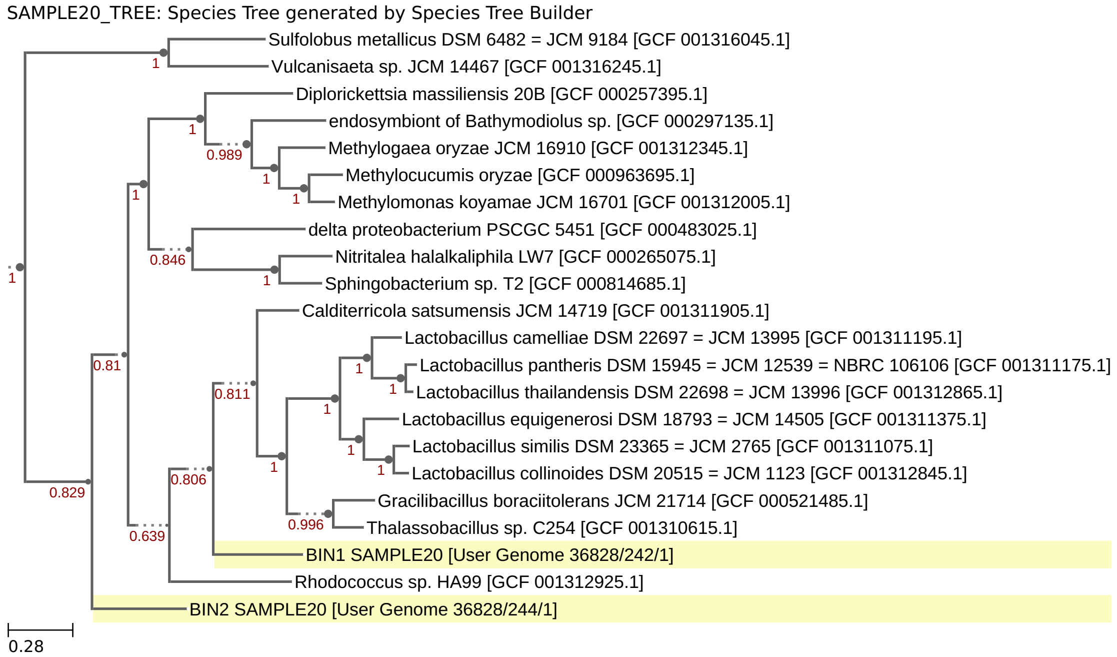
***A3 Figure 16:*** *Maximum likelihood Species tree for Sample 17, bins 1-4. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE18_TREE: Species Tree generated by Species Tree Builder
BIN4 SAMPLE18 [User Genome 36828/228/1]
BIN8 SAMPLE18 [User Genome 36828/236/1]
BIN6 SAMPLE18 [User Genome 36828/232/1]
BIN2 SAMPLE18 [User Genome 36828/224/1]
Pseudolabrys sp. Root1462 [GCF 001426945.1]
Afipia sp. P52-10 [GCF 000516555.1]
Afipia broomeae ATCC 49717 [GCF 000314675.2]
Afipia felis ATCC 53690 [GCF 000314735.2]
Rhodopseudomonas palustris HaA2 [GCF 000013365.1]
Rhodopseudomonas palustris CGA009 [GCF 000195775.1]
Rhodopseudomonas palustris BisA53 [GCF 000014825.1]
Rhodopseudomonas palustris JSC-3b [GCF 000504425.1]
Bradyrhizobium sp. URHD0069 [GCF 000701345.1]
Bradyrhizobium elkanii WSM2783 [GCF 000472865.1]
Bradyrhizobium elkanii USDA 76 [GCF 000379145.1]
Bradyrhizobium erythrophlei [GCF 900105845.1]
Bradyrhizobium lablabi [GCF 900104755.1]
Bradyrhizobium lablabi [GCF 001440475.1]
Bradyrhizobium icense [GCF 001693385.1]
Bradyrhizobium sp. ARR65 [GCF 000472385.1]
Bradyrhizobium sp. Tv2a-2 [GCF 000472425.1]
Bradyrhizobium oligotrophicum S58 [GCF 000344805.1]
Bradyrhizobium japonicum USDA 6 [GCF 000284375.1]
Bradyrhizobium sp. AS23.2 [GCF 001908185.1]
BIN3 SAMPLE18 [User Genome 36828/226/1]
BIN5 SAMPLE18 [User Genome 36828/230/1]
BIN7 SAMPLE18 [User Genome 36828/234/1]
BIN1 SAMPLE18 [User Genome 36828/222/1]
1
1
0.999
1
1
1
1
1
1
1
1
0.991
0.998
1
1
1
1
1
1
0.973
1
1
1
0.571
1
0.34

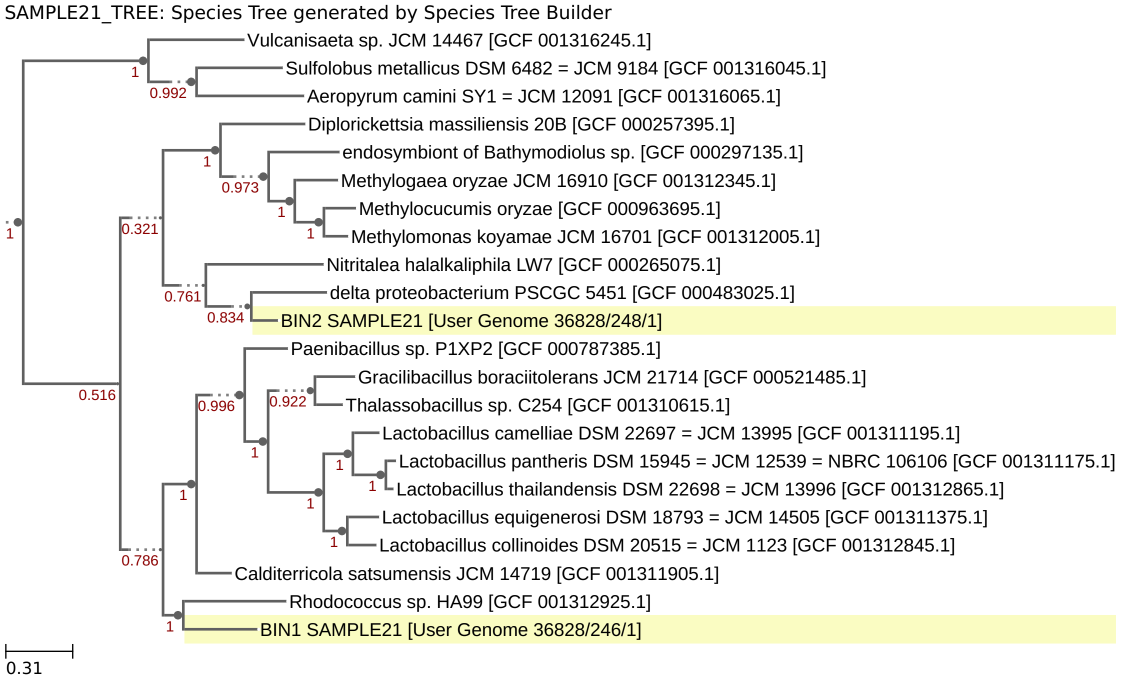
***A3 Figure 17:** Maximum likelihood Species tree for Sample 18, bins 1-8. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE19_TREE: Species Tree generated by Species Tree Builder
Vulcanisaeta sp. JCM 14467 [GCF 001316245.1]
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
Aeropyrum camini SY1 = JCM 12091 [GCF 001316065.1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
BIN1 SAMPLE19 [User Genome 36828/238/1]
Rhodococcus sp. HA99 [GCF 001312925.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
SAMPLE19 BIN2 [User Genome 36828/240/1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Sphingobacterium sp. T2 [GCF 000814685.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
endosymbiont of Bathymodiolus sp. [GCF 000297135.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
1
0.993
1
0.978
0.991
0.465
1
1
1
1
1
0.999
0.996
0.776
1
1
1
0.993
1
1
0.35

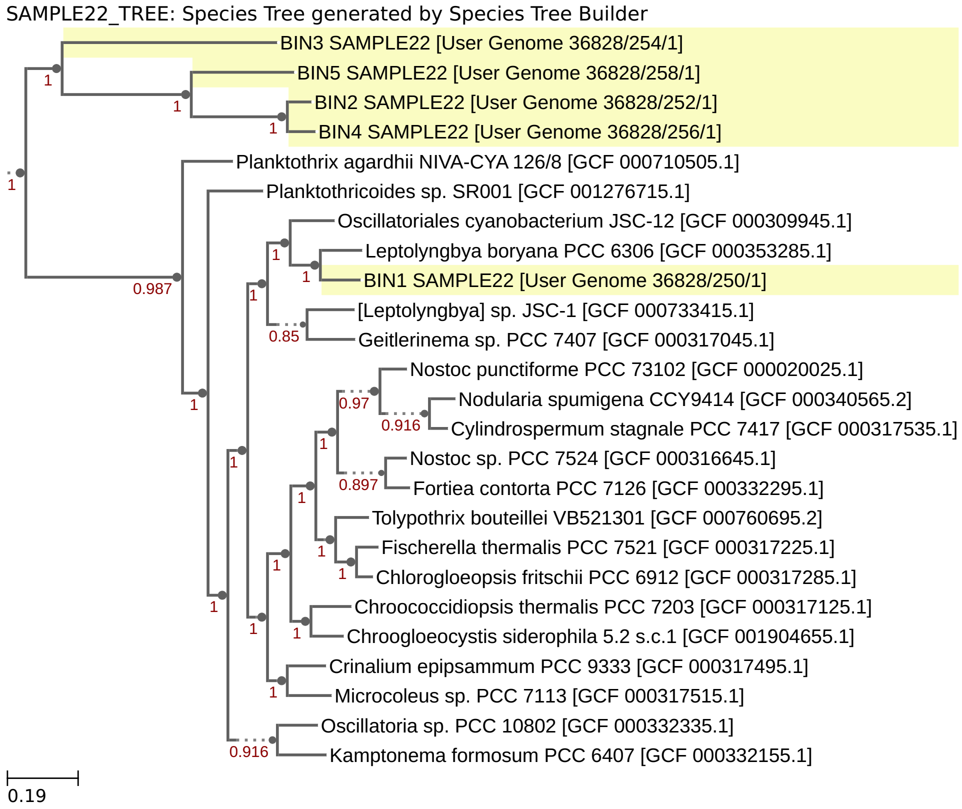
***A3 Figure 18:** Maximum likelihood Species tree for Sample 19, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE20_TREE: Species Tree generated by Species Tree Builder
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
Vulcanisaeta sp. JCM 14467 [GCF 001316245.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
endosymbiont of Bathymodiolus sp. [GCF 000297135.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
Sphingobacterium sp. T2 [GCF 000814685.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus similis DSM 23365 = JCM 2765 [GCF 001311075.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
BIN1 SAMPLE20 [User Genome 36828/242/1]
Rhodococcus sp. HA99 [GCF 001312925.1]
BIN2 SAMPLE20 [User Genome 36828/244/1]
1
1
0.989
1
1
1
1
0.846
1
0.81
1
0.811
1
1
1
1
0.806
0.829
0.996
0.639
0.28

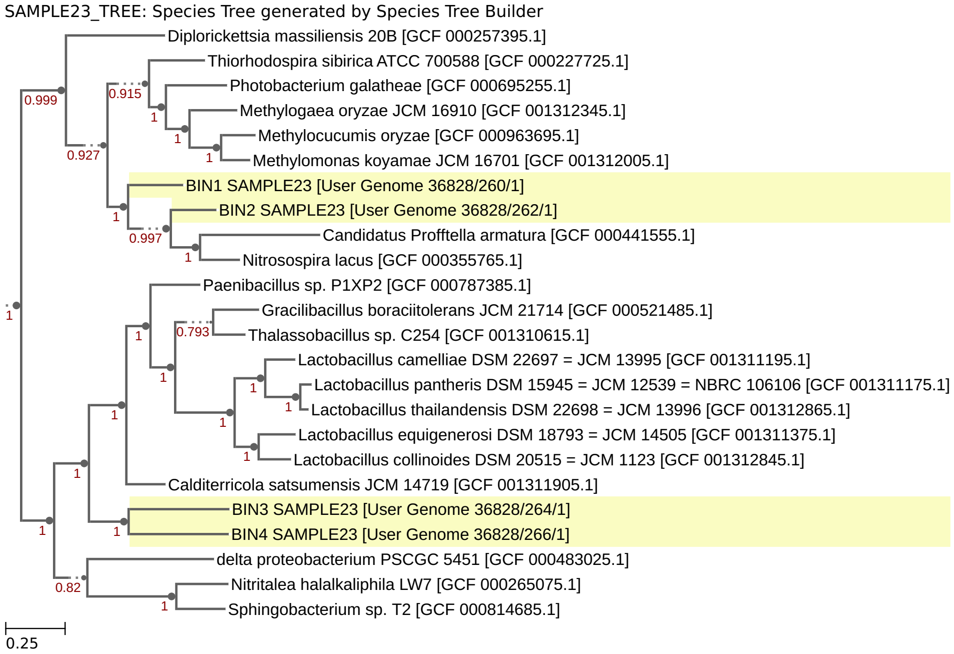
***A3 Figure 19:*** *Maximum likelihood Species tree for Sample 20, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE21_TREE: Species Tree generated by Species Tree Builder
Vulcanisaeta sp. JCM 14467 [GCF 001316245.1]
Sulfolobus metallicus DSM 6482 = JCM 9184 [GCF 001316045.1]
Aeropyrum camini SY1 = JCM 12091 [GCF 001316065.1]
Diplorickettsia massiliensis 20B [GCF 000257395.1]
endosymbiont of Bathymodiolus sp. [GCF 000297135.1]
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
Methylocucumis oryzae [GCF 000963695.1]
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
BIN2 SAMPLE21 [User Genome 36828/248/1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
Thalassobacillus sp. C254 [GCF 001310615.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
Rhodococcus sp. HA99 [GCF 001312925.1]
BIN1 SAMPLE21 [User Genome 36828/246/1]
1
0.992
1
0.973
1
1
1
0.321
0.761
0.834
0.516
0.996
0.922
1
1
1
1
1
1
0.786
1
0.31

***A3 Figure 20:*** *Maximum likelihood Species tree for Sample 21, bins 1-2. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE22_TREE: Species Tree generated by Species Tree Builder
BIN3 SAMPLE22 [User Genome 36828/254/1]
1
BIN5 SAMPLE22 [User Genome 36828/258/1]
1
BIN2 SAMPLE22 [User Genome 36828/252/1]
1
BIN4 SAMPLE22 [User Genome 36828/256/1]
1
Planktothrix agardhii NIVA-CYA 126/8 [GCF 000710505.1]
Planktothricoides sp. SR001 [GCF 001276715.1]
Oscillatoriales cyanobacterium JSC-12 [GCF 000309945.1]
1
Leptolyngbya boryana PCC 6306 [GCF 000353285.1]
1
BIN1 SAMPLE22 [User Genome 36828/250/1]
0.987
1
[Leptolyngbya] sp. JSC-1 [GCF 000733415.1]
0.85
Geitlerinema sp. PCC 7407 [GCF 000317045.1]
Nostoc punctiforme PCC 73102 [GCF 000020025.1]
1
0.97
Nodularia spumigena CCY9414 [GCF 000340565.2]
0.916
Cylindrospermum stagnale PCC 7417 [GCF 000317535.1]
1
Nostoc sp. PCC 7524 [GCF 000316645.1]
1
0.897
Fortiea contorta PCC 7126 [GCF 000332295.1]
1
Tolypothrix bouteillei VB521301 [GCF 000760695.2]
Fischerella thermalis PCC 7521 [GCF 000317225.1]
1
1
Chlorogloeopsis fritschii PCC 6912 [GCF 000317285.1]
1
Chroococcidiopsis thermalis PCC 7203 [GCF 000317125.1]
1
1
Chroogloeocystis siderophila 5.2 s.c.1 [GCF 001904655.1]
Crinalium epipsammum PCC 9333 [GCF 000317495.1]
1
Microcoleus sp. PCC 7113 [GCF 000317515.1]
Oscillatoria sp. PCC 10802 [GCF 000332335.1]
0.916
Kamptonema formosum PCC 6407 [GCF 000332155.1]
0.19

***A3 Figure 21:*** *Maximum likelihood Species tree for Sample 22, bins 1-5. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*


SAMPLE23_TREE: Species Tree generated by Species Tree Builder
Diplorickettsia massiliensis 20B [GCF 000257395.1]
Thiorhodospira sibirica ATCC 700588 [GCF 000227725.1]
Photobacterium galatheae [GCF 000695255.1]
0.999
0.915
1
Methylogaea oryzae JCM 16910 [GCF 001312345.1]
1
Methylocucumis oryzae [GCF 000963695.1]
1
0.927
Methylomonas koyamae JCM 16701 [GCF 001312005.1]
BIN1 SAMPLE23 [User Genome 36828/260/1]
1
BIN2 SAMPLE23 [User Genome 36828/262/1]
0.997
Candidatus Profftella armatura [GCF 000441555.1]
1
Nitrosospira lacus [GCF 000355765.1]
Paenibacillus sp. P1XP2 [GCF 000787385.1]
1
Gracilibacillus boraciitolerans JCM 21714 [GCF 000521485.1]
1
0.793
Thalassobacillus sp. C254 [GCF 001310615.1]
Lactobacillus camelliae DSM 22697 = JCM 13995 [GCF 001311195.1]
1
1
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106 [GCF 001311175.1]
1
Lactobacillus thailandensis DSM 22698 = JCM 13996 [GCF 001312865.1]
1
1
Lactobacillus equigenerosi DSM 18793 = JCM 14505 [GCF 001311375.1]
1
Lactobacillus collinoides DSM 20515 = JCM 1123 [GCF 001312845.1]
1
Calditerricola satsumensis JCM 14719 [GCF 001311905.1]
BIN3 SAMPLE23 [User Genome 36828/264/1]
1
1
BIN4 SAMPLE23 [User Genome 36828/266/1]
delta proteobacterium PSCGC 5451 [GCF 000483025.1]
0.82
Nitritalea halalkaliphila LW7 [GCF 000265075.1]
1
Sphingobacterium sp. T2 [GCF 000814685.1]
0.25

***A3 Figure 22:*** *Maximum likelihood Species tree for Sample 23, bins 1-4. The species tree is based off alignments of COG groups to publicly available genomes on the KBASE platform.*

# Appendix 4

***A4 Table 1:*** *Kaiju taxonomic classification of metagenome reads for Sample 1. The percentage of the metagenome classified and the taxonomic assignment is given, from the class to species level. The percentage of the reads that were unclassified is also provided.*

| % | class | % | family | % | genus | % | order | % | phylum | % | species |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20.994911 | Gammaproteobacteria | 3.931496 | Flavobacteriaceae | 2.959363 | Nitrosopumilus | 6.542552 | Rhizobiales | 54.67688 | Proteobacteria | 2.706551 | Woeseia oceani |
| 16.85726 | Alphaproteobacteria | 3.080831 | Rhodobacteraceae | 2.706551 | Woeseia | 5.263172 | Chromatiales | 10.4805 | Bacteroidetes | 1.679244 | Candidatus Nitrosopumilus sediminis |
| 8.074667 | Deltaproteobacteria | 2.959363 | Nitrosopumilaceae | 1.47278 | Pseudomonas | 4.648056 | Burkholderiales | 5.654521 | Actinobacteria | 1.042004 | Candidatus Nitrosopumilus adriaticus |
| 7.43486 | Betaproteobacteria | 2.706551 | Woeseiaceae | 0.867306 | Halioglobus | 4.169882 | Flavobacteriales | 5.388607 | Firmicutes | 0.65466 | Candidatus Solibacter usitatus |
| 4.995546 | Actinobacteria | 2.381379 | Planctomycetaceae | 0.814504 | Geobacter | 3.443874 | Rhodobacterales | 3.769552 | Planctomycetes | 0.583466 | Fuerstia marisgermanicae |
| 4.169882 | Flavobacteriia | 2.323833 | Rhodospirillaceae | 0.773717 | Bradyrhizobium | 3.386523 | Planctomycetales | 3.233599 | Thaumarchaeota | ------------ | ------------ |
| 3.386523 | Planctomycetia | 1.711983 | Burkholderiaceae | 0.708511 | Streptomyces | 3.02729 | Myxococcales | 2.002861 | Cyanobacteria | 1.760236 | Viruses |
| 2.806206 | Cytophagia | 1.597047 | Pseudomonadaceae | 0.677328 | Burkholderia | 2.959363 | Nitrosopumilales | 1.480751 | Acidobacteria | 16.12795 | cannot be assigned to a species |
| 2.66413 | Clostridia | 1.545995 | Sphingomonadaceae | 0.65466 | Candidatus Solibacter | 2.901428 | Rhodospirillales | 1.462165 | Euryarchaeota | 75.44589 | species < 0.5% of all reads |
| 2.281646 | Bacilli | 1.423672 | Bradyrhizobiaceae | 0.65081 | Mycobacterium | 2.806206 | Cytophagales | 1.365232 | Chloroflexi | ------------ | ------------ |
| 1.444824 | Bacteroidia | 1.192517 | Hyphomicrobiaceae | 0.583466 | Fuerstia | 2.685127 | Alteromonadales | 1.046709 | Verrucomicrobia | 70.54341 | unclassified |
| 0.832079 | Spirochaetia | 1.190262 | Comamonadaceae | 0.550183 | Azospirillum | 2.107571 | Sphingomonadales | 0.832079 | Spirochaetes | | |
| 0.709328 | Opitutae | 1.147958 | Rhizobiaceae | 0.528953 | Desulfovibrio | 2.095362 | Cellvibrionales | 0.675267 | Nitrospirae | | |
| 0.675267 | Nitrospira | 1.126923 | Ectothiorhodospiraceae | 0.51725 | Nitrospira | 1.984431 | Clostridiales | 0.549639 | Deinococcus-Thermus | | |
| 0.665197 | Acidobacteriia | 1.048925 | Chromatiaceae | 0.513984 | Sphingomonas | 1.908261 | Pseudomonadales | ------------ | ------------ | | |
| 0.65466 | Solibacteres | 1.02649 | Alteromonadaceae | 0.511768 | Paenibacillus | 1.857753 | Bacillales | 1.760236 | Viruses | | |
| 0.567369 | Methanomicrobia | 0.993518 | Desulfobacteraceae | ------------ | ------------ | 1.856936 | Desulfuromonadales | 1.761208 | cannot be assigned to a phylum | | |
| 0.549639 | Deinococci | 0.969528 | Geobacteraceae | 1.760236 | Viruses | 1.721043 | Oceanospirillales | 3.860186 | phylum < 0.5% of all reads | | |
| 0.536263 | Sphingobacteriia | 0.867306 | Halieaceae | 9.951588 | cannot be assigned to a genus | 1.47764 | Desulfobacterales | ------------ | ------------ | | |
| ------------ | ------------ | 0.863535 | Phyllobacteriaceae | 72.79704 | genus < 0.5% of all reads | 1.352556 | Corynebacteriales | 70.54341 | unclassified | | |
| 1.760236 | Viruses | 0.849304 | Desulfuromonadaceae | ------------ | ------------ | 1.150097 | Xanthomonadales | | | | |
| 8.730026 | cannot be assigned to a class | 0.840789 | Cytophagaceae | 70.54341 | unclassified | 0.99414 | Enterobacterales | | | | |
| 9.209482 | class < 0.5% of all reads | 0.817693 | Rhodocyclaceae | | | 0.918553 | Thiotrichales | | | | |
| ------------ | ------------ | 0.806222 | Polyangiaceae | | | 0.889664 | Bacteroidales | | | | |
| 70.543408 | unclassified | 0.754081 | Bacillaceae | | | 0.817693 | Rhodocyclales | | | | |
| | | 0.73258 | Streptomycetaceae | | | 0.757347 | Micrococcales | | | | |
| | | 0.675267 | Nitrospiraceae | | | 0.73258 | Streptomycetales | | | | |
| | | 0.672895 | Mycobacteriaceae | | | 0.715394 | Desulfovibrionales | | | | |
| | | 0.670446 | Isosphaeraceae | | | 0.675267 | Nitrospirales | | | | |
| | | 0.665197 | Acidobacteriaceae | | | 0.668696 | Synechococcales | | | | |
| | | 0.65466 | Solibacteraceae | | | 0.665197 | Acidobacteriales | | | | |
| | | 0.653493 | Xanthomonadaceae | | | 0.65466 | Solibacterales | | | | |
| | | 0.643539 | Alcaligenaceae | | | 0.641012 | Vibrionales | | | | |
| | | 0.641012 | Vibrionaceae | | | 0.614028 | Bacteroidetes Order II. Incertae sedis | | | | |
| | | 0.626198 | Paenibacillaceae | | | 0.580472 | Methylococcales | | | | |
| | | 0.614028 | Rhodothermaceae | | | 0.56943 | Spirochaetales | | | | |
| | | 0.610528 | Hymenobacteraceae | | | 0.560876 | Syntrophobacterales | | | | |
| | | 0.588249 | Cellvibrionaceae | | | 0.560798 | Caulobacterales | | | | |
| | | 0.580472 | Methylococcaceae | | | 0.548433 | Marinilabiliales | | | | |
| | | 0.573279 | Peptococcaceae | | | 0.536263 | Sphingobacteriales | | | | |
| | | 0.567913 | Acetobacteraceae | | | 0.515811 | Opitutales | | | | |
| | | 0.561148 | Clostridiaceae | | | ------------ | ------------ | | | | |
| | | 0.560798 | Caulobacteraceae | | | 1.760236 | Viruses | | | | |
| | | 0.558621 | Cyclobacteriaceae | | | 7.556406 | cannot be assigned to a order | | | | |
| | | 0.556015 | Oxalobacteraceae | | | 17.72192 | order < 0.5% of all reads | | | | |
| | | 0.549289 | Spirochaetaceae | | | ------------ | ------------ | | | | |
| | | 0.546723 | Desulfovibrionaceae | | | 70.54341 | unclassified | | | | |
| | | 0.536263 | Sphingobacteriaceae | | | | | | | | |
| | | 0.523082 | Piscirickettsiaceae | | | | | | | | |
| | | 0.515811 | Opitutaceae | | | | | | | | |
| | | 0.510173 | Erythrobacteraceae | | | | | | | | |
| | | 0.503058 | Oceanospirillaceae | | | | | | | | |
| | | 0.503019 | Colwelliaceae | | | | | | | | |
| | | ------------ | ------------ | | | | | | | | |
| | | 1.760236 | Viruses | | | | | | | | |
| | | 9.324418 | cannot be assigned to a family | | | | | | | | |
| | | 34.16439 | family < 0.5% of all reads | | | | | | | | |
| | | ------------ | ------------ | | | | | | | | |
| | | 70.543408 | unclassified | | | | | | | | |

**A4 Table 2:** *Kaiju taxonomic classification of metagenome reads for Sample 2. The percentage of the metagenome classified and the taxonomic assignment is given, from the class to species level. The percentage of the reads that were unclassified is also provided.*

| % | class | % | family | % | genus | % | order | % | phylum | % | species |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 17.56398 | Gammaproteobacteria | 3.544211 | Flavobacteriaceae | 1.794189 | Woeseia | 5.42386 | Rhizobiales | 49.03289 | Proteobacteria | 1.794189 | Woeseia oceani |
| 13.95572 | Alphaproteobacteria | 2.242425 | Rhodobacteraceae | 1.617759 | Nitrosopumilus | 4.139128 | Burkholderiales | 11.09193 | Bacteroidetes | 0.88622 | Candidatus Nitrosopumilus sediminis |
| 9.508138 | Deltaproteobacteria | 2.161755 | Planctomycetaceae | 1.289043 | Pseudomonas | 4.097165 | Chromatiales | 6.818052 | Firmicutes | 0.650965 | Candidatus Solibacter usitatus |
| 6.7027 | Betaproteobacteria | 2.157875 | Rhodospirillaceae | 0.997309 | Geobacter | 3.795227 | Flavobacteriales | 5.515596 | Actinobacteria | 0.588354 | Candidatus Nitrosopumilus adriaticus |
| 4.878619 | Actinobacteria | 1.794189 | Woeseiaceae | 0.704138 | Desulfococcus | 3.189243 | Planctomycetales | 3.779418 | Planctomycetes | 0.503038 | Draconibacterium orientale |
| 3.795227 | Flavobacteriia | 1.668777 | Desulfobacteraceae | 0.698294 | Streptomyces | 3.069819 | Cytophagales | 2.16597 | Cyanobacteria | ------------------------------------------- | |
| 3.612905 | Clostridia | 1.617759 | Nitrosopumilaceae | 0.68397 | Desulfovibrio | 2.847546 | Myxococcales | 2.136989 | Euryarchaeota | 3.792065 | Viruses |
| 3.189243 | Planctomycetia | 1.514 | Burkholderiaceae | 0.650965 | Candidatus Solibacter | 2.677966 | Rhodospirillales | 1.837398 | Thaumarchaeota | 15.79 | cannot be assigned to a species |
| 3.069819 | Cytophagia | 1.398312 | Pseudomonadaceae | 0.638749 | Bradyrhizobium | 2.645679 | Clostridiales | 1.707291 | Chloroflexi | 75.99517 | species < 0.5% of all reads |
| 2.644529 | Bacilli | 1.276588 | Sphingomonadaceae | 0.618438 | Paenibacillus | 2.528986 | Rhodobacterales | 1.48612 | Acidobacteria | ------------------------------------------- | |
| 2.013828 | Bacteroidia | 1.202146 | Bradyrhizobiaceae | 0.616474 | Burkholderia | 2.295358 | Alteromonadales | 1.131536 | Verrucomicrobia | 74.19126 | unclassified |
| 1.018722 | Spirochaetia | 1.15611 | Geobacteraceae | 0.603923 | Mycobacterium | 2.252868 | Desulfobacterales | 1.018722 | Spirochaetes | | |
| 0.831035 | Methanomicrobia | 1.058147 | Comamonadaceae | 0.602821 | Clostridium | 2.153515 | Bacillales | 0.721048 | Nitrospirae | | |
| 0.777862 | Opitutae | 0.991178 | Ectothiorhodospiraceae | 0.571013 | Halioglobus | 2.081037 | Desulfuromonadales | 0.624857 | Ignavibacteriae | | |
| 0.721048 | Nitrospira | 0.975513 | Chromatiaceae | 0.529433 | Nitrospira | 1.690813 | Pseudomonadales | 0.594103 | Chlorobi | | |
| 0.659587 | Acidobacteriia | 0.960998 | Rhizobiaceae | 0.524259 | Bacillus | 1.680369 | Sphingomonadales | 0.585672 | Deinococcus-Thermus | | |
| 0.650965 | Solibacteres | 0.960854 | Hyphomicrobiaceae | 0.503038 | Draconibacterium | 1.617759 | Nitrosopumilales | ------------------------------------------- | | | |
| 0.624857 | Ignavibacteria | 0.914771 | Cytophagaceae | ------------------------------------------- | | 1.537999 | Cellvibrionales | 3.792065 | Viruses | | |
| 0.605983 | Sphingobacteriia | 0.880184 | Desulfuromonadaceae | 3.792065 | Viruses | 1.440707 | Oceanospirillales | 2.086211 | cannot be assigned to a phylum | | |
| 0.594103 | Chlorobia | 0.866484 | Bacillaceae | 10.13678 | cannot be assigned to a genus | 1.273044 | Corynebacteriales | 3.874124 | phylum < 0.5% of all reads | | |
| 0.586295 | Phycisphaerae | 0.856999 | Alteromonadaceae | 72.42734 | genus < 0.5% of all reads | 1.246457 | Bacteroidales | ------------------------------------------- | | | |
| 0.585672 | Deinococci | 0.821311 | Rhodocyclaceae | ------------------------------------------- | | 1.012207 | Xanthomonadales | 74.19126 | unclassified | | |
| 0.500355 | Epsilonproteobacteria | 0.799467 | Peptococcaceae | 74.19126 | unclassified | 0.945429 | Enterobacterales | | | | |
| ------------------------------------------- | | 0.75391 | Clostridiaceae | | | 0.926795 | Desulfovibrionales | | | | |
| 3.792065 | Viruses | 0.752569 | Paenibacillaceae | | | 0.860065 | Syntrophobacterales | | | | |
| 7.8979 | cannot be assigned to a class | 0.740497 | Polyangiaceae | | | 0.821311 | Rhodocyclales | | | | |
| 9.218847 | class < 0.5% of all reads | 0.723012 | Streptomycetaceae | | | 0.815993 | Thiotrichales | | | | |
| ------------------------------------------- | | 0.721048 | Nitrospiraceae | | | 0.777766 | Micrococcales | | | | |
| 74.19126 | unclassified | 0.706198 | Desulfovibrionaceae | | | 0.757647 | Marinilabiliales | | | | |
| | | 0.704569 | Spirochaetaceae | | | 0.730677 | Spirochaetales | | | | |
| | | 0.682869 | Isosphaeraceae | | | 0.723012 | Streptomycetales | | | | |
| | | 0.682581 | Phyllobacteriaceae | | | 0.721048 | Nitrospirales | | | | |
| | | 0.660066 | Rhodothermaceae | | | 0.711515 | Synechococcales | | | | |
| | | 0.659587 | Acidobacteriaceae | | | 0.683683 | Thermoanaerobacterales | | | | |
| | | 0.650965 | Solibacteraceae | | | 0.660066 | Bacteroidetes Order II. Incertae sedis | | | | |
| | | 0.649192 | Hymenobacteraceae | | | 0.659587 | Acidobacteriales | | | | |
| | | 0.626007 | Mycobacteriaceae | | | 0.650965 | Solibacterales | | | | |
| | | 0.616857 | Cyclobacteriaceae | | | 0.624857 | Ignavibacteriales | | | | |
| | | 0.611013 | Syntrophaceae | | | 0.605983 | Sphingobacteriales | | | | |
| | | 0.605983 | Sphingobacteriaceae | | | 0.594103 | Chlorobiales | | | | |
| | | 0.594103 | Chlorobiaceae | | | 0.589887 | Opitutales | | | | |
| | | 0.589887 | Opitutaceae | | | 0.574702 | Vibrionales | | | | |
| | | 0.583708 | Xanthomonadaceae | | | 0.545241 | Oscillatoriales | | | | |
| | | 0.57954 | Desulfobulbaceae | | | 0.537624 | Methanosarcinales | | | | |
| | | 0.574702 | Vibrionaceae | | | 0.506295 | Methylococcales | | | | |
| | | 0.571013 | Halieaceae | | | ------------------------------------------- | | | | | |
| | | 0.567564 | Alcaligenaceae | | | 3.792065 | Viruses | | | | |
| | | 0.543277 | Porphyromonadaceae | | | 7.826906 | cannot be assigned to a order | | | | |
| | | 0.538535 | Flammeovirgaceae | | | 17.66002 | order < 0.5% of all reads | | | | |
| | | 0.51099 | Acetobacteraceae | | | ------------------------------------------- | | | | | |
| | | 0.506295 | Methylococcaceae | | | 74.19126 | unclassified | | | | |
| | | 0.503038 | Prolixibacteraceae | | | | | | | | |
| | | ------------------------------------------- | | | | | | | | | |
| | | 3.792065 | Viruses | | | | | | | | |
| | | 9.568976 | cannot be assigned to a family | | | | | | | | |
| | | 36.60933 | family < 0.5% of all reads | | | | | | | | |
| | | ------------------------------------------- | | | | | | | | | |
| | | 74.19126 | unclassified | | | | | | | | |

**A4 Table 3:** *Kaiju taxonomic classification of metagenome reads for Sample 4. The percentage of the metagenome classified and the taxonomic assignment is given, from the class to species level. The percentage of the reads that were unclassified is also provided.*

| % | class | % | family | % | genus | % | order | % | phylum | % | species |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 18.892527 | Gammaproteobacteria | 2.975519 | Flavobacteriaceae | 2.269193 | Nitrosopumilus | 5.067026 | Rhizobiales | 48.930908 | Proteobacteria | 2.031046 | Woeseia oceani |
| 13.286392 | Alphaproteobacteria | 2.412206 | Planctomycetaceae | 2.031046 | Woeseia | 4.496702 | Chromatiales | 9.786172 | Bacteroidetes | 1.288511 | Candidatus Nitrosopumilus sediminis |
| 8.923099 | Deltaproteobacteria | 2.269193 | Nitrosopumilaceae | 1.350077 | Pseudomonas | 4.09883 | Burkholderiales | 6.764503 | Firmicutes | 0.788541 | Candidatus Nitrosopumilus adriaticus |
| 6.541338 | Betaproteobacteria | 2.104234 | Rhodobacteraceae | 0.933187 | Geobacter | 3.522599 | Planctomycetales | 5.220268 | Actinobacteria | 0.610999 | Candidatus Solibacter usitatus |
| 4.619978 | Actinobacteria | 2.046846 | Rhodospirillaceae | 0.683226 | Desulfococcus | 3.204253 | Flavobacteriales | 4.196221 | Planctomycetes | 0.570612 | Fuerstia marisgermanicae |
| 3.605439 | Clostridia | 2.031046 | Woeseiaceae | 0.675927 | Desulfovibrio | 2.931481 | Cytophagales | 2.524197 | Thaumarchaeota |  | ------------------------------------------- |
| 3.522599 | Planctomycetia | 1.59999 | Desulfobacteraceae | 0.662624 | Streptomyces | 2.669419 | Myxococcales | 2.344205 | Euryarchaeota | 4.813991 | Viruses |
| 3.204253 | Flavobacteriia | 1.484926 | Burkholderiaceae | 0.657438 | Halioglobus | 2.624997 | Clostridiales | 2.127333 | Cyanobacteria | 15.673867 | cannot be assigned to a species |
| 2.931481 | Cytophagia | 1.462739 | Pseudomonadaceae | 0.617771 | Mycobacterium | 2.538748 | Rhodospirillales | 1.50063 | Chloroflexi | 74.222434 | species < 0.5% of all reads |
| 2.611551 | Bacilli | 1.290192 | Sphingomonadaceae | 0.615562 | Bradyrhizobium | 2.458789 | Alteromonadales | 1.403286 | Acidobacteria |  | ------------------------------------------- |
| 1.587119 | Bacteroidia | 1.140792 | Bradyrhizobiaceae | 0.610999 | Candidatus Solibacter | 2.391989 | Rhodobacterales | 1.068997 | Verrucomicrobia | 74.617857 | unclassified |
| 0.923823 | Spirochaetia | 1.063186 | Chromatiaceae | 0.605909 | Clostridium | 2.269193 | Nitrosopumilales | 0.923823 | Spirochaetes |  |  |
| 0.88161 | Methanomicrobia | 1.060161 | Geobacteraceae | 0.602595 | Paenibacillus | 2.138282 | Desulfobacterales | 0.745561 | Nitrospirae |  |  |
| 0.745561 | Nitrospira | 1.05776 | Comamonadaceae | 0.592943 | Burkholderia | 2.115087 | Bacillales | 0.57599 | Chlorobi |  |  |
| 0.721117 | Opitutae | 1.056559 | Ectothiorhodospiraceae | 0.570612 | Fuerstia | 1.812252 | Desulfuromonadales | 0.561007 | Deinococcus-Thermus |  |  |
| 0.668915 | Phycisphaerae | 0.918396 | Rhizobiaceae | 0.54612 | Nitrospira | 1.768263 | Cellvibrionales | 0.548521 | Ignavibacteriae |  |  |
| 0.634579 | Acidobacteriia | 0.905478 | Alteromonadaceae | 0.515193 | Planctomyces | 1.766054 | Pseudomonadales |  | ------------------------------------------- |  |  |
| 0.610999 | Solibacteres | 0.875896 | Hyphomicrobiaceae | 0.514809 | Bacillus | 1.692194 | Sphingomonadales | 4.813991 | Viruses |  |  |
| 0.57599 | Chlorobia | 0.87239 | Cytophagaceae |  | ------------------------------------------- | 1.557153 | Oceanospirillales | 2.063606 | cannot be assigned to a phylum |  |  |
| 0.561007 | Deinococci | 0.852124 | Bacillaceae | 4.813991 | Viruses | 1.268245 | Corynebacteriales | 3.900782 | phylum < 0.5% of all reads |  |  |
| 0.558318 | Sphingobacteriia | 0.809287 | Rhodocyclaceae | 9.985708 | cannot be assigned to a genus | 1.058864 | Bacteroidales |  | ------------------------------------------- |  |  |
| 0.548521 | Ignavibacteria | 0.787917 | Peptococcaceae | 70.145071 | genus < 0.5% of all reads | 1.021118 | Xanthomonadales | 74.617857 | unclassified |  |  |
|  | ------------------------------------------- | 0.753676 | Clostridiaceae |  | ------------------------------------------- | 0.98049 | Enterobacterales |  |  |  |  |
| 4.813991 | Viruses | 0.745561 | Nitrospiraceae | 74.617857 | unclassified | 0.913402 | Desulfovibrionales |  |  |  |  |
| 8.470047 | cannot be assigned to a class | 0.742535 | Paenibacillaceae |  |  | 0.864802 | Thiotrichales |  |  |  |  |
| 9.559742 | class < 0.5% of all reads | 0.733507 | Isosphaeraceae |  |  | 0.843336 | Syntrophobacterales |  |  |  |  |
|  | ------------------------------------------- | 0.715738 | Desulfuromonadaceae |  |  | 0.809287 | Rhodocyclales |  |  |  |  |
| 74.617857 | unclassified | 0.699266 | Desulfovibrionaceae |  |  | 0.745561 | Nitrospirales |  |  |  |  |
|  |  | 0.685243 | Streptomycetaceae |  |  | 0.713865 | Micrococcales |  |  |  |  |
|  |  | 0.676647 | Polyangiaceae |  |  | 0.709879 | Synechococcales |  |  |  |  |
|  |  | 0.657438 | Halieaceae |  |  | 0.694416 | Thermoanaerobacterales |  |  |  |  |
|  |  | 0.639285 | Mycobacteriaceae |  |  | 0.685243 | Streptomycetales |  |  |  |  |
|  |  | 0.634579 | Acidobacteriaceae |  |  | 0.656718 | Spirochaetales |  |  |  |  |
|  |  | 0.633426 | Spirochaetaceae |  |  | 0.634579 | Acidobacteriales |  |  |  |  |
|  |  | 0.632034 | Rhodothermaceae |  |  | 0.632034 | Bacteroidetes Order II. Incertae sedis |  |  |  |  |
|  |  | 0.627615 | Phyllobacteriaceae |  |  | 0.610999 | Solibacterales |  |  |  |  |
|  |  | 0.623966 | Hymenobacteraceae |  |  | 0.597793 | Vibrionales |  |  |  |  |
|  |  | 0.610999 | Solibacteraceae |  |  | 0.57599 | Chlorobiales |  |  |  |  |
|  |  | 0.597793 | Vibrionaceae |  |  | 0.573637 | Methanosarcinales |  |  |  |  |
|  |  | 0.596208 | Syntrophaceae |  |  | 0.558318 | Sphingobacteriales |  |  |  |  |
|  |  | 0.592559 | Xanthomonadaceae |  |  | 0.548521 | Ignavibacteriales |  |  |  |  |
|  |  | 0.590157 | Cyclobacteriaceae |  |  | 0.547609 | Opitutales |  |  |  |  |
|  |  | 0.57599 | Chlorobiaceae |  |  | 0.544871 | Methylococcales |  |  |  |  |
|  |  | 0.564753 | Alcaligenaceae |  |  | 0.53032 | Oscillatoriales |  |  |  |  |
|  |  | 0.558318 | Sphingobacteriaceae |  |  | 0.521868 | Marinilabiliales |  |  |  |  |
|  |  | 0.547609 | Opitutaceae |  |  |  | ------------------------------------------- |  |  |  |  |
|  |  | 0.544871 | Methylococcaceae |  |  | 4.813991 | Viruses |  |  |  |  |
|  |  | 0.534354 | Desulfobulbaceae |  |  | 7.878882 | cannot be assigned to a order |  |  |  |  |
|  |  | 0.527247 | Cellvibrionaceae |  |  | 17.342049 | order < 0.5% of all reads |  |  |  |  |
|  |  | 0.51793 | Flammeovirgaceae |  |  |  | ------------------------------------------- |  |  |  |  |
|  |  |  | ------------------------------------------- |  |  | 74.617857 | unclassified |  |  |  |  |
|  |  | 4.813991 | Viruses |  |  |  |  |  |  |  |  |
|  |  | 9.603827 | cannot be assigned to a family |  |  |  |  |  |  |  |  |
|  |  | 36.94604 | family < 0.5% of all reads |  |  |  |  |  |  |  |  |
|  |  |  | ------------------------------------------- |  |  |  |  |  |  |  |  |
|  |  | 74.617857 | unclassified |  |  |  |  |  |  |  |  |

**A4 Table 4:** *Kaiju taxonomic classification of metagenome reads for Sample 5. The percentage of the metagenome classified and the taxonomic assignment is given, from the class to species level. The percentage of the reads that were unclassified is also provided.*

| % | class | % | family | % | genus | % | order | % | phylum | % | species |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 16.6543 | Deltaproteobacteria | 4.770622 | Flavobacteriaceae | 1.801326 | Desulfococcus | 6.509687 | Desulfobacterales | 46.16938 | Proteobacteria | 1.48241 | Woeseia oceani |
| 13.86351 | Gammaproteobacteria | 4.254628 | Desulfobacteraceae | 1.553713 | Geobacter | 5.056057 | Flavobacteriales | 14.89442 | Bacteroidetes | 1.149484 | Desulfococcus multivorans |
| 9.637166 | Alphaproteobacteria | 2.505626 | Planctomycetaceae | 1.48241 | Woeseia | 4.615395 | Rhizobiales | 7.700351 | Firmicutes | 0.938441 | Draconibacterium orientale |
| 5.348028 | Actinobacteria | 2.243466 | Desulfobulbaceae | 1.227501 | Desulfuromonas | 4.18798 | Desulfuromonadales | 5.988232 | Actinobacteria | 0.902722 | Desulfobacula toluolica |
| 5.056057 | Flavobacteriia | 2.145844 | Desulfuromonadaceae | 1.137757 | Mycobacterium | 3.756448 | Cytophagales | 4.341194 | Planctomycetes | 0.813023 | Desulfatibacillum alkenivorans |
| 4.843671 | Betaproteobacteria | 1.957942 | Geobacteraceae | 1.050296 | Desulfovibrio | 3.688278 | Planctomycetales | 2.093251 | Cyanobacteria | 0.795835 | Desulfocapsa sulfexigens |
| 4.305386 | Clostridia | 1.48241 | Woeseiaceae | 0.979127 | Pseudomonas | 3.182086 | Clostridiales | 2.092625 | Euryarchaeota | 0.758908 | Candidatus Solibacter usitatus |
| 3.756448 | Cytophagia | 1.419119 | Bradyrhizobiaceae | 0.938441 | Draconibacterium | 3.151425 | Chromatiales | 1.977815 | Chloroflexi | 0.680219 | Desulfobacterium autotrophicum |
| 3.688278 | Planctomycetia | 1.386042 | Rhodobacteraceae | 0.902722 | Desulfobacula | 3.084599 | Burkholderiales | 1.64283 | Acidobacteria | 0.64374 | Desulfococcus oleovorans |
| 3.549253 | Bacteroidia | 1.156825 | Mycobacteriaceae | 0.892248 | Pelobacter | 2.467983 | Myxococcales | 1.261161 | Spirochaetes | 0.623016 | Desulfuromonas sp. DDH964 |
| 2.795761 | Bacilli | 1.142412 | Burkholderiaceae | 0.86387 | Bradyrhizobium | 2.289435 | Bacillales | 1.195856 | Verrucomicrobia | 0.599562 | Anaerolinea thermophila |
| 1.261161 | Spirochaetia | 1.111976 | Cytophagaceae | 0.813023 | Desulfatibacillum | 2.084658 | Bacteroidales | 0.825332 | Ignavibacteriae | 0.578435 | Desulfuromonas soudanensis |
| 0.910734 | Methanomicrobia | 1.077286 | Desulfovibrionaceae | 0.795835 | Desulfocapsa | 1.953958 | Alteromonadales | 0.662942 | Chlorobi | 0.568319 | Desulfobulbus propionicus |
| 0.829002 | Opitutae | 1.058263 | Pseudomonadaceae | 0.758908 | Candidatus Solibacter | 1.787226 | Corynebacteriales | 0.659764 | Nitrospirae | 0.555294 | Caldithrix abyssi |
| 0.825332 | Ignavibacteria | 1.007282 | Peptococcaceae | 0.72113 | Clostridium | 1.571662 | Rhodobacterales | 0.555965 | Deinococcus-Thermus | 0.547774 | Fuerstia marisgermanicae |
| 0.775648 | Sphingobacteriia | 0.975547 | Spirochaetaceae | 0.698482 | Lutibacter | 1.444588 | Marinilabiliales | 0.555294 | Calditrichaeota | 0.510937 | Lutibacter profundi |
| 0.758908 | Solibacteres | 0.959164 | Syntrophaceae | 0.69029 | Paenibacillus | 1.441544 | Desulfovibrionales | ------------------------------------------ | | 0.503775 | Salinivirga cyanobacteriivorans |
| 0.723189 | Acidobacteriia | 0.938441 | Prolixibacteraceae | 0.687739 | Streptomyces | 1.399917 | Syntrophobacterales | 1.333493 | Viruses | ------------------------------------------ | |
| 0.695885 | Anaerolineae | 0.934367 | Porphyromonadaceae | 0.680219 | Desulfobacterium | 1.286047 | Pseudomonadales | 2.065008 | cannot be assigned to a phylum | 1.333493 | Viruses |
| 0.662942 | Chlorobia | 0.928325 | Bacillaceae | 0.599562 | Anaerolinea | 1.277543 | Cellvibrionales | 3.985082 | phylum < 0.5% of all reads | 14.48625 | cannot be assigned to a species |
| 0.659764 | Nitrospira | 0.920716 | Rhodospirillaceae | 0.568319 | Desulfobulbus | 1.240616 | Rhodospirillales | ------------------------------------------ | | 71.52836 | species < 0.5% of all reads |
| 0.646963 | Phycisphaerae | 0.900305 | Clostridiaceae | 0.566081 | Bacillus | 1.18113 | Sphingomonadales | 71.34026 | unclassified | ------------------------------------------ | |
| 0.597368 | Chitinophagia | 0.892696 | Sphingomonadaceae | 0.555294 | Caldithrix | 1.090491 | Oceanospirillales | | | 71.34026 | unclassified |
| 0.555965 | Deinococci | 0.855679 | Hyphomicrobiaceae | 0.547774 | Fuerstia | 1.003656 | Spirochaetales | | | | |
| 0.555294 | Calditrichae | 0.83321 | Paenibacillaceae | 0.533764 | Planctomyces | 0.882311 | Xanthomonadales | | | | |
| 0.523693 | Epsilonproteobacteria | 0.785048 | Cyclobacteriaceae | 0.503775 | Salinivirga | 0.848562 | Enterobacterales | | | | |
| ------------------------------------------ | | 0.779184 | Isosphaeraceae | ------------------------------------------ | | 0.825332 | Ignavibacteriales | | | | |
| 1.333493 | Viruses | 0.775648 | Sphingobacteriaceae | 1.333493 | Viruses | 0.816424 | Micrococcales | | | | |
| 6.275502 | cannot be assigned to a class | 0.771843 | Hymenobacteraceae | 9.46748 | cannot be assigned to a genus | 0.796551 | Thermoanaerobacterales | | | | |
| 7.911305 | class < 0.5% of all reads | 0.765845 | Comamonadaceae | 66.64942 | genus < 0.5% of all reads | 0.775648 | Sphingobacteriales | | | | |
| ------------------------------------------ | | 0.758908 | Solibacteraceae | ------------------------------------------ | | 0.758908 | Solibacterales | | | | |
| 71.34026 | unclassified | 0.729903 | Rhizobiaceae | 71.34026 | unclassified | 0.723189 | Acidobacteriales | | | | |
| | | 0.723189 | Acidobacteriaceae | | | 0.710477 | Streptomycetales | | | | |
| | | 0.710477 | Streptomycetaceae | | | 0.695885 | Anaerolineales | | | | |
| | | 0.699914 | Chromatiaceae | | | 0.662942 | Chlorobiales | | | | |
| | | 0.695885 | Anaerolineaceae | | | 0.659764 | Nitrospirales | | | | |
| | | 0.686799 | Alteromonadaceae | | | 0.658779 | Synechococcales | | | | |
| | | 0.667642 | Flammeovirgaceae | | | 0.626328 | Opitutales | | | | |
| | | 0.662942 | Chlorobiaceae | | | 0.604799 | Methanosarcinales | | | | |
| | | 0.659764 | Nitrospiraceae | | | 0.599204 | Bacteroidetes Order II. Incertae sedis | | | | |
| | | 0.65578 | Polyangiaceae | | | 0.597368 | Chitinophagales | | | | |
| | | 0.635817 | Ectothiorhodospiraceae | | | 0.577763 | Thiotrichales | | | | |
| | | 0.626328 | Opitutaceae | | | 0.569617 | Rhodocyclales | | | | |
| | | 0.599204 | Rhodothermaceae | | | 0.555294 | Calditrichales | | | | |
| | | 0.597368 | Chitinophagaceae | | | 0.531303 | Oscillatoriales | | | | |
| | | 0.569617 | Rhodocyclaceae | | | 0.529736 | Vibrionales | | | | |
| | | 0.555294 | Calditrichaceae | | | ------------------------------------------ | | | | | |
| | | 0.529736 | Vibrionaceae | | | 1.333493 | Viruses | | | | |
| | | 0.504491 | Methanosarcinaceae | | | 6.761418 | cannot be assigned to a order | | | | |
| | | 0.503775 | Salinivirgaceae | | | 16.14649 | order < 0.5% of all reads | | | | |
| | | 0.501671 | Xanthomonadaceae | | | ------------------------------------------ | | | | | |
| | | ------------------------------------------ | | | | 71.34026 | unclassified | | | | |
| | | 1.333493 | Viruses | | | | | | | | |
| | | 8.514582 | cannot be assigned to a family | | | | | | | | |
| | | 35.14166 | family < 0.5% of all reads | | | | | | | | |
| | | ------------------------------------------ | | | | | | | | | |
| | | 71.34026 | unclassified | | | | | | | | |

**A4 Table 5:** *Kaiju taxonomic classification of metagenome reads for Sample 7. The percentage of the metagenome classified and the taxonomic assignment is given, from the class to species level. The percentage of the reads that were unclassified is also provided.*

| % | class | % | family | % | genus | % | order | % | phylum | % | species |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 21.08804 | Gammaproteobacteria | 4.331581 | Nitrosopumilaceae | 4.331581 | Nitrosopumilus | 6.482216 | Rhizobiales | 54.29175 | Proteobacteria | 2.815322 | Woeseia oceani |
| 16.16458 | Alphaproteobacteria | 2.842116 | Planctomycetaceae | 2.815322 | Woeseia | 5.585598 | Chromatiales | 6.939337 | Bacteroidetes | 2.038815 | Candidatus Nitrosopumilus sedimin |
| 8.26172 | Deltaproteobacteria | 2.815322 | Woeseiaceae | 1.528692 | Pseudomonas | 4.767477 | Burkholderiales | 5.784874 | Firmicutes | 1.882659 | Candidatus Nitrosopumilus adriatic |
| 7.449125 | Betaproteobacteria | 2.603863 | Rhodospirillaceae | 0.842361 | Nitrospira | 4.331581 | Nitrosopumilales | 5.645465 | Actinobacteria | 0.689345 | Candidatus Solibacter usitatus |
| 5.005143 | Actinobacteria | 2.567148 | Rhodobacteraceae | 0.829216 | Bradyrhizobium | 4.089393 | Planctomycetales | 5.145432 | Thaumarchaeota | 0.686457 | Fuerstia marisgermanicae |
| 4.089393 | Planctomycetia | 2.037391 | Flavobacteriaceae | 0.825573 | Geobacter | 3.209103 | Rhodospirillales | 4.566359 | Planctomycetes | 0.528627 | Pirellula staleyi |
| 2.966957 | Clostridia | 1.854442 | Burkholderiaceae | 0.798947 | Halioglobus | 2.897294 | Rhodobacterales | 2.292808 | Cyanobacteria | ------------------------------------------ | |
| 2.341288 | Bacilli | 1.652026 | Pseudomonadaceae | 0.770856 | Streptomyces | 2.876612 | Myxococcales | 1.776448 | Euryarchaeota | 1.368434 | Viruses |
| 2.202966 | Flavobacteriia | 1.456685 | Bradyrhizobiaceae | 0.727652 | Burkholderia | 2.639155 | Alteromonadales | 1.520026 | Acidobacteria | 16.14101 | cannot be assigned to a species |
| 2.182787 | Cytophagia | 1.389031 | Sphingomonadaceae | 0.689345 | Candidatus Solibacter | 2.202966 | Flavobacteriales | 1.436422 | Chloroflexi | 73.84934 | species < 0.5% of all reads |
| 1.018151 | Nitrospira | 1.275327 | Desulfobacteraceae | 0.686457 | Fuerstia | 2.182787 | Cytophagales | 1.018151 | Nitrospirae | ------------------------------------------ | |
| 0.906707 | Bacteroidia | 1.240369 | Rhizobiaceae | 0.646518 | Mycobacterium | 2.161981 | Clostridiales | 0.812219 | Verrucomicrobia | 71.03285 | unclassified |
| 0.760934 | Spirochaetia | 1.214455 | Chromatiaceae | 0.632535 | Azospirillum | 1.947005 | Pseudomonadales | 0.760934 | Spirochaetes | | |
| 0.689345 | Solibacteres | 1.205119 | Ectothiorhodospiraceae | 0.599378 | Planctomyces | 1.945917 | Cellvibrionales | 0.571705 | Deinococcus-Thermus | | |
| 0.677288 | Methanomicrobia | 1.155342 | Comamonadaceae | 0.584558 | Desulfovibrio | 1.934236 | Bacillales | 0.513137 | Gemmatimonadetes | | |
| 0.654639 | Acidobacteriia | 1.084591 | Hyphomicrobiaceae | 0.555546 | Desulfococcus | 1.800939 | Sphingomonadales | ------------------------------------------ | | | |
| 0.571705 | Deinococci | 1.018151 | Nitrospiraceae | 0.548261 | Magnetospirillum | 1.722652 | Desulfobacterales | 1.368434 | Viruses | | |
| 0.545582 | Opitutae | 0.996507 | Alteromonadaceae | 0.532311 | Paenibacillus | 1.706659 | Oceanospirillales | 1.849 | cannot be assigned to a phylum | | |
| 0.513137 | Gemmatimonadetes | 0.948907 | Geobacteraceae | 0.528627 | Pirellula | 1.660315 | Desulfuromonadales | 3.707502 | phylum < 0.5% of all reads | | |
| ------------------------------------------ | | 0.884561 | Rhodocyclaceae | 0.513137 | Gemmatimonas | 1.380617 | Corynebacteriales | ------------------------------------------ | | | |
| 1.368434 | Viruses | 0.83604 | Isosphaeraceae | 0.51096 | Thioalkalivibrio | 1.072701 | Xanthomonadales | 71.03285 | unclassified | | |
| 10.89422 | cannot be assigned to a class | 0.832732 | Phyllobacteriaceae | ------------------------------------------ | | 1.018151 | Nitrospirales | | | | |
| 9.647863 | class < 0.5% of all reads | 0.798947 | Halieaceae | 1.368434 | Viruses | 0.976035 | Enterobacterales | | | | |
| ------------------------------------------ | | 0.796268 | Bacillaceae | 9.618181 | cannot be assigned to a genus | 0.952465 | Thiotrichales | | | | |
| 71.03285 | unclassified | 0.793337 | Streptomycetaceae | 68.51556 | genus < 0.5% of all reads | 0.884561 | Rhodocyclales | | | | |
| | | 0.718818 | Polyangiaceae | ------------------------------------------ | | 0.794468 | Desulfovibrionales | | | | |
| | | 0.689345 | Solibacteraceae | 71.03285 | unclassified | 0.793337 | Streptomycetales | | | | |
| | | 0.679214 | Alcaligenaceae | | | 0.734392 | Synechococcales | | | | |
| | | 0.678 | Peptococcaceae | | | 0.732675 | Syntrophobacterales | | | | |
| | | 0.677581 | Desulfuromonadaceae | | | 0.702282 | Micrococcales | | | | |
| | | 0.67193 | Mycobacteriaceae | | | 0.689345 | Solibacterales | | | | |
| | | 0.671385 | Cytophagaceae | | | 0.654639 | Acidobacteriales | | | | |
| | | 0.654639 | Acidobacteriaceae | | | 0.64882 | Vibrionales | | | | |
| | | 0.65376 | Paenibacillaceae | | | 0.640112 | Bacteroidales | | | | |
| | | 0.64882 | Vibrionaceae | | | 0.610012 | Methylococcales | | | | |
| | | 0.610012 | Methylococcaceae | | | 0.583093 | Bacteroidetes Order II. Incertae sedis | | | | |
| | | 0.606369 | Xanthomonadaceae | | | 0.580664 | Oscillatoriales | | | | |
| | | 0.603858 | Desulfovibrionaceae | | | 0.570408 | Thermoanaerobacterales | | | | |
| | | 0.595945 | Acetobacteraceae | | | 0.520463 | Pseudonocardiales | | | | |
| | | 0.585939 | Clostridiaceae | | | 0.51816 | Nitrosomonadales | | | | |
| | | 0.583093 | Rhodothermaceae | | | 0.513137 | Gemmatimonadales | | | | |
| | | 0.553327 | Oxalobacteraceae | | | 0.511504 | Spirochaetales | | | | |
| | | 0.521426 | Colwelliaceae | | | 0.50715 | Caulobacterales | | | | |
| | | 0.520463 | Pseudonocardiaceae | | | ------------------------------------------ | | | | | |
| | | 0.51816 | Nitrosomonadaceae | | | 1.368434 | Viruses | | | | |
| | | 0.513137 | Gemmatimonadaceae | | | 8.128674 | cannot be assigned to a order | | | | |
| | | 0.507317 | Cellvibrionaceae | | | 16.76981 | order < 0.5% of all reads | | | | |
| | | 0.50715 | Caulobacteraceae | | | ------------------------------------------ | | | | | |
| | | 0.50514 | Myxococcaceae | | | 71.03285 | unclassified | | | | |
| | | 0.501289 | Syntrophaceae | | | | | | | | |
| | | 0.500619 | Piscirickettsiaceae | | | | | | | | |
| | | ------------------------------------------ | | | | | | | | | |
| | | 1.368434 | Viruses | | | | | | | | |
| | | 9.866397 | cannot be assigned to a family | | | | | | | | |
| | | 34.65817 | family < 0.5% of all reads | | | | | | | | |
| | | ------------------------------------------ | | | | | | | | | |
| | | 71.03285 | unclassified | | | | | | | | |