Author:
Varliero, Gilda

Title:
Arctic microbial exploration: a bioinformatics approach

Author:
Varliero, Gilda

Title:
Arctic microbial exploration: a bioinformatics approach

Arctic microbial exploration: a bioinformatics approach

Gilda Varliero

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy (PhD) in the Faculty of Life Sciences.

School of Biological Sciences

September 2020

Word count: 43,933

Abstract

The Arctic environment is a microbe-driven biome where microorganisms actively shape the environment and mediate biogeochemical cycles. Due to the life-challenging conditions that this environment poses (e.g. subzero temperatures, water and nutrient depletion), microorganisms have developed coping metabolic pathways and enzyme adaptations, constituting a reservoir for the bioprospecting of new molecules and cold-adapted enzymes. The advent of cheap and diffused sequencing technologies has boosted the study of environmental biodiversity and genome reconstruction. In this thesis I use different sequencing technologies (i.e. Illumina and Nanopore), bioinformatics approaches (i.e. amplicon and whole shotgun metagenomics, and metatranscriptomics) and pipelines for the characterization of Arctic microorganisms. In particular, this thesis presents i) PhyloPrimer, a new online user-friendly software for the semi-automated design of taxon-specific primers to perform taxonomic and functional driven studies, ii) the characterization of microbial community distribution in unexplored englacial channels using the universal 16S ribosomal RNA gene. I then present iii) the LongMeta pipeline used to screen whole shotgun metagenomic sequencing data and used to explore microbial communities and their role in rock weathering and nitrogen fixation processes in proglacial systems. Finally, iv) I explore cold-adapted microbial communities in the active layer of proglacial permafrost to create the cold-adapted predicted protein (CAPP) database to provide sequence data to cold-adapted protein bioengineering studies.

Overall, this thesis presents the application and development of several sequencing technologies and computational pipelines available for microbial characterization and bioprospecting, exploring the microbial communities of rapidly changing environments in the Arctic: the glacier, its proglacial system and soil permafrost.

Acknowledgements

My PhD research was possible thanks to the funding support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 675546. Additionally, I thank the NERC, INTERACT and the University of Bristol Alumni Foundation that provided additional funding contributing to field activities and conference expenses.

I am deeply grateful to Gary Barker for his advice and encouragements during the past four years, for always leaving his door open even during the busiest times, and to be the most equipped and efficient field assistant ever! I am also thankful to Marian Yallop who with her kindness was always ready for a chat and to explain me new lab techniques and projects. Thanks also to Alex Anesio for being such a wonderful collaborator and for asking me to take part to great (and fun) field activities and projects.
A huge thanks also to everyone in the Cereal Genomics Lab and the Bristol Genomics Facility for their support and advice. In particular Jane, Christy, Keith, Mark and Amanda for their help and guidance in the lab, Paul for his never-ending patience, and Alex for letting me monopolize his server.
I also had the luck to be part of the MicroArctic, an amazing PhD network made from incredible and supportive people that made me feel welcomed right from the start of my PhD path (www.microarctic.eu).

All of the hard work was made less hard thanks to the coffees, pints, parties and experiences shared with very special people: Chiara (x2), Zoe, Ashley, Vicky, Mike, Alice, James, Aude, Abs and the other Bristolians. Thanks also to all of my friends in Italy who are always so close to me in spite of the distances and especially in the time of need. I am deeply thankful to my family that always sustained me during my academic studies and always has been my number one fan. Finally, a huge thank to Robi who lived my PhD almost as much as I did, always having my back and supporting me in every possible way.

Author's declaration

I declare that the work in this dissertation 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 with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: [redacted] DATE: 12/09/2020

Table of contents

List of figures

List of tables

Chapter 1

General introduction

1.1 Glacial and proglacial systems

1.1.1 Glaciers, ice caps and ice sheets

At a first glance, glaciers may look like enormous inactive blocks of ice, but in reality, are complex dynamic systems which, thanks to their movement, play a key role in shaping the environment. Glacier expansion or retreat is a complex phenomenon where the glacier mass balance relies on the relation between the accumulation zone, where the glacier gains mass by snow deposition and subsequent compaction into firn and ice, and the ablation zone, where the glacier loses mass through snow and ice melts, avalanches, calving and wind. These two zones are divided by the equilibrium line and their balance changes in relation to the amount of snow and ice that melts during the melting season: if more than the usual amount of ice melts during a season, the equilibrium line moves up, leading to an increase in the ablation area and to a decrease in the accumulation area, and vice-versa (Cogley, 2011). Whereas it is possible to have a clear ice stratigraphy in the accumulation zone, as all the ice was formed by snow deposition and compacting, the ice in the ablation zone presents mixed layers because this ice was formed in the accumulation zone and brought to the ablation zone by a non-linear ice uplifting and sliding (Hudleston, 2015) (Figure 1.1). Glaciers slide under their own weight thanks to the formation of a water layer between the ice and the bedrock due to the high pressure and the friction exercised by the ice mass. This water also comes from the glacier surface where the snow and the ice are seasonally melted and form supraglacial rivers, streams and lakes that percolate under the glaciers through englacial channels, crevasses and moulins (Brown, 2002; Fountain and Walder, 1998; Hotaling et al., 2017) (Figure 1.1). It is not uncommon for these formations to originate from cryoconite holes which are water-filled holes caused by the deposit of particles on the glacier surface which causes the progressive melting of the surrounding ice (Fountain et al., 2004; MacDonell and Fitzsimons, 2008).

Ice caps and ice sheets (e.g. Greenland Ice Sheet) undergo the same processes and dynamics of glaciers. Contrary to glaciers, they are dome-shaped and therefore the ice movement is in different radial directions. The difference between ice caps and ice sheets is their extension: the first is smaller than 50,000 km^2, whereas the other is bigger (Laybourn-Parry et al., 2012).

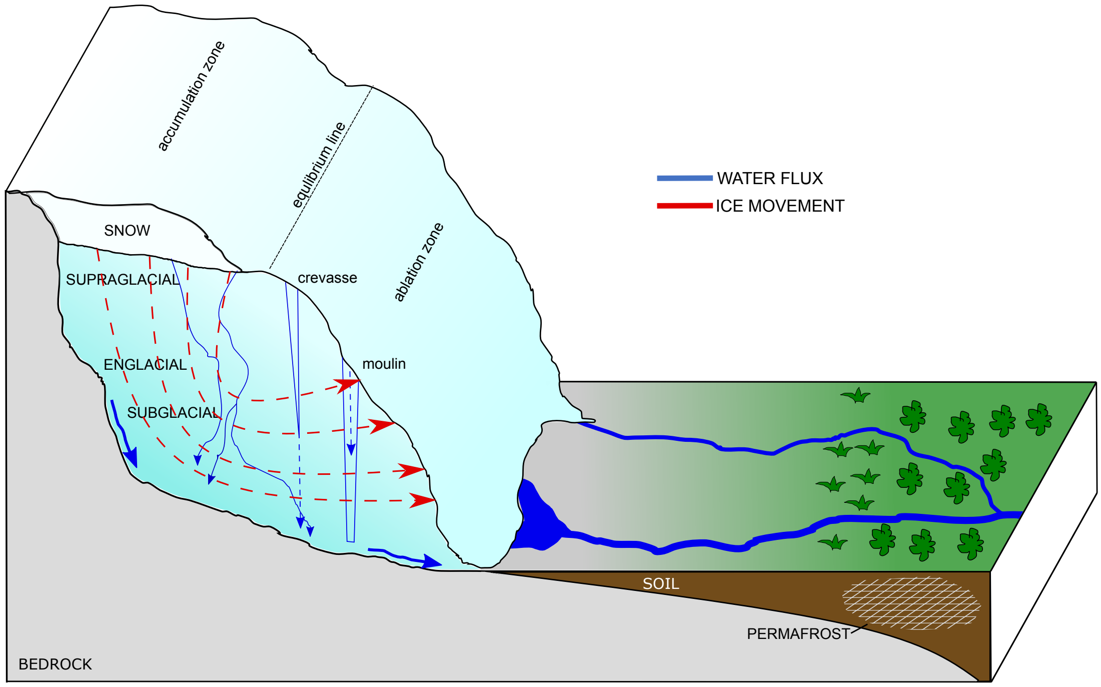


Figure 1.1: Scheme of the glacial and proglacial environment.

Together with the glaciers they occupy the 9.6% of the Earth's surface (Marshall, 2005).

1.1.2 Proglacial systems

The supraglacial, englacial and subglacial compartments of glaciers and ice sheets are interconnected by a complex system of englacial channels. These channels also connect the ice realm with the proglacial systems where the glacial melted water is finally discharged (Carrivick and Heckmann, 2017). Proglacial environments are sited in front of an ice front (i.e. glacier, ice cap and ice sheet) and represent transition zones between glacial and non-glacial habitats (Slaymaker, 2011). A forefield is a type of proglacial system defined as the area between the glacier edge and the terminal moraine which is the accumulation of debris and rocks pushed along by the ice movement. This environment is constituted of bare bedrocks which are transformed into soil thanks to chemical, physical and biological weathering phenomena. Soil complexity increases going further from the ice front with a consequent increase in the soil depth, organic matter content and a decrease of the dimension of soil particles, leading to the growth of an increasingly complex vegetation along the proglacial systems (Bradley et al., 2014; Fernández-Martínez et al., 2016; Carrivick and Heckmann, 2017) (Figure 1.1).

1.1.3 Permafrost

Proglacial systems present permafrost in their soil and sediments (Heckmann et al., 2016; Hodgkins et al., 2004). Permafrost is defined as any type of ground (e.g. soil and sediment) that has been frozen for more than two years. Its upper layer undergoes seasonal cycles of thawing and freezing and is called active layer. Permafrost is widely spread in polar regions and at high altitude where the mean annual air temperature is below -1 °C. Permafrost is estimated to occur in the 24% of the North hemisphere lands and in the entire Antarctic region (Dobinski, 2011). The Arctic permafrost alone is estimated to sequester between 1330 and 1580 Pg of carbon which corresponds to twice the carbon present in the atmosphere, most of which has been trapped in the soil for thousands of years (Schuur et al., 2015).

Permafrost-affected environments also present thermokarst bogs and lakes which are specific formations where the permafrost ice melts forming bogs and lakes (in 't Zandt et al., 2020).

1.2 Environmental microorganisms

Microorganisms are extremely diversified and ubiquitous. The number of microbial species is estimated to be up to 10^{12} (Locey and Lennon, 2016; Lennon and Locey, 2020; Louca et al., 2019) but only 10^5 have publicly available sequences. From hot springs to hypersaline cold ice veins, microorganisms have been observed in all of Earth's habitats (Whitman et al., 1998).

Microbial ubiquity is due to their high dispersion and growth rate, and to the ability to adapt to challenging and changing environmental conditions (Smith et al., 2011). The latter is due to a wide metabolic differentiation which allows microorganisms to survive challenging conditions and use different molecules as source of electrons and metabolic energy, and to a high mutation rate which allows them to adapt and quickly occupy new biological niches (Kivisaar, 2003; Ram and Hadany, 2014). The capability to occupy many different ecological niches enables microorganisms to thrive in otherwise inhospitable habitats, taking advantage of the harsh settings, and to have a leading role in environmental shaping and biogeochemical cycles.

1.2.1 Microbial metabolism

Organisms can be classified in relation to their metabolism. The classification is based on three main aspects: the energy source used for cell energy production, the electron donor used for metabolic redox reactions and the carbon source. Organisms are called phototrophs or chemotrophs if they use a light or chemical source for energy production, they are called organotrophs or lithotrophs if they use organic or inorganic compounds as electron donors, and heterotrophs or autotrophs if they are able to obtain carbon from organic or inorganic compounds (i.e. CO_2). Bacteria and Archaea comprise organism representatives of all metabolic classes.

Photosynthetic eukaryotic organisms (e.g. plants and algae) and prokaryotic organisms (e.g. Cyanobacteria) have a photolithoautotrophic metabolism and are called primary producers. They obtain energy for biosynthetic reactions from light and inorganic electric donors (e.g. H_2O) and obtain carbon source by fixing inorganic carbon (i.e. CO_2). These organisms are essential to the environment: they sequester carbon from the atmosphere and convert it into organic compounds that can be used by other organisms (e.g. heterotrophs). Oxygenic photosynthesis also produces oxygen which is at the base of all the aerobic organism metabolisms (Singh et al., 2010).

The other class of primary producers are the chemolithotrophs and their metabolism has been observed only in prokaryotic organisms. They are the only organisms able to use inorganic compounds both as energy source and electron donors. Chemolithoautotrophs use inorganic compounds (e.g. CO_2) as carbon source and therefore take part into the organic carbon stock construction of the environment, whereas chemolithoheterotrophs use organic carbon sources instead. The latter are mixotrophs and are not strictly primary producers but they have been observed to switch between autotrophy and heterotrophy, or also chemotrophy and phototrophy, depending on the available environmental resources and conditions (Ward, 2019; Eiler, 2006).

The centrality of lithotrophs in the ecosystem functioning is due to their ability to oxidate a wide variety of different inorganic compounds, converting reduced chemical species to oxidized and more bioavailable compounds (Barton et al., 2010). Common litotrophic metabolisms involve, for example, the oxidation of hydrogen into water, ammonia into nitrite, nitrite into nitrate and iron(II) into iron(III) (Lesniewski et al., 2012; Bryce et al., 2018).

All the animals, the majority of fungi and some prokaryotic organisms are chemoorganoheterotrophs. These organisms obtain energy from chemical reactions, electrons from organic chemicals and carbon from organic carbon. These organisms are limited to the presence of organic resources for their survival and they are extremely important to the environmental functioning where they are responsible for the organic matter mineralization where complex organic compounds are decomposed and oxidized into more bioavailable molecules (Bridgham and Ye, 2015). During heterotrophic respiration the organic carbon is respired with consequent CO_2 emission. This process is contrasted with environmental carbon sequestration where less labile carbon molecules, such as lignin and cellulose, can lead to environmental carbon sequestration with the formation, in millions of years, of petroleum and coat deposits (Davidson and Janssens, 2006; Thakur et al., 2018).

The variety, complexity and plasticity of microbial metabolisms give them a pivotal role in the biogeochemical cycles of the major elements necessary to the construction and functioning of living cells (i.e. carbon, nitrogen, hydrogen, oxygen, nitrogen and sulfur) and enzyme cofactors (e.g. iron and manganese), where biological redox reactions create biological fluxes of these elements into the Earth's geochemical cycles (Falkowski et al., 2008; Raiswell and Canfield, 2012; Zhang et al., 2020).

Prokaryotic and eukaryotic organisms take part to the carbon cycle driving the conversion between inorganic and organic carbon, as reported in the previous paragraphs.

Nitrogen cycle is also driven by microorganisms where some of its biogeochemical steps are uniquely mediated by these organisms, such as the nitrogen fixation and the nitrification processes. Nitrogen fixation is performed by diazotrophs which fix atmospheric nitrogen to ammonia. Ammonia can then be oxidized to other nitrogen forms with the nitrification process where ammonia is oxidized to nitrite (nitritation), nitrite is oxidized to nitrate (nitratation), or ammonia is directly oxidized to nitrate (comammox). The nitrogen cycle closes with the denitrification and annamox processes where nitrite, ammonia and other nitrogen species are respired and reconverted to atmospheric nitrogen (Jetten, 2008; Stein and Klotz, 2016).

Ammonium and nitrate are the principal nitrogen sources for plants. Plant nitrogen uptake can be facilitated by plant association with nitrogen fixing bacteria and archaea where they acquire the ability to obtain ammonium as nitrogen source from atmospheric nitrogen (Wernegreen, 2012).

Microbial symbiosis with macroorganisms are not unusual in nature where the extreme microbial metabolic diversity can advantage other organisms. For example, gut microbial communities form deep microbe-host interactions in a wide range of macroorganisms, from insects (Engel and Moran, 2013; Pernice et al., 2014) to humans (Nicholson et al., 2012). They regulate the health status of the host, providing nutrients and helping to maintain the full functionality of the gut mucosal barrier (Thursby and Juge, 2017; Feng et al., 2018; Federico et al., 2017).

Further to the associations between microorganisms and pluricellular organisms (e.g. animals and plants), microbial associations shape and condition local microbial community structures. Within a microbial community, microorganisms create mutualistic and complex networks to survive and uptake the needed resources from the environment. Cross-feeding of molecules is common between environmental organisms as it is convenient from a metabolic and energetic point of view (D'Souza et al., 2018; Cavaliere et al., 2017), especially in environments, as the Arctic, that impose harsh life conditions.

Microorganisms can be found in different metabolic stages depending on environmental conditions. Dormancy is the ability to enter a reversible state of low metabolism activity to avoid adverse conditions and to resume the usual metabolic levels once the environmental conditions are back to normal (Lennon and Jones, 2011). Microorganisms, mainly belonging to the phyla Firmicutes, Actinomyces, Cyanobacteria and Gammaproteobacteria, can also enter a dormant state by the formation of resisting structures (e.g. spores and cysts) (Paul et al., 2019).

1.2.2 Microbial dispersion and biogeography

Their highly plastic and diversified metabolisms give microorganisms the ability to occupy all the environmental niches. Additionally, their ubiquity is due to their small size which give them the advantage to be easily transported globally, being diffused by air, water currents and meteoric phenomena as both free-living and host-associated cells (Wilkinson et al., 2012; Müller et al., 2014). For this reason, for years it has been assumed that all the microorganisms were spread everywhere and that the high connectivity could not allow endemisms, following the idea ‘Everything is everywhere, but the environment selects’ (Baas-Becking, 1934; De Wit and Bouvier, 2006). However, thanks to the advent of extensive use of sequencing technologies, nowadays microbial distributions can be explored more in details and that paradigm is shifting towards a more accurate study and comprehension of microbial biogeography (Fontaneto and Hortal, 2012; O'Malley, 2007). Microbial biogeographic studies aim to reconstruct microbial diversity patterns across different spatial and time scales.

It has been shown how microbial dispersion and environmental connectivity influence microbial distribution where communities in closed and small spaces (e.g. gut microbiome) are more likely to present similar species across different samples compared to wide and spread environmental communities (Livermore and Jones, 2015). Further, several studies have observed how microbial dispersion is different from macroorganism dispersion concluding that, however, that does not mean that there is no development of microbial patterns and different communities (e.g. Karimi et al., 2018; Fierer and Jackson, 2006) and that even with high dispersion rates, there can be the creation of endemisms due to microbial microdiversification (i.e. organism diversification in closely related organisms) into different ecological niches (Larkin and Martiny, 2017). Organism speciation aims to occupy new ecological niches and has been shown to be more frequent in generalistic (i.e. able to adapt to a wide range of habitats) rather than specialistic (i.e. able to adapt to only specific habitats) organisms (Sriswasdi et al., 2017).

Notoriously unified environments such as oceans and atmosphere have shown microbial patterns (e.g. Salazar et al., 2016 and Els et al., 2019b). Hellweger et al. also highlights how microorganisms evolve faster than how fast they are dispersed in the environment (2014). Soil studies have shown biogeographic patterns (Karimi et al., 2018; Fierer and Jackson, 2006). The majority of studies show a core community shared across different areas but also the presence of endemic and site-specific organisms (Delgado-Baquerizo et al., 2018). The patterns that drive diversification are not always clear, being also conditioned by taxon-specific characteristics. However, pH and vegetation content are the ones that have been mainly identified as differentiation drivers (Malard and Pearce, 2018).

The Arctic region shows microbial biogeography patterns across several environments. Malard et al., for example, reported how soil microbial communities are constituted by common microorganisms across the Arctic but also showed locally-driven microbial differentiation (2019). Similarly, these kind of trends were found across biogeographically distributed microbial communities sampled from cryoconite holes (Segawa et al., 2017), red snow (Lutz et al., 2015) and permafrost active layer (Ren et al., 2018).

1.2.3 Glacial and proglacial microbial communities

The Arctic environment was recognized as a biome in 2012 (Anesio and Laybourn-Parry, 2012). This biome is dominated by psychrophilic microorganisms that actively influence and shape this environment. Psychrophiles are cold-adapted organisms and need to overcome many challenges set by the subzero temperatures (Cavicchioli, 2016). Additionally to low temperatures, the Arctic environment also poses other challenging life conditions, such as high salinity, water deprivation and nutrient scarcity (Boetius et al., 2015).

The ice surface of glaciers, ice sheets and ice caps are dominated by prokaryotic (e.g Cyanobacteria) and eukaryotic (e.g. ice algae) autotrophs which are the undisputed primary producers of this nutrient-depleted system (Musilova et al., 2015; Lutz et al., 2017; Yallop et al., 2012). These organisms are well adapted to this habitat as they can use the sun light as energy source and the CO_2 as carbon source. Important for the microbial life on the ice surface are also the cryoconite holes which can occupy up to the 10% of glacier ice surface (Anesio et al., 2009). Cryoconites have have been widely studied because they constitute life hotspots on the glacier surface as these water-filled formations are enriched with nutrients and therefore present high microbial activity (e.g. Bagshaw et al., 2007; Cook et al., 2016; Edwards et al., 2011). However, microbial communities are developed also on the bare ice (Anesio and Laybourn-Parry, 2012). In particular, ice and snow algae have been shown to dominate and shape this environment where they are also visibly present and form darker ice formations (Hoham and Remias, 2020; Williamson et al., 2019).

Below the glacial surface, the englacial system represents the biggest area of the glacier, extending to the glacial bottom. However, this glacial portion is not very well characterized due to the technical sampling difficulties and to the assumption that its life conditions are too challenging to life. However, several studies have now shown how this realm can sustain life both in ice veins and englacial channels (e.g. Dani et al., 2012; Miteva, 2008; Martinez-Alonso et al., 2019; Zeng et al., 2013).

Water flows from the surface to the subsurface of the glacier where the water maintains a liquid form thanks to the high pressure exercised by the glacier mass. The pressure and the water flowing allow the typical glacier movement, and the bedrock grinding with the consequent release of H^+ and inorganic compounds into the system. This realm is therefore dominated by chemolitotrophs able to use these reduced compounds (Boyd et al., 2014; Dieser et al., 2014; Kayani et al., 2018).

These ice systems are a net sink of carbon where both photoautotrophs in the surface and chemoautotrophs in the subglacial system organicate atmospheric carbon (i.e. CO_2) creating carbon stocks that the heterotrophic organisms, not so active in this system, only partially use (Stibal et al., 2012a).

Glaciers, ice caps and ice sheets release between 10^{17} and 10^{21} microorganisms each year into proglacial systems (Castello and Rogers, 2005). Microorganisms that are dominant in the glacial environment (i.e. autotrophs and chemolitotrophs) constitute a first inoculum into the young-exposed ground (i.e. bedrocks) in proximity of the ice edge. Autotrophs and litotrophs have an advantage compared to the other organisms in this soil- and nutrient- depleted environment as they can uptake reduced molecules or use the sun light as energy source.

In particular, lithotrophs drive biotic rock dissolution and promote the soil formation, releasing important metabolic nutrients and enzyme cofactors such as iron and sulfur into the environment (Napieralski et al., 2019; Olsson-Francis et al., 2015; Frey et al., 2010). Progressing from the ice edge, this phenomenon, along with the progressive fixation of atmospheric nitrogen into the soil creates geochemical gradients in the system with an enrichment of carbon, nitrogen and metals in the soil and establishes more complex communities where heterotrophs are more advantaged of autotrophs (Bradley et al., 2016). The development of a rhizosphere and of a progressively complex vegetation further promote, in turn, soil formation (Knelman et al., 2012; Rime et al., 2015). Proglacial systems are model systems to understand microbial successions and how microbial communities progress along different geochemical variables.

Associated with soil formation, permafrost can occur in the proglacial systems. Permafrost represents a dark and water-depleted environment to life where aerobic and anaerobic heterotrophs such as methanogens, nitrogen fixers and sulfate reducers, are mainly found (Steven et al., 2006; Jansson and Taş, 2014). This environment is enriched in organic carbon content but the latter is not always bioavailable as it is trapped in the ice crystals (Schuur et al., 2015; Ward et al., 2017). Organic carbon, however, becomes more available with the permafrost thawing (e.g. in the active layer) where microbial communities have showed quick responses in incubation studies (e.g. MacKelprang et al., 2011; Wei et al., 2018), *in situ* temporal studies (e.g. Vigneron et al., 2019) and along progressive soil depths (e.g. Deng et al., 2015; Müller et al., 2018).

1.2.4 Microbial shifts in a warming Arctic

The past several decades registered a significant increase in the Earth's temperature. Polar environments have been severely affected by it, registering ice mass losses worldwide (ACIA, 2005; Hanna et al., 2013; Meier et al., 2007; King et al., 2019; Braun et al., 2019). Due to the increased glacial water discharge, the sea level rise has been of 3 mm per years since 1993 but it is due to increase (Nerem et al., 2018; Dangendorf et al., 2017). An accurate prediction of the environmental changes is made difficult by the environmental abiotic and biotic positive feedback responses to the global warming, such as an albedo increase and the greenhouse gas release from the permafrost.

Ice melting causes an increase in the Earth's ground exposure. Dark surfaces (i.e. rocks and bedrocks) have a lower albedo than the white surfaces (i.e. ice). The direct consequence is a higher amount of absorbed heat in the Earth's surface and a further increase in the ice melting. Furthermore, ice melting leads to a higher water amount on the glacier surface which leads to an increase in the ice algae population which is less limited by water scarcity (Williamson et al., 2019). An increase in the algal ice coverage brings to a darker ice surface and, again, to a lower albedo leading to a further ice melting (Tedstone et al., 2019).

The permafrost thawing releases nutrients and carbon that have been trapped in the soil-ice matrix for the past ages, giving a positive feedback to the global warming and potentially increasing the greenhouse gas emissions (e.g. CO_2 and CH_4) in the atmosphere by 22-40% in the next century (Loranty et al., 2018; Yang et al., 2019; Whiteman et al., 2013). Further to direct release of greenhouse gases, many models predict a microbial response where microbes could exit their dormant state and respire the newly-released and biologically available organic carbon, further increasing the CO_2 emissions and the permafrost positive feedback to the environmental warming (Makhalanyane et al., 2015; Ganzert et al., 2007; MacKelprang et al., 2011).

A similar effect is estimated to happen due to the melting of the cryosphere where the glacier mountains alone have been estimated to trap 130 Tg of organic carbon; the warming of these systems would bring to a release of the stored carbon to the atmosphere both directly and through microbial mediated respiration (Wang et al., 2020). Further, the increase of the discharged water also directly impacts the proglacial systems increasing the water and nutrient input into the system and also soil complexity (Dubnick et al., 2017).

Shifts in the microbial communities have also potential implications for food webs, microbial interactions and biogeochemical cycles in the Arctic and worldwide (Deslippe et al., 2012; Vincent, 2010; Yergeau et al., 2012; Bell et al., 2013).

1.3 Cold environment exploration

Water and nutrient depletion, cold temperatures, high salinity and UV exposure are only some of the environmental challenging conditions that Arctic microbial communities must adapt to. The cell and genomic adaptations that these organisms have evolved to cope with these conditions are extremely varied constituting a valuable bioprospecting reservoir. Further, high presence of antibiotic producing genes has been reported in glacier environment where the competition is elevated due to the harsh environment (Segawa et al., 2013; Bell et al., 2013). Exploring the environment is essential to discover new species and novel compounds (Jansson and Prosser, 2013).

The adaptation to subzero temperatures implies an increase of cell membrane fluidity (i.e. increase in the membrane polyunsaturated fatty acids) and of protein flexibility (Åqvist et al., 2017; Margesin and Collins, 2019). The latter has been ascribed to the presence of less amino acid bonds in the protein structure (e.g. hydrogen bonds and salt bridges) given by the reduced presence of amino acids with hydrophobic side chains in the external protein regions and the presence of more amino acids with small and neutral side chains in protein loop regions (De Maayer et al., 2014; Loladze et al., 2002; Zhou and Zhou, 2004). Protein flexibility/stability is balanced in a delicate trade-off with the protein activity where it has been observed that a higher protein flexibility (e.g. cold-adapted proteins) corresponds to a higher activity but also to a consequent lower substrate specificity; whereas to a higher protein stability (e.g. heat-adapted proteins) corresponds a lower activity and a higher substrate specificity (Siddiqui, 2015; Siddiqui and Cavicchioli, 2006; Perl et al., 2000). Furthermore, protein active site and folding are essential to the correct protein functioning and specificity (Weng et al., 2011; Englander and Mayne, 2014). For all these reasons, even if some amino acids (e.g serine, threonine, asparagine, glutamine, histidine, tyrosine, tryptophan, aspartate, glutamate, arginine, proline and lysine) have been observed to form less amino acid bonds in secondary protein structures, amino acid composition is specific to protein sequence and function (Chao et al., 2020; D'Amico et al., 2006). Indeed many studies have compared homologous proteins between psychrophylic, mesophylic and thermophylic organisms and studied amino acid composition differences (e.g. Khrapunov

et al., 2017; Bae and Phillips, 2004; Khan et al., 2016; Du et al., 2017).
Cold-adapted enzymes, which have a low optimum temperature for enzymatic reactions, are relevant in several industry settings. Food, pharmaceutical and chemical are some of the industries that benefit from the use of cold-adapted enzymes which, performing enzymatic reactions at lower temperatures compared to their mesophilic homologous, are more energy sustainable and give less non-specific secondary chemical reactions and products (Mangiagalli et al., 2020; Kaur and Gill, 2019).

Another microbial adaptation to subzero temperatures and, in particular, to water thawing/freezing cycles is the production of cryoprotectants which are molecules that help the cell to avoid osmotic stress and membrane damage due to ice formation. Molecules such as ice-binding proteins (IPSs) and extracellular polymeric substances (EPSs) are widely exploited in industrial setting. Antifreeze proteins (AFPs) and ice-nucleating proteins (INPs) are IPSs used in the food industry as additives to improve food preservation during the food production and transport and to regulate food crystallization. EPSs are biopolymers that create a gel-like secretion used by the cells to facilitate, for example, the attachment to particles and other cells in order to stick to the scarce nutrient sources and to create cross-feeding associations with other organisms (Mangiagalli et al., 2020).

Microorganisms can produce a wide range of secondary metabolites (i.e. molecules not strictly necessary to the organism survival). Secondary metabolites are represented by molecules with different chemical structures and functions, such as terpenoids, alkaloids, peptides and polyketides. They are synthesized through different biosynthetic pathways and enzymes which are encoded by genes found in biosynthetic gene clusters (BGCs) (Weber et al., 2015; Castro et al., 2014). Since the development of the technology for the exploration of extreme habitats, industries have showed a growing interest in the discovery of new natural products (Razavi et al., 2017; Siddiqui and Cavicchioli, 2006). It is estimated that about 70% of compounds, used nowadays in pharmaceutical and biotechnology settings, are biologically derived or mimic natural products (Newman and Cragg, 2016).
Secondary metabolite peptides can be directly transcribed from a coding gene (i.e. ribosomal peptides) or synthesized by modular enzymes (i.e. nonribosomal peptides). Nonribosomal peptides are synthesize by non-ribosomal peptide synthetases (NRPSs) which are modular enzymes where different catalytic domains integrate different amino acid substrates to the growing amino acid chains till obtaining the final peptide (Miller and Gulick, 2016). Similarly, polyketide synthases (PKSs) are modular enzymes that synthesize polyketides. Nonribosomal peptides and polyketides have been widely used, for example, as antimicrobial and anticancer compounds and

their modular enzymes have been subject to many bioengineering studies were single domain modules have been modified (Williams et al., 2013).

Cold-adapted organisms and purified enzymes can also be used in the bioremediation field. Bioremediation aims to remove xenobiotic compounds (e.g. insecticides and plastics) from the environment (i.e. soil and groundwater) in both *in situ* and *ex situ* settings (Sharma et al., 2018; Kumar et al., 2019). Psychrophilic proteins have the advantage to allow bioremediation strategies in cold settings (Miri et al., 2019) or, if performed in reactors, it consents lower energy usage for chemical reactions (Tomei and Daugulis, 2013).

Finally, microbial compounds have been used for the creation and discovery of renewable products, such as bioplastics (Balaji et al., 2013; Karan et al., 2019) and ecosostenible alternative to chemical polymers (Xiao and Zheng, 2016).

1.4 Environmental sequencing

In the last sections I gave an overview of the extraordinary taxonomic and metabolic diversity observed in microbial environmental communities and how they can distribute following geochemical gradients and biogeographic patterns forming tight interaction with the environment and its functioning. It has been estimated than only the 0.1-1% of microorganisms are culturable due mainly to the impossibility to recreate the exact environmental conditions and microbial cross-feeding associations found in nature (Torsvik et al., 1996).

Since the commercialization of the Next-Generation Sequencing (NGS) technologies, the scientific community has stopped relying exclusively on culturing techniques for the study of microbial organisms. Even though the culturing approach is still essential for several biological applications (e.g. physiological studies and database annotation improvement), NGS technologies have revolutionized microbial environmental studies allowing the characterization and exploration of extremely complex and diverse environmental communities (Gutleben et al., 2018; Shokralla et al., 2012; Escobar-Zepeda et al., 2015). These technologies also provide a new tool to bioprospecting studies as they allow the systemic screening of new genes and proteins from unknown organisms (Roumpeka et al., 2017; Madhavan et al., 2017).

Launched in 2006, the most widely used sequencing platform today is Illumina which is a second-generation sequencing technology (Bleidorn, 2017). The input for this technology are sheared DNA fragments 500/1000 bp long. Its output consists of short DNA sequences between 30 and 300 base long (i.e. reads). Reads are usually paired-end which means that DNA fragments are sequenced at both ends (forward and reverse reads).

Third-generation sequencing technologies, whose main characteristic is to output longer sequences, have been increasingly used during the past few years. Among these, Nanopore sequencing was first commercialized in 2014 from the Oxford Nanopore Technologies (ONT) and its sequence lengths have no theoretical limits (Deamer et al., 2016); the longest sequence produced up to date is more than 2 Mb long (Payne et al., 2019). The MinION device can be run simply attaching it to the USB port of a portable computer giving this technology a great relevance for *in situ* sequencing of environmental samples (Brown et al., 2017). The main drawback of the ONT technology is the high sequence error rate. However, this aspect has been improved by new sequencing chemistry solutions and better bioiformatics software for the data analyses (Rang et al., 2018; Amarasinghe et al., 2020).

Sequence information from the DNA (i.e. metagenomics) and RNA (i.e. metatrascriptomics) of environmental communities has become relatively cheap, and easy to obtain. Billions of sequences can be found in online repositories and the numbers are due to increase (Marx, 2013; Selzer et al., 2018). Entrez, for example, is a search engine organized by the National Center for Biotechnology Information (NCBI) that gives freely access to 35 different databases and to 2.7 billion of records (September 2019) (Sayers et al., 2020). Whereas the easy and free accessibility to sequencing resources is an indisputable advantage, it also sets a challenge. Sure enough, the bioinformatics analyses of this data is computational and time demanding and often constitute the real bottleneck of scientific environmental studies and the analysis of environmental data.

1.4.1 Metagenomics

Metagenomics is the study of DNA directly extracted from an environmental sample (environmental DNA or eDNA). Metagenomic studies allow us to unravel the environmental complexity by directly sequencing the sample nucleotide content and gaining an insight into taxonomy and gene function of the entire community (Franzosa et al., 2015; Goodwin et al., 2016; Sharpton, 2014; Quince et al., 2017). Metagenomics can be applied with different approaches such as amplicon metagenomics and whole shotgun metagenomics in relation to the study aims and hypotheses (Figure 1.2) (Liu et al., 2020; Brumfield et al., 2020).

For environmental studies using high-throughput sequencing technologies, data analysis is now the limiting step in furthering our understanding of structure and function of microbial communities. The raw sequence data have to undergo a lot of intermediate steps before any biological meaning can be attributed to the DNA sequences. Each step of the pipeline requires the choice of which software/algorithm to use and which parameters to set. Each of these choices can potentially change the outcome of the research and so they have to be carefully optimized.

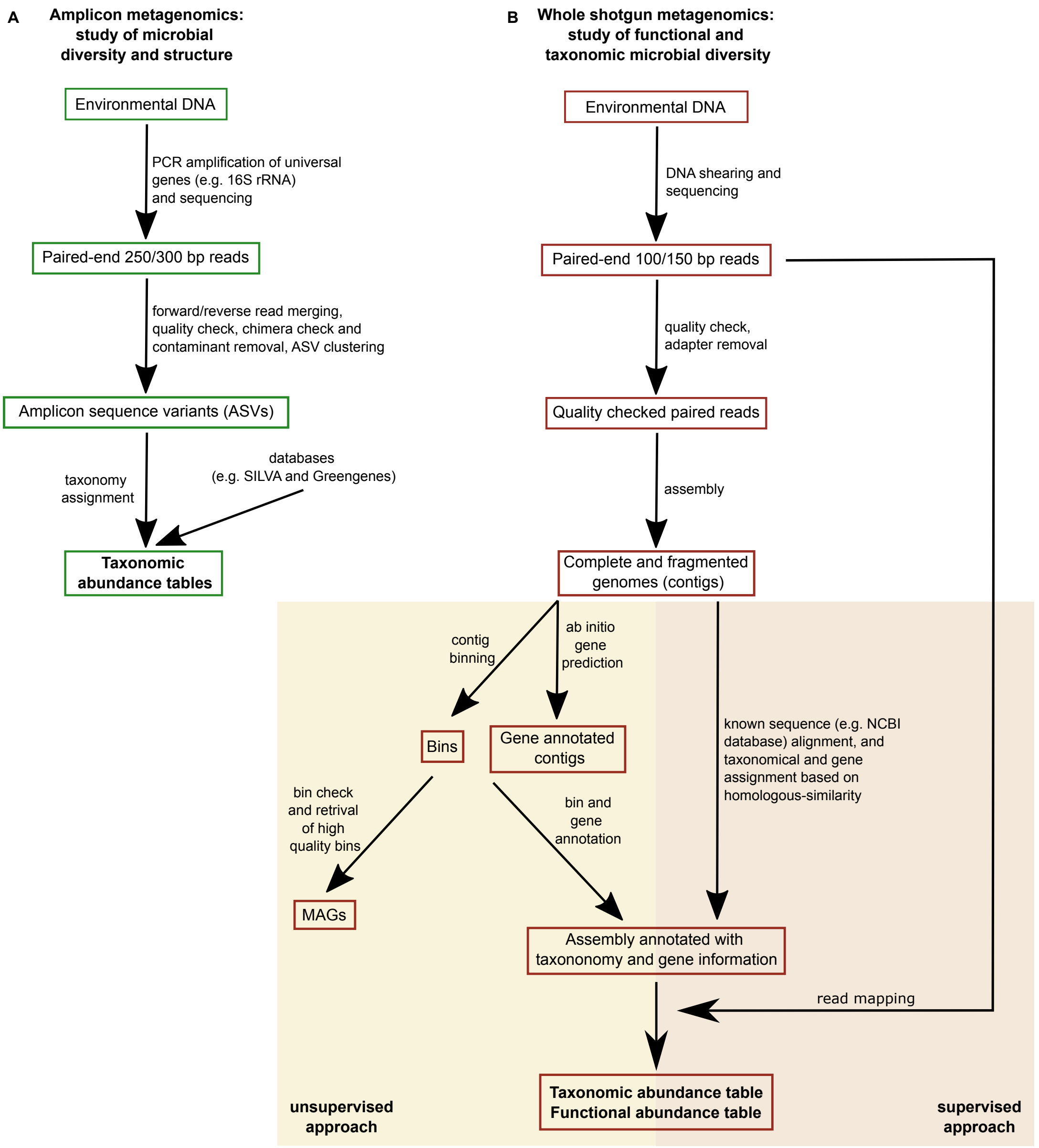

A
Amplicon metagenomics: study of microbial diversity and structure
Environmental DNA
PCR amplification of universal genes (e.g. 16S rRNA) and sequencing
Paired-end 250/300 bp reads
forward/reverse read merging, quality check, chimera check and contaminant removal, ASV clustering
Amplicon sequence variants (ASVs)
taxonomy assignment
databases (e.g. SILVA and Greengenes)
Taxonomic abundance tables
B
Whole shotgun metagenomics: study of functional and taxonomic microbial diversity
Environmental DNA
DNA shearing and sequencing
Paired-end 100/150 bp reads
quality check, adapter removal
Quality checked paired reads
assembly
Complete and fragmented genomes (contigs)
contig binning
ab initio gene prediction
known sequence (e.g. NCBI database) alignment, and taxonomical and gene assignment based on homologous-similarity
Bins
Gene annotated contigs
bin check and retrival of high quality bins
bin and gene annotation
MAGs
Assembly annotated with taxononomy and gene information
read mapping
Taxonomic abundance table
Functional abundance table
unsupervised approach
supervised approach

Figure 1.2: Metagenomic analysis common pipelines for amplicon metagenomic (A) and whole shotgun metagenomic (B) studies assuming paired-end Illumina reads are used.

1.4.1.1 Amplicon metagenomics

Amplicon metagenomics is the study of a microbial community through the amplification of a specific DNA fragment within the population. Specific DNA fragments can be amplified with a polymerase chain reaction (PCR) using the enzyme polymerase and specific targeting primers. The specificity of the primers to the gene of interest is essential as PCR amplification will start from where primers anneal to the DNA template. Potentially any gene and taxon can be amplified provided that its sequence is sufficiently divergent from the others in the community to allow primer specificity. Studying only a specific gene or taxon of a microbial community has the advantage of being cheaper and to give a clearer picture of that specific DNA fragment compared to the whole shotgun metagenomic approach where the information related to a specific gene may be lost, especially if not highly represented in the microbial community. The biggest disadvantage of this approach, though, is that the results can be skewed towards gene variances that are better amplified by the PCR reactions, bias that is not present in the case of the whole shotgun metagenome approach (Chen and Pachter, 2005; Ruijter et al., 2009). However this effect, can be mitigated when primers are correctly chosen or designed to target the genes and taxa of interest.

Genes have been used for targeting specific microbial groups, for examples, to identify nitrogen fixing bacteria (e.g. Garcias-Bonet et al., 2016), sulphur oxidizing producer bacteria (e.g. Zhao et al., 2017) or methanogenic archaea (e.g. Lever and Teske, 2015). Or even to target specific taxon inside a community for detection purposes (e.g. Yu et al., 2005; You and Kim, 2020).
However, most of the amplicon metagenomic studies amplify universal genes and regions (e.g. 16S rRNA gene, 18S rRNA gene and ITS region) for the characterization of microbial diversity and structure in a wide range of microbial communities (Banerji et al., 2018; Hibbett, 2016; Ju and Zhang, 2015; Triadó-Margarit and Casamayor, 2015). The ubiquity of these DNA regions in microbial genomes and the presence of regions with a high sequence variation between different species permit to obtain a picture of the microbial diversity in a specific environment.
The analysis of the amplicon data is not heavily computational demanding and most of the data analysis can be performed on portable devices. The input data are usually paired-end Illumina reads 250/300 bp long. These reads are usually merged (forward and reverse reads are merged together), quality checked, checked for chimeric sequences and contaminants. Until a few years ago the most common approach was then to use operational taxonomic units (OTUs) where the obtained sequences were merged if closely related (i.e. if they had a higher similarity than a set threshold, 97% usually). However, the OTU concept is being replaced by the amplicon sequence variants (ASVs) approach where only identical sequence are merged, giving therefore a higher taxonomic resolution (Callahan et al., 2017) (Figure 1.2A). ASVs (or OTUs) are then taxonomically assigned by alignment to specific and curated sequence databases, such as Greengenes (DeSantis et al., 2006) and SILVA (Yilmaz et al., 2014), which contain exclusively universal gene sequences (Figure 1.2A).
The data analysis can be performed with one of the many pipelines that have been developed, such as DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017), QIIME (Caporaso et al., 2010) and mothur (Schloss et al., 2009).

1.4.1.2 Whole shotgun metagenomics

Whole shotgun metagenomics is the study of a microbial community through the sequencing of the entire DNA content extracted from a sample. Whole shotgun metagenomic output are highly fragmented DNA sequences (usually paired-end 100/150 bp reads) that represent the whole microbial community, meaning that the fragmented microbial genomes from different organisms are all mixed together. Depending on data quality and study aims (e.g. taxonomic and functional profiling and genome reconstruction), sequencing data need to go thought several bioinformatics steps to obtain information (Figure 1.2B).

Taxonomic and functional characterization of the sequenced community can be obtained working on the short reads directly. However, taxonomic and functional assignment is more accurate on longer reads (Tamames et al., 2019) and therefore the usual approach is to assemble the reads into genomes, or due to the complex nature of microbial communities, into contigs (i.e. consensus DNA sequence created by overlapping DNA reads) and scaffolds (i.e. contigs connected by gaps of known base length). There are many tools that perform genome reconstruction but only few are optimized to work with metagenomic sequences, such as MEGAHIT (Li et al., 2015) and metaSPADES (Nurk et al., 2017). Metagenomic datasets set, in fact, many challenges for the assembly algorithms such as genome differential coverages and strain diversity (Sangwan et al., 2016; Vollmers et al., 2017).

In the last years, with the advent of third-generation sequencing technologies, the scenario has changed. Nanopore sequencing technology, thanks to its long reads, has highly improved assembly performances both for genome and metagenome reconstructions where the long Nanopore reads can be used either to scaffold fragmented Illumina assemblies, to perform an hybrid assembly along with Illumina reads, or to directly perform an assembly uniquely with Nanopore reads (Sohn and Nam, 2018; Ayling et al., 2020; Nicholls et al., 2019). The latter option, in particular, is usually followed by several rounds of read polishing where the sequence errors introduced by the error-prone Nanopore reads are corrected (Kono and Arakawa, 2019).

Once the assembly is obtained, taxonomy and genes can be assigned to the sequence with supervised and/or unsupervised approaches.

Taxonomy and gene assignment can be done with a supervised, similarity-based approach. This implies using alignment algorithms, such as BLAST (Altschul et al., 1990), for querying the DNA sequences (i.e. reads and contigs) with known proteins and genes and then assigning taxonomy and functional ontology depending on the attributes of the raw sequences. Known sequences can be found, for example, in publicly available databases such as GenBank (Sayers et al., 2019) and UniProt (Bateman, 2019). This approach is, for instance, used from MEGAN (Huson et al.,

2007) and MG-RAST (Glass and Meyer, 2011) where all the sequences that aligned to a DNA sequence are screened and assigned with Lowest Common Ancestor (LCA) algorithm where taxonomy is assigned to the lowest common ancestor shared among the aligned sequences (Huson et al., 2007).

Other tools perform taxonomy and gene assignment with an unsupervised approach, screening DNA sequences for specific base pair patterns. For example, binning software, such as CONCOCT (Alneberg et al., 2014) and MetaBAT (Kang et al., 2019), can bin and separate the sequences into different taxonomic clusters (i.e. genomes) by looking at sequence tri- and tetranucleotide composition and the differential read coverage. Ab initio gene prediction is obtained by looking for Open Reading Frames (ORFs) with software such as Prokka (Seemann, 2014) and Glimmer (Kelley et al., 2012).

The unsupervised approach is successfully used when the assembly is high quality and ORFs and taxonomic bins, which rely on sequence accuracy, can be successfully be predicted. On the contrary, supervised approaches are more tolerant towards sequence errors but functional and taxonomic assignment will only rely on known homologous proteins and genes.

In order to quantify the gene and taxon proportions and relative abundances usually the initial reads for each samples are mapped back to the annotated assembly with mapping software such as bowtie2 (Langmead and Salzberg, 2012) and bwa (Li and Durbin, 2010).

Further, with deep-coverage whole shotgun metagenomic dataset, it is possible to obtain complete or partially complete metagenome assembled genomes (MAGs). Not relying exclusively on the sequencing of microbial cultures and isolates to obtain genome sequences has been giving a boost to the understanding of diversity and to the biodiscovery and bioprospecting field (Roumpeka et al., 2017). Sure enough, the construction of MAGs does not rely on known databases and therefore allows the finding of new organisms, and the new genes and biosynthetic gene clusters associated to those.

The first almost complete genomes obtained from environmental samples were in 2004 (Tyson GW et al., 2004). Since then, thanks to the advent of long sequence technologies, the recovery of complete metagenome assembled genomes (cMAGs), circularized (in case of prokaryotes) and ungapped complete genome, has become increasingly easier. Especially powerful are the approaches that combine both short and high-quality Illumina reads and long error-prone Nanopore reads (Giguere et al., 2020; De Maio et al., 2019). Thanks to the use of both these technologies, for example, Somerville et al. were able to retrieve strain-level MAGs (2019).

Whereas cMAGs represent complete and ungapped genomes, non-complete MAGs are defined from the taxonomic bins where contigs belonging to one organism are grouped together. These

MAGs represent draft genomes and they need bioinformatics validation in order to assess whether they are representative of an organism (Olson et al., 2018; Chen et al., 2020; Shaiber and Eren, 2019). This can be obtained with bioinformatics software such as CheckM (Parks et al., 2015) and BUSCO (Simão et al., 2015), which can assess the MAG quality by looking for single-copy genes, which are genes that are expected to be found in single copy in all genomes, in the MAG contigs. Looking at the number of single-copy genes in a MAG, these software can estimate the MAG completeness, contamination and strain heterogeneity.

1.4.2 Metatranscriptomics

Metatranscriptomics is the study of RNA extracted from an environmental sample (environmental RNA or eRNA). Whereas metagenomic data represent all the organisms (e.g. dead, dormant and alive), metatranscriptomic data give information only on alive organisms that were actively transcribing RNA (e.g. rRNA and mRNA) when the environmental sample was collected. As described in the previous sections, microorganisms can enter and exit different metabolic states of dormancy in relation to the environmental conditions, where also a variable portion of microorganisms are dead (Lennon and Jones, 2011).

The direct characterization of the transcribed rRNA can, for example, unravel which organisms were effectively active at the sample collection, whereas, the study of mRNA can unravel in which environmental processes the organisms are actively involved (Blazewicz et al., 2013). RNA data can also help, for example, to understand organism responses to environmental conditions (Moran et al., 2013; Carvalhais et al., 2012).

Metatranscriptomic data is obtained through the direct sequencing of RNA, or the more stable complementary DNA (cDNA) which is reversed transcribed from the RNA molecules. The use of RNA in molecular analysis is quite difficult and therefore less used because of the higher degradability of the RNA and its difficult preservation (Cristescu, 2019). Furthermore, its sampling is difficult in environmental settings because it is hard to preserve the transcriptional picture of the time when the sample was collected (McCarthy et al., 2015).

Using both metagenomic and metatranscriptomic data can be extremely useful in order to explore both potentially active (i.e. DNA) and active (i.e. RNA) organisms, to explore the microbial potential and effective functionality, and also to assign coding regions to reconstructed genomes (Pochon et al., 2017).

1.5 Thesis aims and outline

The overall aims of this thesis are i) to present different sequencing technologies and bioinformatics approaches for the analyses of environmental microbial communities, ii) to explore several complex and original Arctic environmental sequencing datasets from glaciers, proglacial forefields and permafrost, and iii) to publicly provide new tools and software for ease bioinformatics analyses.

In Chapter 2, I present PhyloPrimer, an online user-friendly and semi-automated platform to design PCR primers and probes for specific taxonomic and functional targets in environmental samples. The tool is suitable to perform detection of specific taxa from environmental samples, testing also for oligo microbial specificity, secondary structure formation and other oligo specifics. In this chapter I show how I implemented this tool and how it can be applied for detection studies.

In Chapter 3, I used an amplicon (16S rRNA gene) metagenomic approach to unravel the taxonomic diversity and structure of englacial channel microbial communities through comparison with those living in meteoric ice. The main aim is to assess the role of the glacial water channels in the microbial dispersion within the glacial realm.

Chapter 4 presents a complex sequencing dataset where the proglacial systems of the Midtre Lovénbreen (Svalbard), Storglaciären (Sweden) and the Greenland ice sheet in proximity of point 601 were sequenced with a whole genome shotgun metagenomic approach. This dataset was analyzed with a new pipeline, LongMeta, that allows to answer questions such as 'Which genes a specific taxonomic group could potentially express?', giving gene*taxon resolution to the metagenomic data. Thanks to this pipeline, I explored several taxonomic trends across these forefield microbial successions. Further, I explored rock weathering and nitrogen fixation genes and how they are distributed between and within several organisms and in a microbial succession. This approach also allowed to assign these functions to organisms never associated with these functions before.

In Chapter 5, I characterized microbial successions in the frozen soil of the Greenland ice sheet proglacial system. Thanks to a hybrid approach where I used both Illumina (i.e. DNA) and Nanopore (i.e. DNA and cDNA) sequences to construct a high reliable deep coverage assembly, I was also able to reconstruct several microbial MAGs and to construct the cold-adapted predicted protein (CAPP) database reporting cold-adapted proteins that could be potentially used for an informed bioengineering approach for enzyme modification.

In Chapter 6 I wrap up the thesis and bring all the different approaches and findings together.

Chapter 2

PhyloPrimer: a taxon-specific oligonucleotide design platform

Abstract

Many environmental and biomedical biomonitoring and detection studies aim to explore the presence of specific organisms or gene functionalities in microbiome samples. In such cases, when the study hypotheses can be answered with the exploration of a small number of genes, a targeted PCR-approach is appropriate. However, due to the complexity of environmental microbial communities, the design of specific primers is challenging and can lead to non-specific results. I designed PhyloPrimer, the first user-friendly platform to semi-automate the design of taxon-specific oligos (i.e. PCR primers and probes) for a gene of interest. The main strength of PhyloPrimer is the ability to retrieve and align GenBank gene sequences matching the user's input, and to explore their relationships through an online dynamic tree. PhyloPrimer then designs oligos specific to the gene sequences selected from the tree and uses the tree non-selected sequences to look for and maximize oligo differences between targeted and non-targeted sequences. Designed oligos are then checked for the presence of secondary structure with the nearest-neighbor (NN) calculation and the presence of off-target matches with *in silico* PCR tests, also processing oligos with degenerate bases. Whilst the main function of PhyloPrimer is the design of taxon-specific oligos, the software can also be used for designing oligos to target a gene without any taxonomic specificity, for designing oligos from preselected sequences and for checking predesigned oligos. I have validated the pipeline on four commercially available microbial mock communities using PhyloPrimer to design genus- and species- specific primers for the detection of *Streptococcus* species in the mock communities. The software performed well on these mock microbial communities and can be found at www.cerealsdb.uk.net/cerealgenomics/phyloprimer.

2.1 Introduction

The Polymerase Chain Reaction (PCR) is a pivotal technique to many molecular protocols and is widely used to exponentially amplify a specific portion of DNA (e.g. gene) using DNA template (e.g. the entire DNA content of an environmental sample), primers, deoxynucleotides (dNTPs), DNA polymerase and reaction buffers (Garibyan and Avashia, 2013). Before starting with any PCR-based procedure, primers and probes (if performing a probe-based qPCR) need to be selected to target the specific DNA region and organisms object of the study. The amplification starts where the primers anneal to the DNA template, for this reason the specificity of the PCR reaction is highly impacted by the specificity of the primers to the DNA template.

The design of new oligonucleotides (i.e. primers or probes), hereafter abbreviated as oligos, is a relatively easy task when working with known axenic cultures or known low complexity communities but can be challenging when dealing with unknown organisms and complex environment communities. Different studies can require different level of oligo-specificity: oligos could be designed to target the same DNA portion in all the community organisms (e.g. universal primers), in a specific group of organisms or in a specific species or strain. The latter two tasks become challenging when the target DNA fragment is present in non-target organisms that are part of the community (Fierer et al., 2005).

Many different primer and probe sequences have been published. These oligos can target a broad variety of different DNA sequences and can present a wide range of target organism's specificity. Universal oligos, such as primers targeting conserved regions of the 16S rRNA gene (e.g. Takahashi et al., 2014), are widely used for the study of microbial diversity. It is also possible to target non-universal genes, such as the nifH gene (e.g. Gaby and Buckley, 2012) and the pmoA gene (e.g. Wang et al., 2017), in order to target only organisms with a specific metabolism and that occupy specific environmental niches. Oligos can also have a more specific target: they can amplify only genes present in organisms of interest even when the gene is present in a wider selection of organisms (e.g. Yu et al., 2005; You and Kim, 2020).

When no predesigned oligos are available, however, it is necessary to develop new ones. Oligo sensitivity is a trade-off between the specificity of the oligo to the DNA template and allowing some oligo-template mismatch if targeting different organisms in order to get an even coverage of all the representative organisms (Parada et al., 2016). Depending on the user needs, there are many web-tools and software freely available for the oligo design. The most widely used tools for primer design are Primer3 and its web interface Primer3Plus (Untergasser et al., 2007; Untergasser et al., 2012), Oligo7 (Rychlik, 2007) and Primeclade (Gadberry et al., 2005). To target unknown genes where only the protein or related gene sequences are known, it is necessary to design degenerate oligos. The latter take advantage of the codon degeneracy property of the amino acid sequences and, having degenerate bases in their sequences, represent a pool of unique primers that target the same amino acid coding sequence. Primer designing tools for degenerate primers can require the input of proteins, such as CODEHOP (Boyce et al., 2009; Rose et al., 2003) or Primer Premier (Singh et al., 1998); or the input of DNA sequences or alignments, such as DegePrimer (Hugerth et al., 2014), HYDEN (Linhart and Shamir, 2005) or FAS-DPD (Iserte et al., 2013).

Environmental communities pose many challenges for the oligo specificity as we often do not know which are the community organisms and therefore it is difficult to foresee the possible nonspecific products (Deiner et al., 2017; Morales and Holben, 2009). *In silico* PCR is an essential step towards the design of specific oligos (Yu and Zhang, 2011). Some of the commonly used tools are UCSC In-Silico PCR (Kent et al., 2002), FastPCR (Kalendar et al., 2009) and Primer-BLAST (Untergasser et al., 2012). The latter allows to check the oligo specificity against the comprehensive NCBI databases (Sayers et al., 2020). Further to their taxonomic specificity, oligos need to be tested for different parameters, such as the absence of homopolymer regions or di-nucleotide repetitions and the presence of a GC clamp (Elbrecht et al., 2018). Primers must also be scanned for the presence of secondary structures such as self-dimers, cross-dimers and hairpins (Chuang et al., 2013) in addition to characteristics of the targeted organisms. For instance, prokaryotic genomes rarely have introns as gene splicing is rare in these organisms (Sorek and Cossart, 2010), whereas introns and multiple splicing sites are widely present in eukaryotic genomes and must be taken in consideration when designing primers (Goel et al., 2013; Shafee and Lowe, 2018).

In case the PCR target is a gene possessed only by a specific organism, the primers can be designed directly on that gene sequence. If more than one gene variant needs to be amplified (e.g. multiple species are targeted), a consensus sequence can be calculated and the oligos can then be designed on it (consensus primers). A consensus sequence is created from a sequence alignment and is defined as a sequence that reports the most frequent base present in the alignment in each position. The construction of this sequence, and consequently the designed oligos, are greatly influenced by the selection of the initial sequences. This pivotal step is usually not implemented in the oligo design software as these require the upload of preselected sequences. To date, only ARB implemented a toolkit that allows the creation of new primers and probes on sequences selected by the ARB phylogenetic tree of ribosomal sequences (Essinger et al., 2015; Ludwig et al., 2004). However, in order to work on other DNA portions, the user needs to create a sequence database to import inside the software.

Therefore, prior to the oligo design, the user has to retrieve the sequences of interest from a database (e.g. NCBI database), making sure that the sequences represent the DNA portion of interest and that they cover the same sequence fragment. This process can be complex and time-consuming especially when working with environmental microbial communities or working with a ubiquitous and divergent gene.

In this chapter, I present PhyloPrimer, a user-friendly and comprehensive online platform to i) select the DNA sequences to use for oligo design, ii) construct a consensus sequence, iii) design microbial oligos (i.e. primers and probes), iv) test for oligo specificity through *in silico* tests and v) test the oligos for the presence of secondary structures with the nearest-neighbour (NN) model for nucleic acids. In addition to provide a unique platform to check oligos (i.e. primer pairs, primer and probe assays, and single oligos) for both secondary structure and non-specific targets, also processing oligos with degenerate bases, the real strength of PhyloPrimer is the DNA sequence selection where the user can explore the diversity of the sequence of interest through a dynamic phylogenetic tree.

PhyloPrimer was born from a collaboration between myself and Dr. Cédric Malandain, working at HYDREKA ENOVEO (a bioremediation company in Lyon, France). This company is involved in several projects that aim to study the microbial communities of polluted sites, to detect specific-pollutant degrading microorganisms that could conduct a biological-driven remediation of the sites, and to understand if other approaches (e.g. chemical oxidation) are necessary (http://enoveo.com/en/services/polluted-sites-and-soils/). The detection of microorganisms able to degrade site-specific pollutants is conducted by targeted qPCRs and, because of the broad variety and diversity of the polluted sites and microorganisms, the design of highly specific oligos is required for different projects. For this reason, we collaborated with Dr. Cédric Malandain to develop a new user-friendly and comprehensive platform that could provide a fast and reproducible pipeline for the design of new oligos.
This chapter presents the tool implementation and how it was used to design primers for mock microbial communities.

2.2 Implementation

PhyloPrimer runs on a remote server provided from the University of Bristol. The current server has 48 CPUs (64-bit Intel(R) Xeon(R) CPU E5-2680 v3 at 2.50GHz). Only 6 PhyloPrimer processes at one time are allowed on the server, the excess processes enter a queue. On average, the oligo design requires 30/40 minutes whereas the oligo check requires 5/10 minutes. The web interface was implemented in HTML and JavaScript. PhyloPrimer is coded in Perl, JavaScript, HTML, CSS and mySQL. Two javascript packages were used: a modified version of PhyloCanvas v 1.7.3 (phylocanvas.org) and canvasJS v 2.3.2 (canvasjs.com). The user can access PhyloPrimer through a web platform at www.cerealsdb.uk.net/cerealgenomics/phyloprimer. PhyloPrimer was tested and implemented using the Safari, Firefox and Chrome browsers. The website uses general data protection regulation (GDPR) cookie acceptance box on the first use.

2.2.1 The web platform

The PhyloPrimer web platform is structured with sequential web pages that can be categorized into four different groups: i) the home page, ii) the input pages, iii) the oligo pages and iv) the result page. From the home page, the user can select one of the three different input pages available for uploading the data (e.g. DNA sequences, DNA alignments and newick tree) where each page corresponds to a different modality to use PhyloPrimer. Once the data are uploaded, the user is redirected to the oligo pages where there are different parameter settings to design either primer assays, primer and probe assays or single oligos. Once the user submits these parameters, the oligo design and the oligo check are performed on the web server. As soon as PhyloPrimer has finished the analyses, the user receives an email with a link to the result page where the user can explore the designed oligos and choose the ones will be used for future work (Figure 2.1).

This software can be used in three different modalities. It can be used to design oligos from DNA sequences interactively selected from a dynamic phylogenetic tree (Dynamic Selection; Figure 2.2A), to design oligos from preselected DNA sequences (Premade Selection; Figure 2.2B) and to check predesigned oligos (Oligo Check; Figure 2.2C).

2.2.1.1 Dynamic Selection

This mode is the strength of PhyloPrimer and was developed to facilitate the selection and retrieval of NCBI sequences to be then used for the oligo design. Therefore this modality should be chosen when the user wants to select DNA sequences, similar to a gene or a DNA fragment of interest, that will be aligned by PhyloPrimer to calculate the consensus sequence from which the oligos will be then designed. In this page, the user can upload either up to ten genes/DNA regions of interest, up to 500 DNA sequences or a newick tree (with a maximum of 500 tree leaves) (Figure 2.2A):

- Up to 10 DNA sequences. It is essential that the uploaded sequences represent the same DNA portion (e.g. same gene or gene fragment belonging to different organisms) for a smooth phylogenetic tree construction and visualization. For this reason, PhyloPrimer also performs a multiple sequence alignment with MAFFT v 7.31 (Katoh and Standley, 2013), with parameters `--localpair --maxiterate 1000`, on the user sequences to check how many gaps are in the alignments, using it as a proxy for checking if the same gene fragment was uploaded, if more than the 50% of any of the sequence alignment was constituted by gaps, the software will produce a warning.

- Up to 500 DNA sequences. In addition to the fasta file, the user can upload two optional files to allow gene and taxonomy visualization in the dynamic tree: a first file reporting the sequence names present in the fasta file and the associated taxid; the second file reporting the sequence names and the correspondent gene information.
- A phylogenetic tree in newick format with a maximum of 500 tree leaves. PhyloPrimer constructed trees are not accurate for phylogenetic visualization purposes but provide a

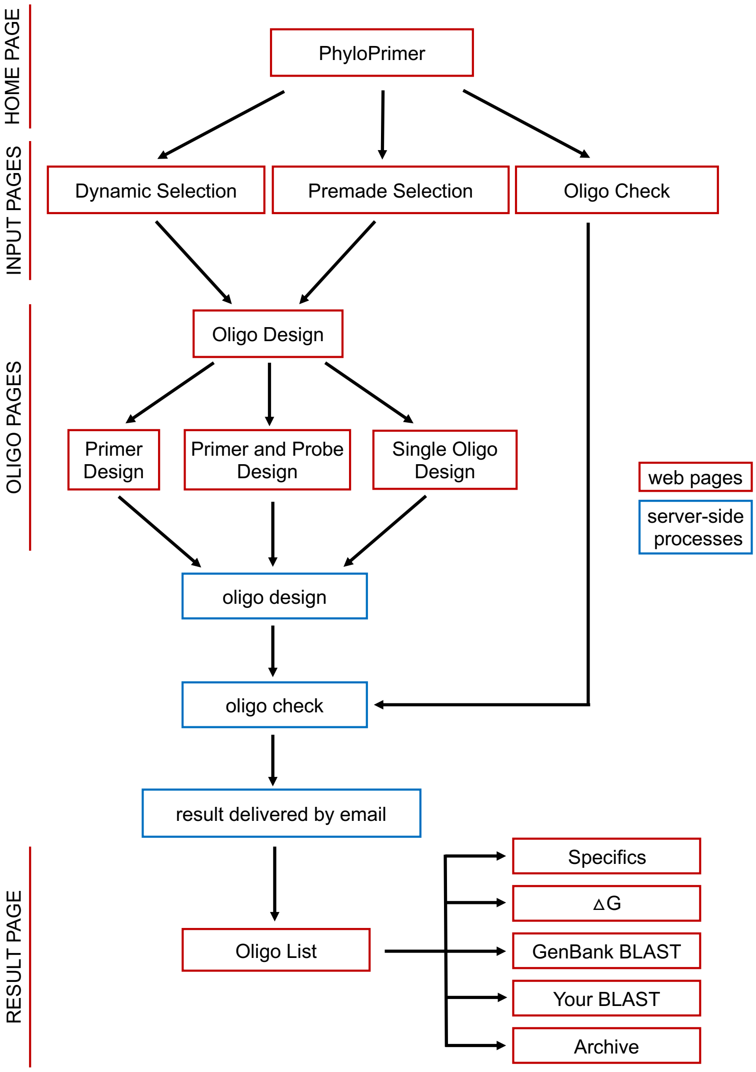


Figure 2.1: PhyloPrimer structure indicating the web pages and the server-side processes.

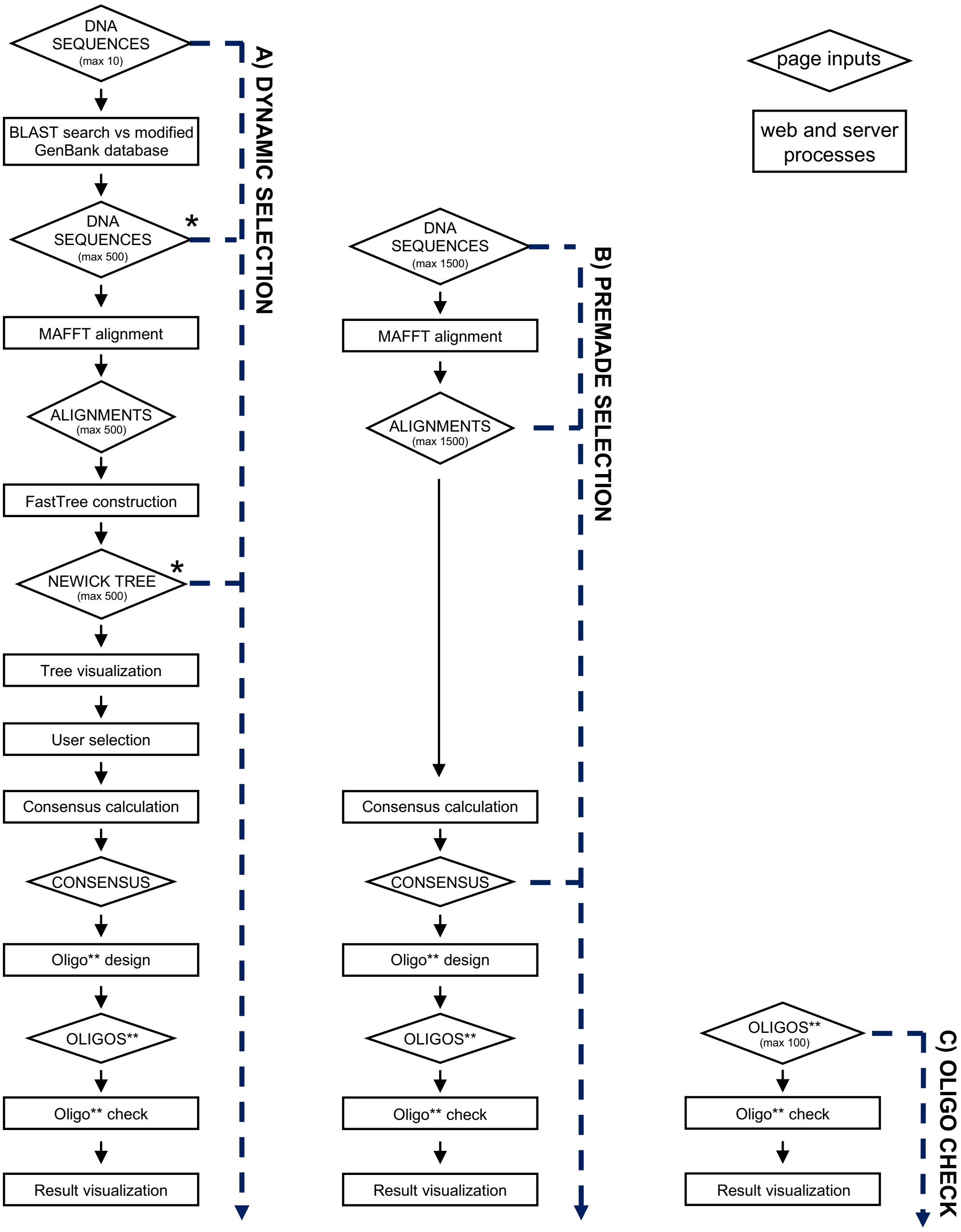


Figure 2.2: Detailed scheme of the three different input pages and workflows: Dynamic Selection (A), Premade Selection (B) and Oligo Check (C). Through the Dynamics Selection page the user can input three different kind of data: up to 10 genes or DNA regions of interest, up to 500 DNA sequences and a newick tree (together with an alignment file). The Premade Selection page permits the uploading of up to 1500 DNA sequences, 1500 DNA alignments or directly the consensus sequence that will be used for the oligo design. In the Oligo Check page only the upload of predesigned oligos is allowed. Different processes on the server-side of PhyloPrimer will start in relation to which data was uploaded. *There can be optional input files for taxonomy and protein information. If a newick file is the input, an additional alignment file must be uploaded. ** Oligos are intended as primers pairs, primer pairs plus a probe or single oligos.

good sequence clustering to explore the sequence similarity. If the user wishes to look at more accurate trees, it is possible to directly upload a newick tree (including of bootstrap values). In addition to the newick file, the user must upload the alignment file that was used for the phylogenetic tree construction and can upload the two optional files described above.

Once the input data are uploaded, PhyloPrimer checks if they are in the correct format and respect all the prerequisites. If everything is correct, PhyloPrimer will indicate the number of uploaded DNA sequences or tree leaves and will start different processes depending on which kind of input was uploaded.

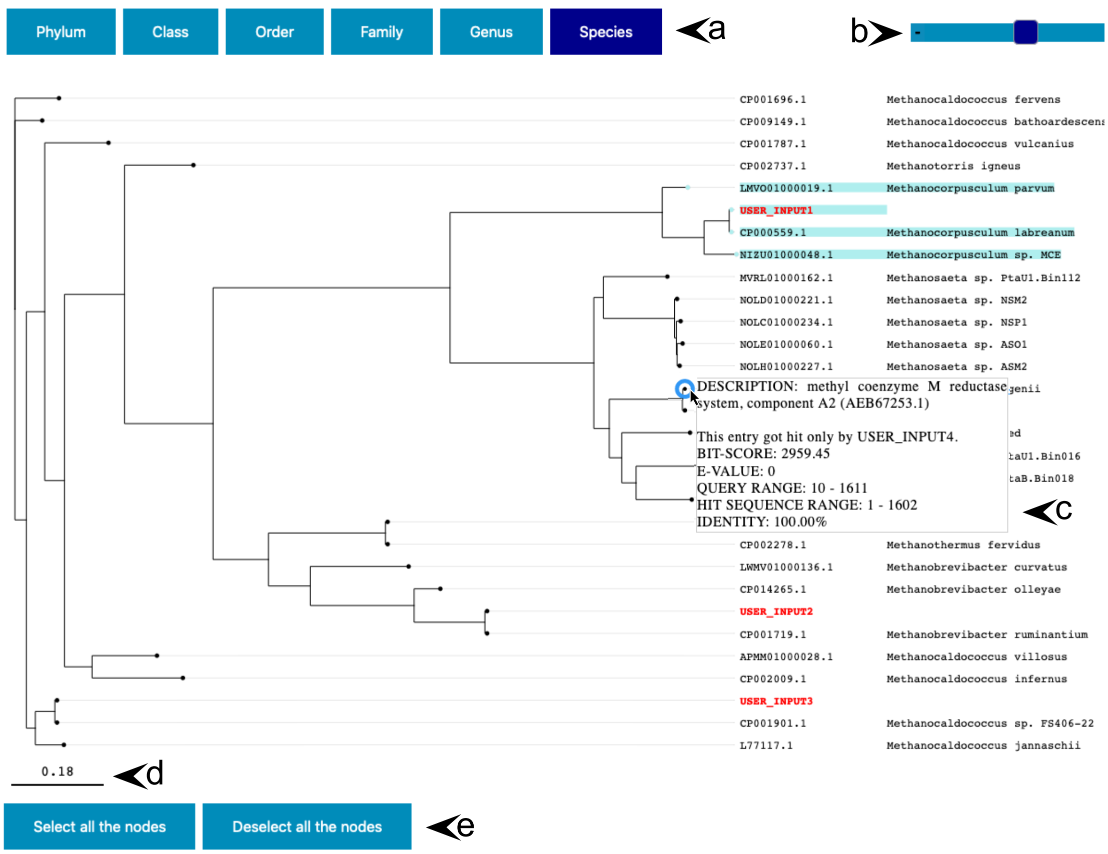
If the DNA sequences were uploaded in the first input area (up to 10 DNA sequences), the user is given the option to modify three BLAST parameters: i) identity percentage (i.e. the percentage of bases shared between the query and the subject sequence), coverage percentage (i.e. the percentage of bases of the query sequence that are covered by the subject sequence) and E-value (i.e. the probability of finding an alignment by chance). Once these parameters are customized (if needed), PhyloPrimer launches a MegaBLAST search with BLAST v 2.6.0+ (Baxevanis, 2020; Morgulis et al., 2008). The BLAST search is performed against a modified GenBank database (Sayers et al., 2019) comprehensive of gene sequences representing the entire prokaryotic domain (more details in Section 2.2.5). The BLAST command line is: `blastn -db GenBank -query input -task megablast -max_target_seqs 2500 -perc_identity identity -q cov_hsp_perc coverage -evalue evalue -outfmt 5 -num_threads 35`, where `evalue`, `identity` and `coverage` can be modified by the user. The modification of these parameters directly influences how many sequences are retrieved from the PhyloPrimer database. Therefore, setting the correct identity and coverage is very important for the sequence retrieval and a clear tree visualization. For examples, a low identity score may retrieve genes that do not belong to the gene family of interest, whereas a low coverage may retrieve only gene fragments, excluding the gene region to amplify. The BLAST search retrieves up to 2500 matches per user sequence. Once the BLAST search has terminated, PhyloPrimer groups together the sequences, both the user and BLAST retrieved sequences, that are represented by more than four entries. Once these clusters have been formed, PhyloPrimer will select only the best 500 matches for the alignment and tree construction. In case more than one gene was uploaded, per each gene the software will retrieve a maximum of unique BLAST matches corresponding to 500 divided by the number of genes. For example, if the user uploaded 5 genes, maximum 100 unique sequences per gene will be visualized on the phylogenetic tree. The limit of 500 tree nodes is arbitrary and is to allow a smooth dynamic tree visualization.

If there are at least four unique sequences (4 sequences is the minimum threshold usually set by aligner tools), the software performs a MAFFT alignment on all the sequences. MAFFT is run with the parameter `--maxiterate 1000` and, in case less than 200 sequences were retrieved, the flag `--localpair` which performs a more accurate alignment. The multiple alignment is then the input for FastTreeMP v 2.1 which infers approximately-maximum-likelihood phylogenetic trees with a GTR+CAT model, where the command is run with the parameters `-nt -gtr -boot 1000` (Price et al., 2009).
The whole process (i.e. BLAST search, MAFFT alignment and tree construction) usually takes between 1 and 5 minutes.

In case the second input option was used (up to 500 DNA sequences), PhyloPrimer will directly run a MAFFT alignment on the uploaded sequences and then construct a tree with FastTree with the same command parameters showed above. Whereas if the newick tree was uploaded, PhyloPrimer directly plots the tree.

The dynamic tree can be used for the dynamic exploration of the user and the GenBank sequences (in case the user uploaded the data in the first input area). Any node can be clicked to select/deselect the terminal leaves which report the sequence header and, where available, the sequence taxonomy. The selected sequences will appear blue-highlighted. Taxonomy rank information can be changed by the the apposite buttons (e.g. Genus, Family, Order, etc). Details on the sequence of the BLAST search can be visualized by hovering on the terminal nodes. The user sequences are reported in the tree in bold red characters (Figure 2.3). The phylogenetic tree is followed by a table reporting the correspondences between the PhyloPrimer name (reported in the tree) and the user original name. If PhyloPrimer found clusters composed by more than four identical sequences, it will group their entries together and will include the cluster details in a table reporting the cluster name, the number of sequences clustered together, the sequence headers and taxonomy information at the species level.

Once the sequences of interest have been selected from the tree, the button Create Consensus will trigger the consensus sequence calculation. If all the sequences were selected from the tree, only one consensus (i.e. positive consensus) is calculated. On the contrary, if only certain sequences were selected, two consensus sequences are calculated. The positive consensus is calculated by only the sequences that were selected by the dynamic tree and the negative consensus by the sequences that were not selected. All the oligos will be designed uniquely on the positive consensus. The negative will be only used to look for and maximize primer differences between the two consensus sequences. (Figure 2.4).
Once constructed, PhyloPrimer reports the consensus sequence and calculates how many degen-


Phylum
Class
Order
Family
Genus
Species
a
b
CP001696.1 Methanocaldococcus fervens
CP009149.1 Methanocaldococcus bathoardescens
CP001787.1 Methanocaldococcus vulcanius
CP002737.1 Methanotorris igneus
LMVO01000019.1 Methanocorpusculum parvum
USER_INPUT1
CP000559.1 Methanocorpusculum labreanum
NIZU01000048.1 Methanocorpusculum sp. MCE
MVRL01000162.1 Methanosaeta sp. PtaU1.Bin112
NOLD01000221.1 Methanosaeta sp. NSM2
NOLC01000234.1 Methanosaeta sp. NSP1
NOLE01000060.1 Methanosaeta sp. ASO1
NOLH01000227.1 Methanosaeta sp. ASM2
c
CP002278.1 Methanothermus fervidus
LWMV01000136.1 Methanobrevibacter curvatus
CP014265.1 Methanobrevibacter olleyae
USER_INPUT2
CP001719.1 Methanobrevibacter ruminantium
APMM01000028.1 Methanocaldococcus villosus
CP002009.1 Methanocaldococcus infernus
USER_INPUT3
CP001901.1 Methanocaldococcus sp. FS406-22
L77117.1 Methanocaldococcus jannaschii
0.18
d
Select all the nodes
Deselect all the nodes
e

DESCRIPTION: methyl coenzyme M reductase system, component A2 (AEB67253.1)

This entry got hit only by USER_INPUT4.
BIT-SCORE: 2959.45
E-VALUE: 0
QUERY RANGE: 10 - 1611
HIT SEQUENCE RANGE: 1 - 1602
IDENTITY: 100.00%

Figure 2.3: Screenshot of the dynamic tree that can be visualized in the Dynamic Selection page. The represented tree was constructed uploading 5 sequences coding for the gene methyl coenzyme M reductase (bold red). The selected entries (highlighted in light blue) in the dynamic tree correspond all to sequences assigned to the genus *Methanocorpusculum*. PhyloPrimer will construct the consensus sequence for the oligo design considering only the four sequences highlighted in the dynamic tree. The different elements are a) taxonomy buttons, b) zoom bar, c) information box related to the hovered terminal node, d) scale bar and e) node select/deselect buttons.

erate bases were included in the positive consensus and will suggest the maximum number of degenerate bases to include in the oligo sequence (this is an oligo design parameter that can be set in the oligo pages): it will suggest to allow 3 degenerate bases in the oligo sequence if more than the 20% of the consensus is constituted by degenerate bases, 2 bases if the degenerate bases are more than 10%, otherwise PhyloPrimer will suggest 1. If more than the 20% of the bases in the positive consensus were degenerate bases, PhyloPrimer will report a warning saying that the user may have to design several oligo assays in order to target all the selected sequences (selecting fewer sequences from the different tree clusters). This may happen when the selected sequences do not have a conserved region (more details in Section 2.2.2).

2.2.1.2 Premade Selection

The Premade Selection page is used when the user has already selected the sequences PhyloPrimer must use for the consensus sequence calculation and the oligo design. Therefore no

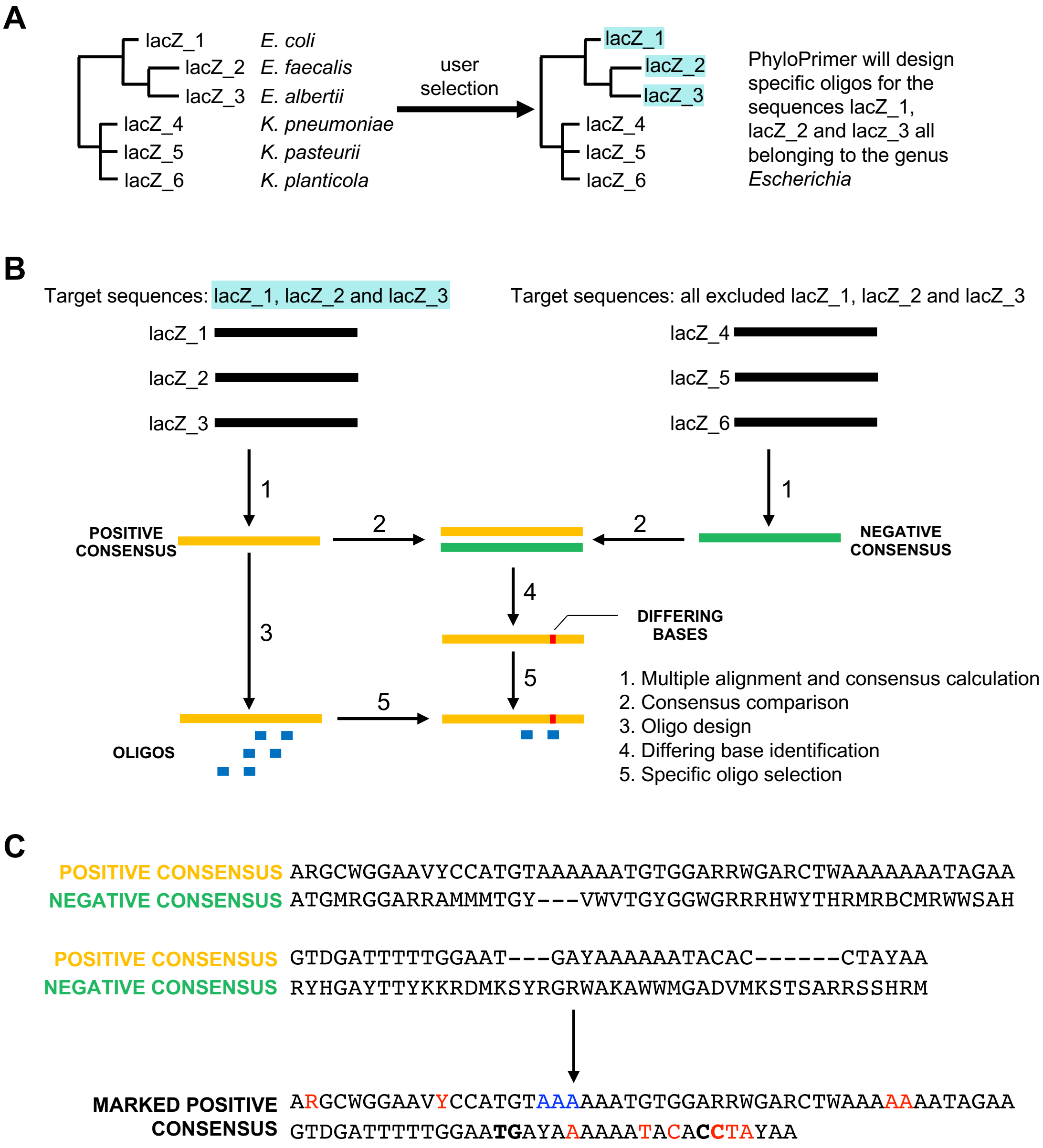

A
lacZ_1
lacZ_2
lacZ_3
lacZ_4
lacZ_5
lacZ_6
E. coli
E. faecalis
E. albertii
K. pneumoniae
K. pasteurii
K. planticola
user selection
lacZ_1
lacZ_2
lacZ_3
lacZ_4
lacZ_5
lacZ_6
PhyloPrimer will design specific oligos for the sequences lacZ_1, lacZ_2 and lacz_3 all belonging to the genus Escherichia
B
Target sequences: lacZ_1, lacZ_2 and lacZ_3
Target sequences: all excluded lacZ_1, lacZ_2 and lacZ_3
lacZ_1
lacZ_2
lacZ_3
lacZ_4
lacZ_5
lacZ_6
1
1
POSITIVE CONSENSUS
2
2
NEGATIVE CONSENSUS
4
DIFFERING BASES
3
5
5
OLIGOS
1. Multiple alignment and consensus calculation
2. Consensus comparison
3. Oligo design
4. Differing base identification
5. Specific oligo selection
C
POSITIVE CONSENSUS ARGCWGGAAVYCCATGTAAAAAATGTGGARRWGARCTWAAAAAAAATAGAA
NEGATIVE CONSENSUS ATGMRGGARRAMMMTGY---VWVTGYGGWGRRRHWYTHRMRBCMRWWSAH
POSITIVE CONSENSUS GTDGATTTTTGGAAT---GAYAAAAAATACAC------CTAYAA
NEGATIVE CONSENSUS RYHGAYTTYKKRDMKSYRGRWAKAWWMGADVMKSTSARRSSHRM
MARKED POSITIVE CONSENSUS ARGCWGGAAVYCCATGTAAAAAATGTGGARRWGARCTWAAAAAAAATAGAA
GTDGATTTTTGGAATGAYAAAAAATACACCTAYAA

Figure 2.4: PhyloPrimer positive and negative consensus workflow. In the Dynamic Selection mode, the consensus design starts with the selection of DNA sequences on the phylogenetic tree (A). Successively, the selected sequences (i.e. lacZ_1, lacZ_2 and lacZ_3) are used to construct the positive consensus and the others are used to calculate the negative consensus (i.e. lacZ_4, lacZ_5 and lacZ_6). The two consensus sequences are compared, positive consensus areas with differing bases are identified and oligos specific to that area are selected (B). In any consensus visualization of the consensus sequence PhyloPrimer is then going to report only the positive consensus with marked letters where differing with the negative sequence (C). The base color code is as follows: red letters indicate positions where the two sequences presented differing bases, blue letters indicate positions where there are bases on the positive consensus but gaps in the negative and bold letters flank regions where there were gaps on the positive consensus but bases on the negative consensus. A degenerate base is marked as differing only if that base does not contain the correspondent base of the negative consensus.

dynamic tree visualization is available through this page as there is no need to explore closely related sequences. The three possible inputs are either up to 1500 DNA sequences, up to 1500 DNA alignments or a consensus sequence (Figure 2.2B).

As in the Dynamic Selection page, all the inputs will be first checked to have the correct format. In case more than one DNA sequence was uploaded, PhyloPrimer will first align all the DNA sequences with MAFFT (using the same commands reported in the previous section). Consecutively it will calculate a consensus sequence that will then be used for the oligo design. In case alignment sequences were uploaded, PhyloPrimer will construct a consensus sequence from them (Figure 2.2B). In case a consensus sequence was uploaded, it will directly be used for the oligo design.

In the case on the Premade Selection workflow PhyloPrimer will have only one consensus sequence (i.e. positive consensus) as no negative reference sequences are present as in the case of the phylogenetic tree. As in the last page, the consensus sequence will be printed on the page and a maximum number of degenerate bases will be suggested for the oligo design.

2.2.1.3 Oligo Check

The Oligo Check page only allows the input of oligo sequences that the user wants to be checked from PhyloPrimer (Figure 2.2C). The oligos will be checked for secondary structures, melting temperature and *in silico* PCR targeted organisms. These are the same checks that the oligos designed by PhyloPrimer undergo. The oligos can be uploaded as a maximum of 100 primer pairs, primer pair/probe assays or single oligos. It is possible to customize parameters that will be used for the melting temperature and secondary structure calculation and *in silico* PCR taxonomy tests.

2.2.1.4 Oligo pages

At this point all the user inputs have already been uploaded and the positive consensus (and negative if needed) has already been constructed by, or imported to, PhyloPrimer. Depending on which kind of oligos must be designed, I devised three different web pages: Primer Design, Primer and Probe Design and Single Oligo Design. From these the user can customize specific oligo and PCR parameters essential to a successful oligo design. Each of the oligo page is structured in the same way:

- The first section asks the user for a job ID and the email address. The latter is the only mandatory field as all the results will be received by email. All the results and sensible information (i.e. email address) will be stored on the server for one month and then deleted permanently.

- The second section allows to set oligo specific characteristics such as oligo length, oligo melting temperature, homopolymer length (i.e. maximum number of same-base repeat), GC content range or maximum number of allowed degenerate bases in the oligo sequence.

- The third section considers oligo pair parameters (e.g. primer pairs) such as the amplicon size or annealing temperature (not present in the Single Oligo Design page).

- In the forth section PhyloPrimer reports the positive consensus sequence that will be used for the oligo design. In case only the positive consensus exists, all the consensus bases will be black. If also a negative consensus was calculated, the bases will have different colors in relation to how the bases of the two consensus sequences differ at each position. All the gaps are removed from the positive consensus. The two bases surrounding an area where a gap region was present in the positive consensus but not in the negative are marked with bold characters. Vice-versa, any positive consensus base corresponding to a gap on the negative consensus is colored with blue. Where, for a certain position, the base between the two consensuses differs, that base is reported in red (Figure 2.4C). The user can highlight on the consensus sequence one area per oligo and PhyloPrimer will design the oligo only in the indicated area. This will shorten the analyses but may lead to a reduced oligo choice as it will reduce the area that the software will span for the research of optimal oligos.

- In this section the user can upload nucleotide sequences that PhyloPrimer must use to perform an *in silico* PCR and check the oligo specificity. This BLAST search will be performed on the side of the main one which is performed against a modified version of the GenBank database (more information in Section 2.2.5).

- The user can also set up the hairpin, self-dimer, and cross-dimer minimum allowed ΔG to avoid the formation of secondary structure during the PCR.

- The user can change the monovalent (e.g. Na^+ and K^+), Mg^{2+}, oligo and dNTP concentration, and PCR elongation temperature to reflect the PCR conditions specific to the used polymerase and PCR protocol. The more precise are these values the more precisely PhyloPrimer can predict melting temperature and secondary structure formation.

- PhyloPrimer only reports the 100 best oligo assays in the result page. This section reports several criteria that, if selected, PhyloPrimer uses to score the oligo assays and therefore to select the ones to report in the result page. Scoring points can be assigned to oligos with a T_m difference between the forward and reverse primers lower than 1 °C, oligos that have ΔG values higher than -1 kcal mol^{-1} and no degenerate bases in their sequence. Taxonomic specificity can be increased by selecting whether the oligos must be specific at the species, genus or higher taxonomic level in reference to the sequences that were selected in the dynamic tree; or also by giving extra scoring points to oligos that fall where there are differing bases between the positive and negative consensus, especially if they are at the 3' oligo end.

This page uses GDPR cookies to memorize the user preferences and speed up the parameter selection.
Once all the necessary parameters have been customized, PhyloPrimer shows the submission page and starts the oligo design and check on the server side. If the Dynamic Selection modality was used, the user is also given the possibility to go back to the initial tree and select more clusters from the same exact tree. In this way it is possible to construct multiple oligos, one (or more) for each tree cluster. This approach can also be used to try out different oligo parameters for the same cluster and sequences. Especially when a gene is very diversified, it may indeed be necessary to design multiple primers to tackle all the diversity.

2.2.1.5 Result page

When PhyloPrimer terminates the oligo design and check, it automatically sends the user a link to a result page. This page is divided into two main sections: on the left of the page there is the Oligo List which reports the 100 oligo assays considered the best by the PhyloPrimer scoring system. The oligo assay can be either represented by a primer pair, a primer pair plus a probe or a single oligo, depending on what was requested by the user. If the list shows oligos that were designed by PhyloPrimer all the oligo sequences will be preceded by F, R, P or O if the oligo is a forward primer, reverse primer, probe or a single oligo, respectively, then by a number which is the oligo position and, separated by a line, the oligo length (e.g. F57-18-TAGACGGGCTGACGTATG: this is a forward primer which 5' end is at the 57 bases on the consensus and it is 18 bases long). All the reported positions are intended as the positions at the 5' oligo end. If the oligos were only checked and not designed by PhyloPrimer, they will be reported with a sequence name only if it was provided by the user. The Oligo List also presents a search window that can be use for highlighting the oligos that reflect different criteria: melting temperature between a certain range, oligo length and taxonomy (at species and genus level) correspondent to the GenBank sequences that matched against all the oligos of an oligo assay.
On the right side of the page is all the information regarding the oligos that have been selected from the Oligo List. The information is organized into different pages: Specifics, ΔG, GenBank BLAST, Your BLAST (if the user uploaded a fasta file for an additional BLAST check), and Archive.

The section Specifics first reports the consensus sequence and how oligos are positioned on it (not present in the Oligo Check mode). In case both a positive and a negative consensus were calculated, it reports also the differing bases between the two sequences so that the user can visualize whether the oligos fall inside a differing region (Figure 2.4C). Oligo specifics (e.g. oligo length, GC content and melting temperature) are also reported in a detailed table.

The ΔG section reports all the self-dimer, cross-dimer and hairpin formations that were found for the selected oligos. The page reports in details all the oligo structures with an associated ΔG value lower than 0 kcal mol^{-1}. More details on how the secondary structures are calculated in Section 2.2.4.

The GenBank BLAST section reports the result of that BLAST search performed against a modified GenBank database (more details in Section 2.2.5). If the result page is reporting primer assays, the pie chart will report the taxonomy of the GenBank sequences aligned to both the forward and reverse sequences. In case of primer pair/probe assays, the pie chart will show matches that were aligned to the forward primer, reverse primers and probe together. If dealing with single oligos, the pie chart will report all the the matches. The pie chart aims exclusively to give an indicative idea about which organisms could be potentially targeted by a PCR and when reading it, it must be kept in mind that species represented in the pie chart are skewed towards species present in the PhyloPrimer sequence database. A BLAST match is considered valid if the oligo and the GenBank sequence matched with less than two mismatches. The mismatches are classified as external and internal mismatches. The external mismatches can not be consecutive whereas two consecutive internal mismatches are allowed (Figure 2.5). If two or more consecutive external mismatches or three or more mismatches are present, Phyloprimer assumes the oligo does not anneal with the DNA template and therefore no exponential amplification would occur. Below the pie chart there is a table reporting all the BLAST matches and some of the match specifics. The matches are ordered by which oligos they aligned to and by the alignment quality. The table presents the alignment characteristics for each oligo pair. PhyloPrimer reports the external and internal mismatches, the amplicon size, the subject sequence and the alignment start and end, plus the associated taxonomy. As in the GenBank BLAST, the Your BLAST section reports a pie chart and a table with the BLAST results. This BLAST search was performed on a database created by the fasta file that was uploaded by the user, therefore this page will only be visible if the additional file was uploaded in the oligo pages.

From the Archive section, the user can download an archive with all the input files and all the files that were outputted during the oligo design and check processes. Whereas only 100 oligo assays are reported in the result page, the archive files report all the PhyloPrimer oligos and

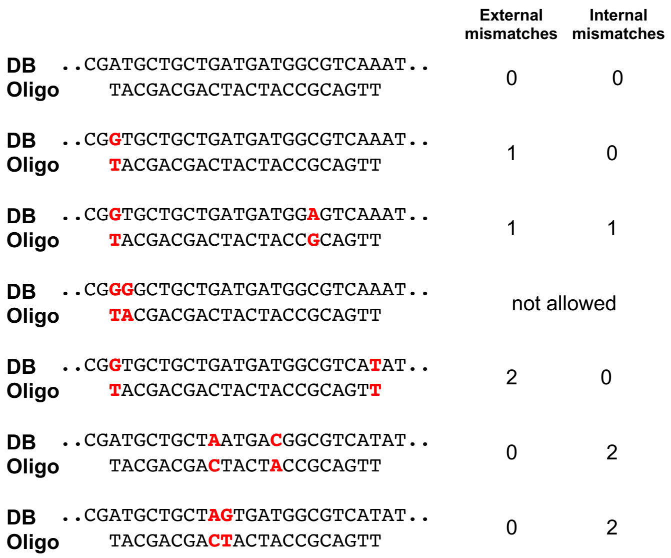

External mismatches
Internal mismatches
DB ..CGATGCTGCTGATGATGGCGTCAAAT..
Oligo TACGACGACTACTACCGCAGTT
0
0
DB ..CGGTGCTGCTGATGATGGCGTCAAAT..
Oligo TACGACGACTACTACCGCAGTT
1
0
DB ..CGGTGCTGCTGATGATGGAGTCAAAT..
Oligo TACGACGACTACTACCGCAGTT
1
1
DB ..CGGGGCTGCTGATGATGGCGTCAAAT..
Oligo TACGACGACTACTACCGCAGTT
not allowed
DB ..CGGTGCTGCTGATGATGGCGTCATAT..
Oligo TACGACGACTACTACCGCAGTT
2
0
DB ..CGATGCTGCTAATGACGGCGTCATAT..
Oligo TACGACGACTACTACCGCAGTT
0
2
DB ..CGATGCTGCTAGTGATGGCGTCATAT..
Oligo TACGACGACTACTACCGCAGTT
0
2

Figure 2.5: Allowed mismatches in the BLAST result where an oligo matched with a database sequence (DB2). The latter can be either a GenBank sequence or a sequence that was uploaded by the user. The red letters represent mismatches in the alignment.

therefore may be worth exploring to check if other suitable oligos were designed.

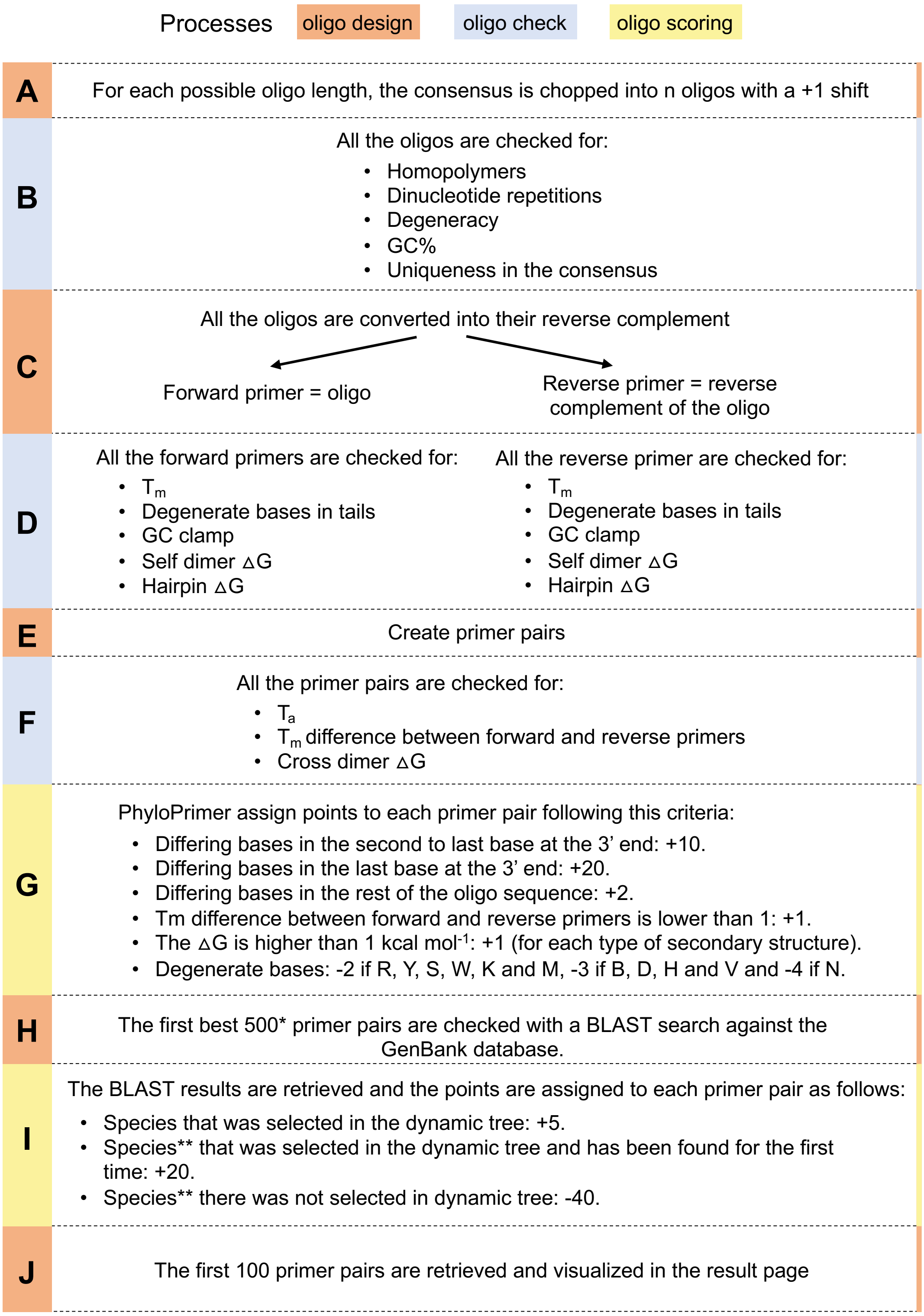
2.2.2 Oligo design and scoring system

PhyloPrimer uses a consensus approach for the oligo design or, in other words, it designs the oligos from a consensus sequence. The consensus sequence can be uploaded to PhyloPrimer by the user through the Premade Selection page or it can be calculated by PhyloPrimer itself. The software constructs the consensus with the DNA sequences or alignments uploaded through the Premade Selection page or with the sequences that were selected by the user on the dynamic tree (Dynamic Selection mode). In order for PhyloPrimer to find suitable oligos, the consensus must have one or more conserved regions, DNA regions that are in common among all the selected/uploaded sequences. If no conserved regions were present, the consensus sequence will be represented by long stretches of degenerate bases and the software will not be able to design any oligo from it. There can be different reasons for this: i) the sequence selection was too broad for the target gene family, ii) the selected sequences did not include only sequences from the same gene family, iii) the sequences represented different DNA regions of the same gene or iv) the studied gene family is very divergent. In general, it is more likely to have a conserved region in the consensus when working with closely related sequences, for example, when developing oligos for a specific species rather than for an entire gene family. However, when the aim is to develop oligos at gene level, the presence of a conserved gene region between different organisms highly depends on the gene sequence. It is essential to know the gene family object of the study and to check the consensus sequence that PhyloPrimer reports. In case the consensus presents a lot of degeneracy, it will be necessary to adjust the maximum number of degenerate bases allowed inside the oligo sequence in the oligo design pages. If this does not help, the design of different oligos for different cluster of organisms should be considered.

In PhyloPrimer the conserved region of the consensus sequence is determined by the maximum number of degenerate bases that is allowed inside the oligo sequences. For instance, if the user sets the maximum degenerate base value to 1, PhyloPrimer will discard all the oligos that have more than 1 degenerate base in the sequence or, in other words, won't consider the areas of the consensus that have an incidence of degenerate bases higher than 1 base every oligo length (between 18-22 bp by default). For example, if the maximum degenerate base value was set to 2, PhyloPrimer would discard the oligos and therefore the consensus area that had more than 2 bases every oligo length, and so on.

PhyloPrimer will start the oligo design only once the positive consensus has been obtained. For each possible oligo length (between 18 and 22 bases by default), the software extracts from the consensus sequence all the possible subsequences of that length (Figure 2.6A). This first step creates the starting pool of oligos that the following steps will check and discard if not respecting all the design parameters. The first check step discards by default the oligos that are not unique in the consensus sequence, that have homopolymer repetition longer than 3 bases, dinucleotide repetition longer than 6 bases, a GC content lower than 40% or higher than 60%, and will check and discard the oligos that do not have between 2 and 4 Gs/Cs in the last 5 bases of 3' oligo end (GC clamp). PhyloPrimer will also check if the oligos have a higher number of degenerate bases than the limit and that only the correct degenerate bases are present (all except from N by default). The default number of degenerate bases is set by PhyloPrimer in relation to how many degenerate bases were found inside the consensus sequence but can be changed by the user (Figure 2.6B).

PhyloPrimer then calculates the reverse complement of all the oligos and considers the original oligos as putative forward primers and the oligo reverse complements as putative reverse primers (Figure 2.6C). All the forward and reverse primers are progressively checked to have a valid melting temperature (between 54 °C and 64 °C by default) and, in case the presence of degenerate bases is allowed, not to have degenerate bases in the last 5 bases of the 5' oligo end and last 2 bases of the 3' end oligo tails (by default). The software also checks for the presence of self-dimer and hairpin secondary structures and discards any oligos with a secondary structure associated to a ΔG value lower than -5 and -3 kcal mol^{-1}, respectively (Figure 2.6D). If the primers had values that reflected the set criteria, PhyloPrimer uses them for finding suitable primer pairs (Figure 2.6E). The primer pairs are first selected considering the distance between their 5' ends on the consensus (between 200 and 600 bases by default). All the suitable primer pairs are then



Processes: oligo design, oligo check, oligo scoring

A For each possible oligo length, the consensus is chopped into n oligos with a +1 shift

B All the oligos are checked for:

- Homopolymers
- Dinucleotide repetitions
- Degeneracy
- GC%
- Uniqueness in the consensus

C All the oligos are converted into their reverse complement

Forward primer = oligo

Reverse primer = reverse complement of the oligo

D All the forward primers are checked for:

- T_m
- Degenerate bases in tails
- GC clamp
- Self dimer ΔG
- Hairpin ΔG

All the reverse primer are checked for:

- T_m
- Degenerate bases in tails
- GC clamp
- Self dimer ΔG
- Hairpin ΔG

E Create primer pairs

F All the primer pairs are checked for:

- T_a
- T_m difference between forward and reverse primers
- Cross dimer ΔG

G PhyloPrimer assign points to each primer pair following this criteria:

- Differing bases in the second to last base at the 3’ end: +10.
- Differing bases in the last base at the 3’ end: +20.
- Differing bases in the rest of the oligo sequence: +2.
- Tm difference between forward and reverse primers is lower than 1: +1.
- The ΔG is higher than 1 kcal mol^{-1}: +1 (for each type of secondary structure).
- Degenerate bases: -2 if R, Y, S, W, K and M, -3 if B, D, H and V and -4 if N.

H The first best 500* primer pairs are checked with a BLAST search against the GenBank database.

I The BLAST results are retrieved and the points are assigned to each primer pair as follows:

- Species that was selected in the dynamic tree: +5.
- Species** that was selected in the dynamic tree and has been found for the first time: +20.
- Species** there was not selected in dynamic tree: -40.

J The first 100 primer pairs are retrieved and visualized in the result page

Figure 2.6: Primer design workflow where oligo design, check and scoring processes are indicated. *250 if no negative consensus was present, no differing bases between the two consensus sequences were present or no differences were taken in consideration in the scoring system. **depending on the visualization criteria that were selected, +20 and -40 points are assigned if the different oligos were BLASTed against GenBank entries belonged to genera, families, orders, classes, phyla and domains that were attributed to the sequences selected from the phylogenetic tree.

checked for the presence of cross-dimer formations and discarded if the ΔG values are lower than -5 kcal mol^{-1}. Furthermore the primer pairs are also discarded if the melting temperature difference between forward and reverse primers is higher than 5 °C or the annealing temperature does not range between 50 °C and 60 °C (Figure 2.6F).

At this point, all the remaining primer pairs have all the requirements that were set by the user through the oligo pages. All the following steps aim to retrieve the best primer pairs that will be visualized in the result page. This is achieved assigning points to each primer pair as follows: 1 point is assigned to the primer pair if the melting temperature of the forward and reverse primers differed for less than 1 °C, for each secondary structure 1 point is assigned if the ΔG value is higher than -1 kcal mol^{-1}. Moreover, 20 points if a base polymorphic between the positive and the negative consensus is present in the last base of the 3' end and 10 points if it is present in the second to last position of the oligo 3' end, two points are also assigned for each base difference between the positive and the negative consensus (Figure 2.6G).

PhyloPrimer then selects the first 500 primer pairs that scored the highest points according to the scoring system (Figure 2.6H). The oligos belonging to those first 500 primer pairs will be BLASTed against the GenBank database. PhyloPrimer then checks the BLAST results and considers only the BLAST matches that were matched by both the forward and the reverse primers of a primer pairs, if that sequence belongs to one of the species that were selected from the dynamic tree, PhyloPrimer assigns 10 points to the primer pair, if the species was not among the selected species it will deduct 40 points and every time there is a new correct species is going to add 20 points to the total. By default, PhyloPrimer will not assign more points to primers that belong to the same genus (or higher ranks) of the selected tree entries. But if these visualization parameters are checked, PhyloPrimer will assign 20 points to the entries that belong to the same taxonomy and deduct 40 to those that do not. This is for facilitating the design of oligos that are specific for a specific genus (or higher taxonomic group) rather than only specific to certain species. In case an additional file was uploaded by the user for an additional BLAST check, PhyloPrimer will also BLAST all the oligos against that database but the outcome will not be object of the scoring system (Figure 2.6I).

The described scoring criteria are all active by default but any of those can be deselected by the user on the Oligo Design page. PhyloPrimer then selects the first 100 primer pairs and these primer pairs will be the ones showed in the last result page (Figure 2.6J). When degenerate bases are present inside the oligo sequences, the melting temperature and the GC content are calculated as the mean of these values in each of the possible oligo.

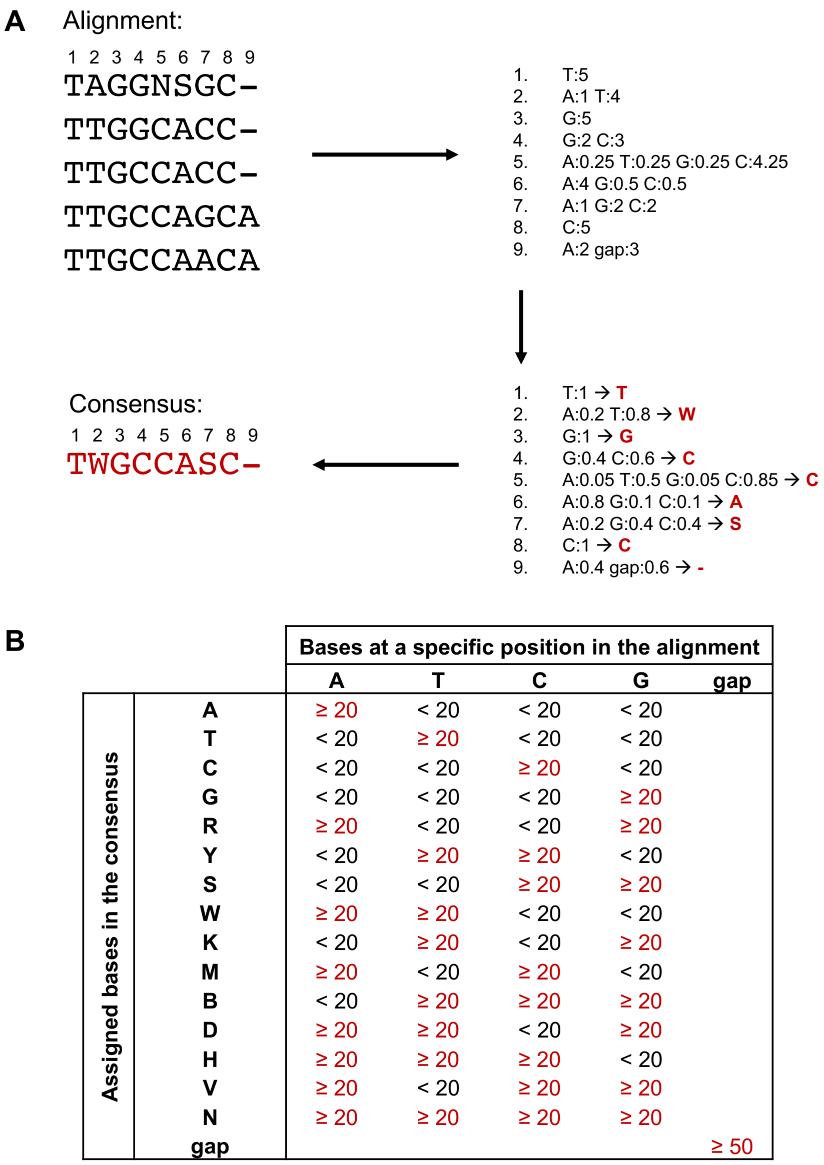
Above I described only the process for the design of primer pairs. The workflow to design single oligos is similar except from the oligo pairs' formation which is not performed. The workflow to design primer pair/probe assays presents the following modifications:

- PhyloPrimer also designs the probe. The latter is designed with the same process as for the forward primer in case it is an anti-sense probe, as for the reverse primers in case it is a sense probe or as for both the primers if the binding DNA template strand is not specified.
- Probe sequences with a guanine in the last base of the 5' end are discarded.
- During the process of creating the primer pairs, PhyloPrimer looks for a probe that falls inside a consensus region between the two primers.
- The primer pair/probe assays are checked for the melting temperature difference between the probe and the primers. The assay is discarded if the difference between the melting temperature of the probe and the one of the primers (the average between forward and reverse primers) is lower than 10 °C (by default).

Finally, in the Oligo Check mode, PhyloPrimer performs the *in silico* PCR tests and the same oligo checks that are performed in the other modes (Figure 2.6D and F), varying in relation to which kind of oligo was uploaded by the user.

2.2.3 Consensus calculation

The consensus is a nucleotide sequence calculated from alignment sequences and reports the most frequent base for each alignment position (Figure 2.7A). PhyloPrimer constructs the consensus sequence with a custom script. The software goes through the alignment position by position (column by column) and for each position looks at the base composition for all the alignments. For each position, 1 point is assigned to each A, C, G, T or gap (- or .) that is found. When there are degenerate bases in the alignment sequences, fractioned points are assigned to the correspondent bases of the degenerate base (according to the IUPAC nucleotide code). For instance, if the degenerate base N is present, 1/4 point is assigned to A, 1/4 to C, 1/4 to T and 1/4 to G; in case of a B, 1/3 is assigned to G, 1/3 to C and 1/3 to T; if S is present 1/2 point is assigned to C and 1/2 to G. Once the software knows the frequency of each of the four base (and the gap) at all the positions, it determines which base to assign to each position following the criteria illustrated in Figure 2.7B. For example, if for a certain position, 50% or more of the alignment sequences had a gap, a gap will be reported in the consensus, if one of the bases represents the 20% or more and each of the other less than 20%, that position will be assigned to that letter. If there are more than two bases present with 20% or more abundance, that position will be assigned to the corresponding degenerate base.



Assigned bases in the consensus	Bases at a specific position in the alignment				
	A	T	C	G	gap
A	≥ 20	< 20	< 20	< 20	
T	< 20	≥ 20	< 20	< 20	
C	< 20	< 20	≥ 20	< 20	
G	< 20	< 20	< 20	≥ 20	
R	≥ 20	< 20	< 20	≥ 20	
Y	< 20	≥ 20	≥ 20	< 20	
S	< 20	< 20	≥ 20	≥ 20	
W	≥ 20	≥ 20	< 20	< 20	
K	< 20	≥ 20	< 20	≥ 20	
M	≥ 20	< 20	≥ 20	< 20	
B	< 20	≥ 20	≥ 20	≥ 20	
D	≥ 20	≥ 20	< 20	≥ 20	
H	≥ 20	≥ 20	≥ 20	< 20	
V	≥ 20	< 20	≥ 20	≥ 20	
N	≥ 20	≥ 20	≥ 20	≥ 20	
gap					≥ 50

Figure 2.7: Consensus calculation workflow. PhyloPrimer processes the alignment sequences into a consensus sequence (A) by assigning a base to each consensus position in relation to the base frequency in each alignment position (B).

2.2.4 Melting temperature and ΔG secondary structures

PhyloPrimer calculates oligo melting temperatures (T_m) and secondary structure Gibbs free energies (ΔG) with the nearest-neighbor (NN) model for nucleic acids. This model predicts the thermodynamic behavior of a DNA molecule using the thermodynamic parameters of each nucleotide pair composing the molecule itself. Both the T_m and the ΔG calculation rely on the use of the thermodynamics parameters enthalpy (ΔH) and entropy (ΔS). These parameters were derived from calorimetry and spectroscopic experiments of DNA duplexes for nucleotide base pair motives (SantaLucia and Hicks, 2004), internal mismatches (Allawi and Santalucia, 1997; Allawi and SantaLucia, 1998a; Allawi and SantaLucia, 1998c; Allawi and SantaLucia, 1998b;

Peyret et al., 1999), dangling ends (Bommarito et al., 2000) and hairpin terminal mismatches (unpublished data). The latter were retrieved from the UNAFold database (Markham and Zuker, 2008). The ΔH and ΔS are considered temperature independent when working with nucleic acids and are reported for 1M Na^+ conditions. All the thermodynamic parameters are summarized in Appendix 2.A.1.

The melting temperature (T_m) of a DNA molecule is the temperature in which half of the DNA is paired with its complement and half is single stranded. The correct calculation of this parameter is essential to the correct calculation of the PCR annealing temperature, and it is pivotal for the qPCR probe when wanting to differentiate amplicon expression levels. PhyloPrimer calculates T_m with the formula reported from SantaLucia and Hicks, 2004. More information on the T_m calculation can be found in Appendix 2.A.2.
The annealing temperature, T_a, is calculated as the lowest melting temperature (if more than one oligo is present) minus 5. This is an indicative calculation as the optimal annealing temperature can considerably vary in relation to the polymerase that is used during the PCR.

The ΔG, or Gibbs free energy, estimates if a reaction can occur spontaneously (ΔG higher than 0, exergonic reaction) or not (ΔG lower than 0, endergonic reaction) and therefore indicates how stable a particular DNA structure is at a certain temperature. In this case, ΔG represents the quantity of energy needed to fully break a secondary structure. The lower it is (more negative), the more stable and likely to occur the secondary structure will be and the more energy will be required to break it. ΔG is defined as equal to the enthalpy minus the product of the temperature times the entropy (Gibbs free energy equation). PhyloPrimer calculates the ΔG for three different secondary structure formations: self-dimers (i.e. dimers formed within the oligo itself), cross-dimers (i.e. dimers formed between different oligos) and hairpin loops (i.e. hairpin-like secondary structures formed within the oligo itself). For each of these different structures, different rules must be applied to ΔH and ΔS calculation which are then used to apply Gibbs free energy equation (SantaLucia and Hicks, 2004). More information on ΔG calculation can be found in Appendix 2.A.3.

Melting temperature and ΔG values obtained in this way (SantaLucia and Hicks, 2004; Gibbs free energy equation) are valid only in 1 M Na^+ condition. Because the PCR conditions can span a wide range of different conditions, salt correction formulas must be applied to correct the obtained values (Owczarzy et al., 2004; Owczarzy et al., 2008). Depending on the polymerase used and the PCR protocol, Mg^{2+} and monovalent ions can vary considerably and rarely the 1 M Na^+ condition is respected. Mg^{2+} can be added to the PCR reaction as $MgCl_2$ or $MgSO_4$, it is a DNA polymerase co-factor and it helps the stabilization of the primer-template DNA duplex influencing the negative charges of the DNA backbones. Monovalent ions such as Na^+ and K^+ (KCl) are also used for stabilizing the DNA duplex, whereas Tris-HCl is used for stabilizing the pH.

PhyloPrimer performs salt-correction correction with the parameters reported and customized in the oligo pages therefore calibrating the corrections on the user specific PCR conditions. More information on the used formulas in Appendix 2.A.4.

Worth mentioning, there are also other elements of a PCR reactions that can influence the melting temperature and the Gibbs energy calculation such as $NH4^+$, DMSO, dyes, glycerol, DMSO, formamide, TMAC and betaine. For these components no correction formulas have been proposed yet. But there are some approximation that the user can look at, such as the one suggested by Von Ahsen et al., 2001 where it is suggested to decrease the T_m by 0.75 °C for each volume percentage of dimethyl sulfoxide (DMSO).

When dealing with degenerate oligos, PhyloPrimer calculates melting temperature and ΔG values for all the possible oligos. The final T_m is the average of all the calculated T_m whereas the final ΔG is the lowest ΔG.

2.2.5 Databases

PhyloPrimer uses an external nucleotide sequence database in two points of the pipeline. The first point is when it BLASTs the sequences uploaded in the Dynamic Selection mode to retrieve similar sequences and construct a dynamic phylogenetic tree (DB1), and the second when it checks the oligo specificity through *in silico* PCR (DB2). The two databases are different but in both cases all the sequences are a selection of the GenBank database (Sayers et al., 2019). Genbank is a comprehensive database composed by only nucleotide sequences annotated by the users and a curated annotation pipeline. It gathers different kinds of sequences which are also divided into different sections by taxonomy and sequence type (e.g. shotgun metagenomic sequences, cDNA sequences, full genomes).

The first database (DB1) is constituted by GenBank sequences from EST (Expressed sequence tags), TSA (transcriptome shotgun assembly), HTC (High-throughput cDNA), ENV (Environmental samples), VRL (Viruses), PLN (Plants), PHG (Phages), BCT (Bacteria) and annotated genes from the GenBank genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/). Non-annotated genomes are reported with their entire genome sequence. The choice of reporting annotated genomes wherever possible was made in order to give more detailed information of the BLAST hit region to the user for the dynamic tree visualization. In DB1 entries belonging to all the organisms except from Metazoa (kingdom) and Streptophyta (phylum) are included.

With a custom script, the sequences were grouped together when different entries had exactly the same nucleotide sequence and the same length, creating a non-redundant database.

The second database (DB2) comprises GenBank sequences from EST, TSA, HTC, ENV, VRL, PLN, PHG and BCT, as DB1. However, instead of only the annotated genes, it contains the whole genomes. This database also reports sequences belonging to Metazoa and Streptophytes.

The last database update was made in June 2019. DB1 has 199,157,016 entries, whereas DB2 has 58,709,420. Once published and publicly available, the database will be regularly updated every two months.

2.3 PhyloPrimer tests

PhyloPrimer was used by Jared Wray, a MSc student at the University of Bristol and co-supervised by myself and Dr. Barker, to develop primers to detect organisms belonging to the *Streptococcus* genus and specific *Streptococcus* species (*Streptococcus agalactiae*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Streptococcus mutans* and *Streptococcus mitis*) from known mock communities. I will report and comment how these primer pairs performed in the next sections.

2.3.1 Universal genes and the rpoB gene

Prokaryotic universal genes codify for cell housekeeping functions and therefore are universally present in the genome of all the Prokaryotes (Gil et al., 2004; Acevedo-Rocha et al., 2013). The 16S rRNA gene is the most widely used universal gene to characterize taxonomic microbial diversity in both environmental and biomedical studies (Mancabelli et al., 2017; Prodan et al., 2020; Baird and Hajibabaei, 2012). This gene encodes for the 16S ribosomal RNA which is a component of the 30S subunit of the prokaryotic ribosome. The wide use of this gene for microbial screening studies is due to the presence of both highly conserved sequence regions which allow the design of universal primers, and the presence of variable regions which allow the distinction between different organisms. In Chapter 3, I amplified and sequenced this gene to investigate diversity and differentiation of microbial communities sampled from water englacial channels.

However, because of its high sequence conservation level, the 16S rRNA gene is not suitable when the study aim is to target more taxon specific organisms. Other universal genes, such as protein coding genes, show a higher sequence variability and are more suitable for more targeted studies. The rpoB gene is a universal gene and encodes for the β-subunit of the RNA polymerase

(Adékambi et al., 2009). This is an essential enzyme to all the transcription processes in a cell as it accounts for the synthesis of mRNA, tRNA and rRNA. Its sequence is less conserved across different genomes compared to the 16S rRNA gene. This makes it less suitable to design universal primers but more suitable to design primers that can target specific organisms (Case et al., 2007).

We therefore developed primer pairs to target different species belonging to the *Streptococcus* genus focusing on the rpoB gene.

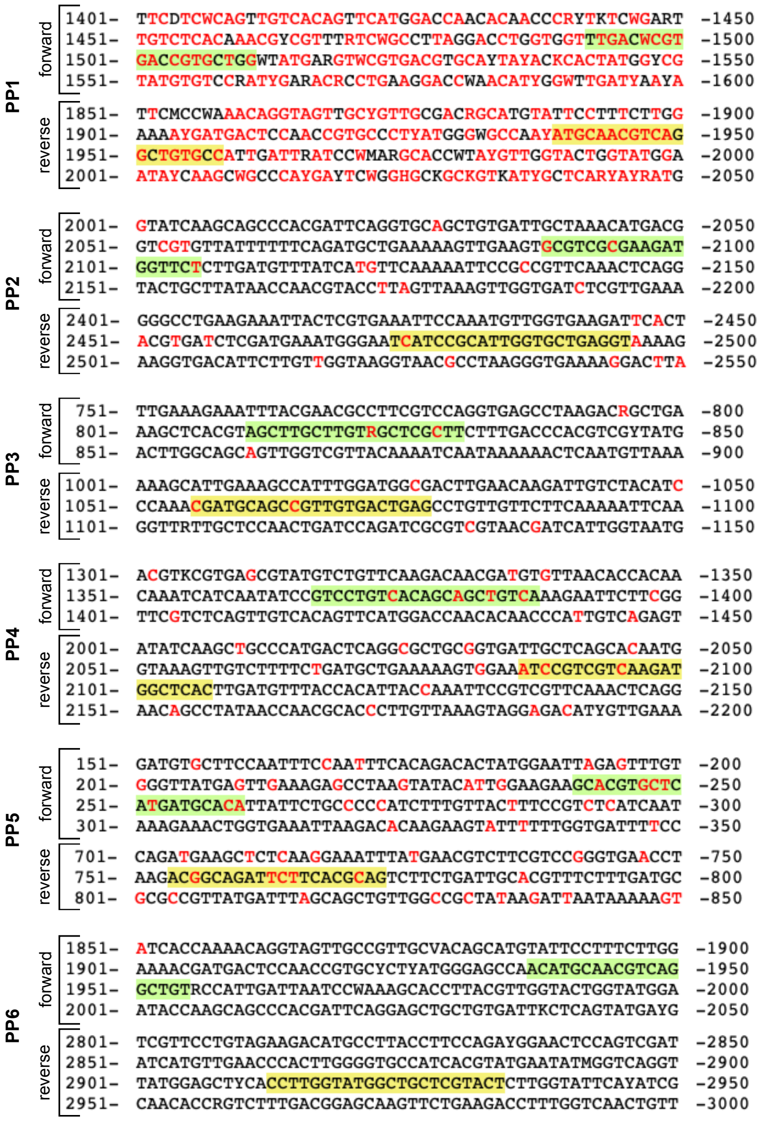
2.3.2 Mock communities and primer design

The primer pairs were designed to amplify all the organisms related to the genus *Streptococcus* (PP1), and five *Streptococcus* species: *Streptococcus agalactiae* (PP2), *Streptococcus pneumoniae* (PP3), *Streptococcus pyogenes* (PP4), *Streptococcus mutans* (PP5) and *Streptococcus mitis* (PP6) (Table 2.1). The primer design was performed with PhyloPrimer (Dynamic Selection mode) with default parameters with exception of melting and annealing temperature (60-75 °C), monovalent ion concentration (0 mM), magnesium ion concentration (2.5 mM), oligo concentration (0.6 µM) and dNTP concentration (1.2 µM). These parameters were modified accordingly to the specifics of the polymerase used for the PCR. The primers were selected from the PhyloPrimer result pages (see Appendix 2.B to find the links to the result pages). Figure 2.8 reports the designed primers and their position on the consensus sequence used to design them. Furthermore, in order to be sure the DNA was amplifiable in all the mock communities, I also used the primers 341F and 518R to amplify the 16S RNA gene as a positive control (Table 2.1) (Muyzer et al., 1993).

The primers were tested with four mock communities: Metagenomic Control Material for Pathogen Detection (ATCC® MSA-4000), 10 Strain Staggered Mix Genomic Material (ATCC® MSA-1001), Skin Microbiome Genomic Mix (ATCC® MSA-1005) and ZymoBIOMICS Microbial Community DNA Standard (D6306, Zymo Research). These communities comprise several organisms, present with different abundances and ranging a wide range of microbial diversity. In

Table 2.1: Primer specifics. All the primers were designed with PhyloPrimer web platform except the 16S rRNA primers which were designed by Muyzer et al., 1993. *PhyloPrimer predicted amplicon length.

Primer pair	Target organisms	Primer sequence		Amplicon size*
		Forward	Reverse	
16S rRNA	Bacteria	CCTACGGGAGGCAGCAG	GGCACAGCCTGACGTTGCAT	200
PP1	*Streptococcus sp.*	TTGACWCGTGACCGTGCTGG	GGCACAGCCTGACGTTGCAT	470
PP2	*Streptococcus agalactiae*	GCGTCGCGAAGATGGTTCT	ACCTCAGCACCAATGCGGATGA	410
PP3	*Streptococcus pneumoniae*	AGCTTGCTTGTRGCTCGCTT	CTCAGTCACAACGGCTGCATCG	270
PP4	*Streptococcus pyogenes*	CAGTTGCACAGGCCAATTCGA	GTGAGCCATCTTGACGACGGAT	380
PP5	*Streptococcus mutans*	GCGAGCGTCTTGTCAAGGAT	ACCACCAAGCGGCTGTTGA	870
PP6	*Streptococcus mitis*	ACATGCAACGTCAGGCTGT	AGTACGAGCAGCCATACCAAGG	1000



```
1401-  TTCDTCWCAGTTGTCACAGTTCATGGACCAACACAACCCRYTKTCWGART  -1450
1451-  TGTCTCACAAACGYCGTTTRTCWGCCTTAGGACCTGGTGGTTTGACWCGT  -1500
1501-  GACCGTGCTGGWTATGARGTWCGTGACGTGCAYTAYACKCACTATGGYCG  -1550
1551-  TATGTGTCCRATYGARACRCCTGAAGGACCWAACATYGGWTTGATYAAYA  -1600

1851-  TTCMCCWAAACAGGTAGTTGCYGTTGCGACRGCATGTATTCCTTTCTTGG  -1900
1901-  AAAAYGATGACTCCAACCGTGCCCTYATGGGWGCCAAYATGCAACGTCAG  -1950
1951-  GCTGTGCCATTGATTRATCCWMARGCACCWTAYGTTGGTACTGGTATGGA  -2000
2001-  ATAYCAAGCWGCCCAYGAYTCWGGHGCKGCKGTKATYGCTCARYAYRATG  -2050

2001-  GTATCAAGCAGCCCACGATTCAGGTGCAGCTGTGATTGCTAAACATGACG  -2050
2051-  GTCGTGTTATTTTTCAGATGCTGAAAAAGTTGAAGTGCGTCGCGAAGAT  -2100
2101-  GGTTCTCTTGATGTTTATCATGTTCAAAAATTCCGCCGTTCAAACTCAGG  -2150
2151-  TACTGCTTATAACCAACGTACCTTAGTTAAAGTTGGTGATCTCGTTGAAA  -2200

2401-  GGGCCTGAAGAAATTACTCGTGAAATTCCAAATGTTGGTGAAGATTCACT  -2450
2451-  ACGTGATCTCGATGAAATGGGAATCATCCGCATTGGTGCTGAGGTAAAAG  -2500
2501-  AAGGTGACATTCTTGTTGGTAAGGTAACGCCTAAGGGTGAAAAGGACTTA  -2550

 751-  TTGAAAGAAATTTACGAACGCCTTCGTCCAGGTGAGCCTAAGACRGCTGA  -800
 801-  AAGCTCACGTAGCTTGCTTGTRGCTCGCTTCTTTGACCCACGTCGYTATG  -850
 851-  ACTTGGCAGCAGTTGGTCGTTACAAAATCAATAAAAAACTCAATGTTAAA  -900

1001-  AAAGCATTGAAAGCCATTTGGATGGCGACTTGAACAAGATTGTCTACATC  -1050
1051-  CCAAACGATGCAGCCGTTGTGACTGAGCCTGTTGTTCTTCAAAAATTCAA  -1100
1101-  GGTTRTTGCTCCAACTGATCCAGATCGCGTCGTAACGATCATTGGTAATG  -1150

1301-  ACGTKCGTGAGCGTATGTCTGTTCAAGACAACGATGTGTTAACACCACAA  -1350
1351-  CAAATCATCAATATCCGTCCTGTCACAGCAGCTGTCAAAGAATTCTTCGG  -1400
1401-  TTCGTCTCAGTTGTCACAGTTCATGGACCAACACAACCCATTGTCAGAGT  -1450

2001-  ATATCAAGCTGCCCATGACTCAGGCGCTGCGGTGATTGCTCAGCACAATG  -2050
2051-  GTAAAGTTGTCTTTTCTGATGCTGAAAAAGTGGAAATCCGTCGTCAAGAT  -2100
2101-  GGCTCACTTGATGTTTACCACATTACCAAATTCCGTCGTTCAAACTCAGG  -2150
2151-  AACAGCCTATAACCAACGCACCCTTGTTAAAGTAGGAGACATYGTTGAAA  -2200

 151-  GATGTGCTTCCAATTTCCAATTTCACAGACACTATGGAATTAGAGTTTGT  -200
 201-  GGGTTATGAGTTGAAAGAGCCTAAGTATACATTGGAAGAAGCACGTGCTC  -250
 251-  ATGATGCACATTATTCTGCCCCCATCTTTGTTACTTTCCGTCTCATCAAT  -300
 301-  AAAGAAACTGGTGAAATTAAGACACAAGAAGTATTTTTTGGTGATTTTCC  -350

 701-  CAGATGAAGCTCTCAAGGAAATTTATGAACGTCTTCGTCCGGGTGAACCT  -750
 751-  AAGACGGCAGATTCTTCACGCAGTCTTCTGATTGCACGTTTCTTTGATGC  -800
 801-  GCGCCGTTATGATTTAGCAGCTGTTGGCCGCTATAAGATTAATAAAAAGT  -850

1851-  ATCACCAAAACAGGTAGTTGCCGTTGCVACAGCATGTATTCCTTTCTTGG  -1900
1901-  AAAACGATGACTCCAACCGTGCYCTYATGGGAGCCAACATGCAACGTCAG  -1950
1951-  GCTGTRCCATTGATTAATCCWAAAGCACCTTACGTTGGTACTGGTATGGA  -2000
2001-  ATACCAAGCAGCCCACGATTCAGGAGCTGCTGTGATTKCTCAGTATGAYG  -2050

2801-  TCGTTCCTGTAGAAGACATGCCTTACCTTCCAGAYGGAACTCCAGTCGAT  -2850
2851-  ATCATGTTGAACCCACTTGGGGTGCCATCACGTATGAATATMGGTCAGGT  -2900
2901-  TATGGAGCTYCACCTTGGTATGGCTGCTCGTACTCTTGGTATTCAYATCG  -2950
2951-  CAACACCRGTCTTTGACGGAGCAAGTTCTGAAGACCTTTGGTCAACTGTT  -3000
```

Figure 2.8: Primer position on the consensus sequence as visualized on the PhyloPrimer result pages for the different primer pairs (PP1, PP2, PP3, PP4, PP5 and PP6). The base color code is as follows: red letters indicate that at that positions the two sequences presented differing bases, blue letters indicate positions where there are bases on the positive consensus but gaps in the negative and bold letters flank regions where there were gaps on the positive consensus but bases on the negative consensus. A degenerate base is marked as differing only if that base does not contain the correspondent base of the negative consensus.

the following tests they will be called community A, B, C and D, respectively (Table 2.2).

Each mock community DNA was used as template for the PCR amplification using the primers 16S rRNA and the PhyloPrimer developed primer pairs (PP1, PP2, PP3, PP4, PP5 and PP6). The 25 µL PCR solution consisted in 12.5 µL for 2x KAPA HiFi Hot Start ReadyMix polymerase (KAPA BIOSYSTEMS), 1.5 µL of 5 µM forward primer, 1.5 µL of 5 µM reverse primer, between 1-3 µL of template DNA (corresponding to 4 ng of DNA) and nuclease-free water up to volume. Additionally to the mock community samples, each PCR reaction was also run with a negative control where the template DNA was substituted with nuclease-free water. The PCR conditions were as follows: 95 °C for 3 minutes, 25 cycles of 98 °C for 20 seconds, 64 °C for 15 seconds and 72 °C for 20 second, and a final extension of 72 °C for 1 minute. The annealing temperature of 64 °C was used for all the primer pairs PP1, PP2, PP4, PP5 and PP6, whereas I used 65 °C for P3 and 62 °C for the 16S rRNA primers.

For each sample, 6 µL of PCR product was then run with 2 µL of gel loading buffer (NEB) on 1.5% w/v horizontal agarose gel (0.5 mg ethidium bromide ml^{-1}) in 1x TEA buffer (Tris acetate EDTA) and run for 30 minutes at 120 mV (Bio-Rad PowerPac 300, Bio-Rad Laboratories). Gel pictures were visualized under UV light and captured with GelDoc-ItTS2 Imager (UVP). No bands were showed in any of the negative control lanes. GelPilot 100 bp Plus Ladder (QIAGEN, Hilden, Germany) was run for amplicon size comparison.

Table 2.2: Mock microbial composition for the communities A, B, C and D where community A corresponds ATCC$^{®}$ MSA-4000, B to ATCC$^{®}$ MSA-1001, C to ATCC$^{®}$ MSA-1005 and D to the ZyMO community.

Species	**Relative abundance** (%)			
	A	**B**	**C**	**D**
Acinetobacter baumannii	0.10	-	-	-
Acinetobacter johnsonii	-	-	16.70	-
Bacillus cereus	-	4.48	-	-
Bacillus subtilis	-	-	-	12.00
Bifidobacterium adolescentis	-	0.04	-	-
Clostridium beijerinckii	-	0.45	-	-
Corynebacterium striatum	-	-	16.70	-
Cryptococcus neoformans	-	-		2.00
Cutibacterium acnes	-	-	16.70	-
Deinococcus radiodurans	-	0.04	-	-
Enterococcus faecalis	0.70	0.04		12.00
Escherichia coli	1.40	4.48	-	12.00
Klebsiella pneumoniae	14.40	-	-	-
Lactobacillus fermentum	-	-	-	12.00
Lactobacillus gasseri	-	0.45	-	-
Listeria monocytogenes	-	-	-	12.00
Micrococcus luteus	-	-	16.70	-
Neisseria meningitidis	28.90	-	-	-
Pseudomonas aeruginosa	0.30	-	-	12.00
Rhodobacter sphaeroides	-	44.78	-	-
Saccharomyces cerevisiae	-	-	-	2.00
Salmonella enterica	-	-	-	12.00
Staphylococcus aureus	15.10	-	-	12.00
Staphylococcus epidermidis	-	44.78	16.70	-
Streptococcus agalactiae	2.90	-	-	-
Streptococcus mitis	-	-	16.70	-
Streptococcus mutans	-	0.45	-	-
Streptococcus pneumoniae	28.90	-	-	-
Streptococcus pyogenes	7.20	-	-	-

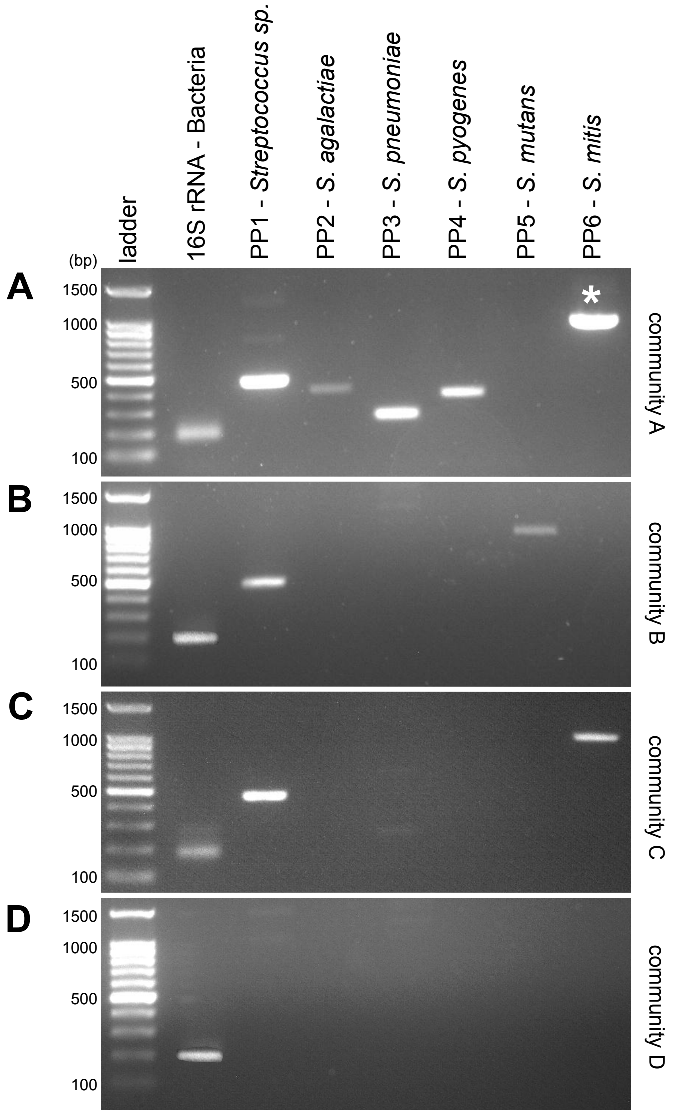


Figure 2.9: Agarose gel pictures of the PCR products amplified with the 16S rRNA primers and the PhyloPrimer designed primer (PP1, PP2, PP3, PP4, PP5 and PP6) on the mock community A, B, C and D. The white star marks the non-specific band found in community A for the primer PP6. All the other primer pairs amplified only the expected communities and no false negatives occurred.

2.3.3 Test results

The 16S rRNA gene was amplified in all the four communities showing that all the DNA communities had amplifiable microbial DNA (Figure 2.9). The primer pair PP1 which was specific for the *Streptococcus* genus amplified same length amplicons (about 500 bp) in all the communities except from community D where no *Streptococcus* species were present (Table 2.2). Primers PP2, PP3 and PP4 which targeted respectively *S. agalactiae*, *S. pneumoniae* and *S. pyogenes* showed PCR products only in community A which was the only community that contained these organisms. The amplicon size also reflected that predicted by PhyloPrimer being around 480, 300 and 400 bp for primer PP2, PP3 and PP4, respectively. The primer pair PP5 only amplified in the community B which was the only one containing *S. mutans*. Finally, the primer pair PP6, specific for *S. mitis*, showed same length bands (around 1000 bp) in both community A and C. Community A did not contain *S. mitis* and therefore species specificity was not achieved with this primer pair. When checked with primer-BLAST (Ye et al., 2012), an *in silico* PCR tool, the primer pair also targeted other *Streptococcus* species such as *S. agalactiae* and *S. pyogenes* which are present in community A (Table 2.2). Figure 2.8 shows that the primer pair PP6 was the only one without differing bases between positive and negative consensus inside its forward and reverse primer sequences. The complete sequence of the consensus sequence has a total of 8 differing bases and can be found at the *S. mitis* PhyloPrimer result page (see Appendix 2.B).

2.4 Discussion

The development of taxonomic specific primers is essential to many environmental and biomedical biomonitoring and detection studies (e.g. Song et al., 2000; Liu et al., 2003; Ai et al., 2019; Santos et al., 2020) where the recent COVID-19 pandemic is a perfect example of how important is the design of species-specific primers to detect a specific organism of interest (Park et al., 2020). I developed PhyloPrimer, an automated platform that integrates a new pipeline which aims to design taxonomic-specific oligos and tests them for secondary structures and target specificity.

We developed and tested six different primer pairs to target *Streptococcus* at both genus and species level (i.e. *S. agalacitae*, *S. pneumoniae*, *S. pyogenes*, *S. mutants* and *S. mitis*). When tested on four mock communities, five primer pairs out of six (PP1, PP2, PP3, PP4 and PP5) showed specificity to the target organisms and the correct predicted amplicon length (Table 2.1 and Figure 2.9). The primer pair PP6, designed to target *S. mitis*, showed an non-specific band in community A which did not contain that organism. This band was probably due to PP6 targeting also other *Streptococcus* species present in community A (Table 2.2). *In silico* PCR performed by primer-BLAST showed how the primer pair PP6 also targets *S. agalactiae* and *S. pyogenes*. This software relies on the updated and comprehensive NCBI database to check for PCR targeted organisms. Whereas, PhyloPrimer also relies on NCBI sequence database, its database is not up-to-date with the NCBI online version. At the time of the *Streptococcus* primer design, the PhyloPrimer database was updated to June 2019. The sequences targeted by PP6 in the primer-BLAST *in silico* PCR were all added to the NCBI dataset starting from June 2019, after the last PhyloPrimer database update and therefore were not present in the database. Most of the *S. pyogenes* and *S. agalactiae* sequences were added to the NCBI dataset from different unpublished projects, and extensive sequencing and analyses of clinically relevant organisms (Davies et al., 2019; Baines et al., 2019).

This result shows how quickly the information present in publicly available databases has increased and will further increase in the future, especially for clinically relevant organisms (Selzer et al., 2018). This sets the necessity to regularly update PhyloPrimer databases once the software will be made publicly available.
In our case, the presence of these new *S. pyogenes* and *S. agalactiae* sequences in the PhyloPrimer database would have allowed a more complete dynamic tree construction and the calculation of positive and negative consensus reporting more differing bases between their sequences. A better characterization of the differing areas between positive and negative consensus allows a more informed oligo selection thanks to the PhyloPrimer oligo scoring system where more points are given to oligos falling inside diversified areas between the two consensus sequences (Figure 2.4 and Figure 2.6). The consensus calculated for the design of PP6 had only 8 differing bases and no primers were designed in those areas because the oligos did not satisfied other design requirements (e.g. melting temperature). The fact that the only primer pair that did not fall inside differing regions (Figure 2.8) was the only one with non-specific results highlights how the PhyloPrimer positive-negative consensus approach is effective for the design of specific primers. Furthermore, if these new sequences were in the database, a more complete taxonomy profiling of the targeted sequences would have appeared in the pie chart, leading to an easier exclusion on these oligos by the user.

PhyloPrimer performed well with different organism and gene settings and showed overall good results when tested on the mock communities. However, the software also comes with some limitations. For example, it was designed with microbial communities in mind. Therefore it is not suitable for the design of eukaryotic oligos as it does not deal with intron and exon regions and the database used for the tree construction is built only with microbial sequences (i.e. all the organisms except from Metazoa and Streptophyta). Also, PhyloPrimer does not design degenerate oligos specifically. PhyloPrimer uses a consensus approach and it designs the oligos from a consensus sequence calculated from a DNA alignment. Therefore it will not introduce degeneracy on purpose and will design oligos containing degenerate bases only if present in the consensus sequence and if necessary to the design of suitable oligos. PhyloPrimer run time is also quite long, sometimes being over an hour when it is dealing with long consensus sequences.

The PhyloPrimer platform is a semi-automated and user-friendly pipeline to go from sequence selection to designed and *in silico* tested oligos. I am confident that this tool could really help with oligo design of complex environmental communities speeding up and providing a solid and reproducible pipeline for the oligo design and *in silico* tests. In particular, the tool is being tested further by HYDREKA ENOVEO where they will be able to apply it to real environmental communities and to design taxon-specific oligos for their biomonitoring and bioremediation projects.

2.5 Conclusion

I developed PhyloPrimer, an online platform for the design and the *in silico* test of microbial oligos. Thanks to its unique approach to increase oligo taxon-specificity (i.e. positive and negative consensus comparison), the software is particularly adapted to design species-specific oligos that can be used in monitoring and detection studies. The software received positive feedback and performed well on mock microbial communities. It is publicly available at www.cerealsdb.uk.net/cerealgenomics/phyloprimer.

Acknowledgements

I thank Dr. Cédric Malandain, from HYDREKA ENOVEO, and Dr. Gary Barker for their comments and suggestions during the first stages of PhyloPrimer development. I thank Ella Boswell and Nina Blampied, from the University of Bristol, and Layla Rahiem and Mirjam Pikaart, from Avans University, who used PhyloPrimer and gave useful feedback in return. Finally, I thank Jared Wray who developed the oligos I presented in this chapter.

A modified version of this chapter is in preparation to be submitted to a peer-reviewed journal: 'Varliero, G., Malandain, C., Wray, J., Barker, G. L., PhyloPrimer: a taxon-specific oligonucleotide design platform'.

Appendix

2.A Thermodynamics calculation formulas and application

2.A.1 Tables for the thermodynamics calculation

Table A2.1: ΔH and ΔS values for NN base pairs (A), internal mismatches (B) and dangling ends (C).

A	NN pairs	**ΔH** (kcal mol^{-1})	**ΔS** (cal K^{-1} mol^{-1})
	AA_TT	-7.6	-21.3
	AT_TA	-7.2	-20.4
	TA_AT	-7.2	-21.3
	CA_GT	-8.5	-22.7
	GT_CA	-8.4	-22.4
	CT_GA	-7.8	-21.0
	GA_CT	-8.2	-22.2
	CG_GC	-10.6	-27.2
	GG_CC	-8.0	-19.9
	initiation	0.2	-5.7
	terminal A_T	2.2	6.9
	symmetry	0.0	-1.4

B	NN pairs	**ΔH** (kcal mol^{-1})	**ΔS** (cal K^{-1} mol^{-1})
	AA_TA	1.2	1.7
	CA_GA	-0.9	-4.2
	GA_CA	-2.9	-9.8
	TA_AA	4.7	12.9
	AC_TC	0.0	-4.4
	CC_GC	-1.5	-7.2
	GC_CC	3.6	8.9
	TC_AC	6.1	16.4
	AG_TG	-3.1	-9.5
	CG_GG	-4.9	-15.3
	GG_CG	-6.0	-15.8
	TG_AG	1.6	3.6
	AT_TT	-2.7	-10.8
	CT_GT	-5.0	-15.8
	GT_CT	-2.2	-8.4
	TT_AT	0.2	-1.5
	AG_TT	1.0	0.9
	AT_TG	-2.5	-8.3
	CG_GT	-4.1	-11.7
	CT_GG	-2.8	-8.0
	GG_CT	3.3	10.4
	GT_CG	-4.4	-12.3
	TG_AT	-0.1	-1.7
	TT_AG	-1.3	-5.3
	AA_TC	2.3	4.6
	AC_TA	5.3	14.6
	CA_GC	1.9	3.7
	CC_GA	0.6	-0.6
	GA_CC	5.2	14.2
	GC_CA	-0.7	-3.8
	TA_AC	3.4	8.0
	TC_AA	7.6	20.2
	AA_TG	-0.6	-2.3
	AG_TA	-0.7	-2.3
	CA_GG	-0.7	-2.3
	CG_GA	-4.0	-13.2
	GA_CG	-0.6	-1.0
	GG_CA	0.5	3.2
	TA_AG	0.7	0.7
	TG_AA	3.0	7.4
	AC_TT	0.7	0.2
	AT_TC	-1.2	-6.2
	CC_GT	-0.8	-4.5
	CT_GC	-1.5	-6.1
	GC_CT	2.3	5.4
	GT_CC	5.2	13.5
	TC_AT	1.2	0.7
	TT_AC	1.0	0.7

C	Dangling ends	**ΔH** (kcal mol^{-1})	**ΔS** (cal K^{-1} mol^{-1})
	AA_XT	0.2	2.3
	AC_XG	-6.3	-17.1
	AG_XC	-3.7	-10.0
	AT_XA	-2.9	-7.6
	CA_XT	0.6	3.3
	CC_XG	-4.4	-12.6
	CG_XC	-4.0	-11.9
	CT_XA	-4.1	-13.0
	GA_XT	-1.1	-1.6
	GC_XG	-5.1	-14.0
	GG_XC	-3.9	-10.9
	GT_XA	-4.2	-15.0
	TA_XT	-6.9	-20.0
	TC_XG	-4.0	-10.9
	TG_XC	-4.9	-13.8
	TT_XA	-0.2	-0.5
	XA_AT	-0.7	-0.8
	XC_AG	-2.1	-3.9
	XG_AC	-5.9	-16.5
	XT_AA	-0.5	-1.1
	XA_CT	4.4	14.9
	XC_CG	-0.2	-0.1
	XG_CC	-2.6	-7.4
	XT_CA	4.7	14.2
	XA_GT	-1.6	-3.6
	XC_GG	-3.9	-11.2
	XG_GC	-3.2	-10.4
	XT_GA	-4.1	-13.1
	XA_TT	2.9	10.4
	XC_TG	-4.4	-13.1
	XG_TC	-5.2	-15.0
	XT_TA	-3.8	-12.6

Table A2.2: ΔH and ΔS values for loop-length correction (A), tri-loop bonus (B) and tetra-loop bonus (C).

A

Loop length	ΔH (kcal mol^{-1})	ΔS (cal K^{-1} mol^{-1})
3	0.0	-11.3
4	0.0	-11.3
5	0.0	-10.6
6	0.0	-12.9
7	0.0	-13.5
8	0.0	-13.9
9	0.0	-14.5
10	0.0	-14.8
11	0.0	-15.5
12	0.0	-16.1
13	0.0	-16.4
14	0.0	-16.4
15	0.0	-16.8
16	0.0	-17.1
17	0.0	-17.4
18	0.0	-17.7
19	0.0	-18.1
20	0.0	-18.4

B

Tri-loops	ΔH (kcal mol^{-1})	ΔS (cal K^{-1} mol^{-1})
AGAAT	-1.5	0.0
AGCAT	-1.5	0.0
AGGAT	-1.5	0.0
AGTAT	-1.5	0.0
CGAAG	-2.0	0.0
CGCAG	-2.0	0.0
CGGAG	-2.0	0.0
CGTAG	-2.0	0.0
GGAAC	-2.0	0.0
GGCAC	-2.0	0.0
GGGAC	-2.0	0.0
GGTAC	-2.0	0.0
TGAAA	-1.5	0.0
TGCAA	-1.5	0.0
TGGAA	-1.5	0.0
TGTAA	-1.5	0.0

C

Tetra-loops	ΔH (kcal mol^{-1})	ΔS (cal K^{-1} mol^{-1})	Tetra-loops	ΔH (kcal mol^{-1})	ΔS (cal K^{-1} mol^{-1})
AAAAAT	0.5	-0.6	GAAAAT	0.5	-3.2
AAAACT	0.7	1.6	GAAACT	1.0	0.0
AAACAT	1.0	1.6	GAACAT	1.0	0.0
ACTTGT	0.0	4.2	GCTTGT	0.0	1.6
AGAAAT	-1.1	1.6	GGAAAT	-1.1	0.0
AGAGAT	-1.1	1.6	GGAGAT	-1.1	0.0
AGATAT	-1.5	1.6	GGATAT	-1.6	0.0
AGCAAT	-1.6	1.6	GGCAAT	-1.6	0.0
AGCGAT	-1.1	1.6	GGCGAT	-1.1	0.0
AGCTTT	0.2	1.6	GGCTTT	-0.1	0.0
AGGAAT	-1.1	1.6	GGGAAT	-1.1	0.0
AGGGAT	-1.1	1.6	GGGGAT	-1.1	0.0
AGGGGT	0.5	0.6	GGGGGT	0.5	-1.0
AGTAAT	-1.6	1.6	GGTAAT	-1.6	0.0
AGTGAT	-1.1	1.6	GGTGAT	-1.1	0.0
AGTTCT	0.8	1.6	GTATAT	-0.5	0.0
ATTCGT	-0.2	1.6	GTTCGT	-0.4	0.0
ATTTGT	0.0	1.6	GTTTGT	-0.4	0.0
ATTTTT	-0.5	1.6	GTTTTT	-0.5	0.0
CAAAAG	0.5	-1.3	TAAAAA	0.5	0.3
CAAACG	0.7	0.0	TAAACA	0.7	1.6
CAACAG	1.0	0.0	TAACAA	1.0	1.6
CAACCG	0.0	0.0	TCTTGA	0.0	4.2
CCTTGG	0.0	2.6	TGAAAA	-1.1	1.6
CGAAAG	-1.1	0.0	TGAGAA	-1.1	1.6
CGAGAG	-1.1	0.0	TGATAA	-1.6	1.6
CGATAG	-1.5	0.0	TGCAAA	-1.6	1.6
CGCAAG	-1.6	0.0	TGCGAA	-1.1	1.6
CGCGAG	-1.1	0.0	TGCTTA	0.2	1.6
CGCTTG	0.2	0.0	TGGAAA	-1.1	1.6
CGGAAG	-1.1	0.0	TGGGAA	-1.1	1.6
CGGGAG	-1.0	0.0	TGGGGA	0.5	0.6
CGGGGG	0.5	-1.0	TGTAAA	-1.6	1.6
CGTAAG	-1.6	0.0	TGTGAA	-1.1	1.6
CGTGAG	-1.1	0.0	TGTTCA	0.8	1.6
CGTTCG	0.8	0.0	TTTCGA	-0.2	1.6
CTTCGG	-0.2	0.0	TTTTGA	0.0	1.6
CTTTGG	0.0	0.0	TTTTTA	-0.5	1.6
CTTTTG	-0.5	0.0	TAAAAG	0.5	-1.6
GAAAAC	0.5	-3.2	TAAACG	1.0	1.6
GAAACC	0.7	0.0	TAACAG	1.0	1.6
GAACAC	1.0	0.0	TCTTGG	0.0	3.2
GCTTGC	0.0	2.6	TGAAAG	-1.0	1.6
GGAAAC	-1.1	0.0	TGAGAG	-1.0	1.6
GGAGAC	-1.1	0.0	TGATAG	-1.5	1.6
GGATAC	-1.6	0.0	TGCAAG	-1.5	1.6
GGCAAC	-1.6	0.0	TGCGAG	-1.0	1.6
GGCGAC	-1.1	0.0	TGCTTG	-0.1	1.6
GGCTTC	0.2	0.0	TGGAAG	-1.0	1.6
GGGAAC	-1.1	0.0	TGGGAG	-1.0	1.6
GGGGAC	-1.1	0.0	TGGGGG	0.5	0.6
GGGGGC	0.5	-1.0	TGTAAG	-1.5	1.6
GGTAAC	-1.6	0.0	TGTGAG	-1.0	1.6
GGTGAC	-1.1	0.0	TTTCGG	-0.4	1.6
GGTTCC	0.8	0.0	TTTTAG	-1.0	1.6
GTTCGC	-0.2	0.0	TTTTGG	-0.4	1.6
GTTTGC	0.0	0.0	TTTTTG	-0.5	1.6
GTTTTC	-0.5	0.0			

Table A2.3: ΔH and ΔS values for loop terminal mismatches.

NN pairs	**ΔH** (kcal mol^{-1})	**ΔS** (cal K^{-1} mol^{-1})	**NN pairs**	**ΔH** (kcal mol^{-1})	**ΔS** (cal K^{-1} mol^{-1})
AA_TA	-3.2	-8.1	GC_TA	0.0	0.6
AA_TC	-0.9	-1.9	GC_TC	0.0	0.6
AA_TG	-2.3	-5.8	GC_TG	-4.5	-11.6
AC_TA	-2.2	-5.2	GC_TT	0.0	0.6
AC_TC	-0.5	-1.0	GG_TA	0.0	1.6
AC_TT	-1.2	-2.9	GG_TC	-5.9	-16.1
AG_TA	-2.7	-6.8	GG_TG	0.0	1.6
AG_TG	-1.3	-2.9	GG_TT	-2.0	-4.8
AG_TT	-2.9	-7.7	GT_TA	-3.5	-9.7
AT_TC	-2.8	-8.1	GT_TC	0.0	0.6
AT_TG	-3.5	-9.7	GT_TG	-2.0	-4.8
AT_TT	-2.4	-6.4	GT_TT	0.0	0.6
CA_GA	-2.8	-5.8	TA_AA	-2.7	-6.8
CA_GC	-2.0	-3.9	TA_AC	-2.6	-7.1
CA_GG	-3.0	-6.8	TA_AG	-2.4	-6.1
CC_GA	-2.4	-5.2	TC_AA	-2.6	-6.8
CC_GC	-1.4	-2.9	TC_AC	-0.5	-1.0
CC_GT	-2.4	-5.5	TC_AT	-2.7	-7.1
CG_GA	-5.1	-13.2	TG_AA	-1.9	-4.2
CG_GG	-2.9	-6.4	TG_AG	-1.5	-3.5
CG_GT	-2.9	-6.1	TG_AT	-2.3	-5.8
CT_GC	-3.1	-8.1	TT_AC	-1.4	-3.5
CT_GG	-5.9	-16.1	TT_AG	-3.7	-10
CT_GT	-5.3	-14.2	TT_AT	-2.3	-6.4
GA_CA	-6.0	-16.1	TA_GA	0.0	1.6
GA_CC	-4.0	-10.6	TA_GC	0.0	0.6
GA_CG	-3.4	-8.4	TA_GG	0.0	1.6
GC_CA	-2.8	-5.8	TA_GT	-2.3	-5.8
GC_CC	-2.5	-6.1	TC_GA	0.0	0.6
GC_CT	-3.3	-8.4	TC_GC	0.0	0.6
GG_CA	-4.0	-9.7	TC_GG	-3.7	-9.4
GG_CG	-5.3	-13.9	TC_GT	0.0	0.6
GG_CT	-3.7	-9.4	TG_GA	0.0	1.6
GT_CC	-2.5	-6.1	TG_GC	-2.9	-6.1
GT_CG	-4.5	-11.6	TG_GG	0.0	1.6
GT_CT	-6.1	-16.8	TG_GT	-2.0	-4.8
GA_TA	0.0	1.6	TT_GA	-2.9	-7.7
GA_TC	0.0	0.6	TT_GC	0.0	0.6
GA_TG	0.0	1.6	TT_GG	-2.0	-4.8
GA_TT	-3.7	-10.0	TT_GT	0.0	0.6

2.A.2 Melting temperature calculation

T_m is calculated with the formula reported from SantaLucia and Hicks (2004). The following formula calculates the T_m at 1 M Na^+ condition.

$$T_m(1\,\mathrm{M}\,\mathrm{Na}^+) = 1000 \times \frac{\Delta H^o}{\Delta S^o + R\ln(\frac{C_T}{x})} \tag{2.1}$$

where:

$T_m(1\,\mathrm{M}\,\mathrm{Na}^+)$ = melting temperature at 1 M Na^+ conditions (K).

ΔH^o = enthalpy (kcal mol^{-1}).

R = gas constant, 1.9872 (cal K^{-1} mol^{-1}).

ΔS^o = entropy (cal K^{-1} mol^{-1}).

C_T = total molar strand concentration (mol L^{-1}). It is approximated to the concentration of the oligos (forward, reverse and probe concentration). It does not consider the template concentration because it is supposed to be significantly lower than the oligo concentration.

x = 4 for nonself-complementary duplexes (e.g. primers) and 1 for self-complementary duplexes.

In order to apply the equation 2.1, we need first to calculate ΔH and ΔS which are the sum of all the ΔH and ΔS that influence the annealing temperature of the oligos to the DNA template. For each value, the following components are separately summed up (Figure A2.1):

- ΔH and ΔS initiation values which consider the fact that duplex initiation is not likely to happen, the ΔG is in fact positive; marked as ΔH^{*1} and ΔS^{*1} in Figure A2.1 (Table A2.1A).
- ΔH and ΔS terminal values in case the oligo duplex ends with an adenine or a thymine; marked as ΔH^{*2} and ΔS^{*2} in Figure A2.1 (Table A2.1A).
- ΔH and ΔS values for all the nearest-neighbor base pairs; marked as ΔH^{*3} and ΔS^{*3} in Figure A2.1 (Table A2.1).

Also, because of how the PCR works, the forward and the reverse primers are considered to have a blunt-end (e.g. $\frac{A\,G}{T\,C}$) at the 5'-end terminal and a dangling end (e.g. $\frac{A}{T\,C}$; Table A2.1C) at the 3' terminal (Figure A2.1A-B). The probes are considered to terminate with both dangling ends (Figure A2.1C). This is an approximation as i) for the first PCR steps the primer's ends terminate both in dangling ends and ii) the dangling base depends on the DNA template the oligo anneals to. PhyloPrimer can predict the dangling ends from the consensus sequence as it

```
               5'-ATGACTCCTCGATCAGGTCGTGGCAGATAGTAGGCTAACGGATAGCATTGCATTGCTGACCAGTTGAGTCCAGGAT-3
CONSENSUS                                                                GACTGGTCAACTCAGGTC
                     ACTCCTCGATCAGGTCGT         TAGGCTAACGGATAGCATTG
               3'-TACTGAGGAGCTAGTCCAGCACCGTCTATCATCCGATTGCCTATCGTAACGTAACGACTGGTCAACTCAGGTCCTA-5
```

A FORWARD PRIMER

```
5'-ACTCCTCGATCAGGTCGT –3' forward primer
3'-TGAGGAGCTAGTCCAGCAC-5' consensus
```

ΔH = ΔH(initiation)^{*1} + ΔH(terminal_AT)^{*2} + ΔH(AC/TG)^{*3} + ΔH(CT/GA)^{*3} + ΔH(TC/AG)^{*3} + ΔH(CC/GG)^{*3} + ΔH(CT/GA)^{*3} + ΔH(TC/AG)^{*3} + ΔH(CG/GC)^{*3} + ΔH(GA/CT)^{*3} + ΔH(AT/TA)^{*3} + ΔH(TC/AG)^{*3} + ΔH(CA/GT)^{*3} + ΔH(AG/TC)^{*3} + ΔH(GG/CC)^{*3} + ΔH(GT/CA)^{*3} + ΔH(TC/AG)^{*3} + ΔH(CG/GC)^{*3} + ΔH(GT/CA)^{*3} + ΔH(TX/AC)^{*3} = +0.2 +2.2 -8.4 -7.8 -8.2 -8 -7.8 -8.2 -10.6 -8.2 -7.2 -8.2 -8.5 -7.8 -8 -8.4 -8.2 -10.6 -8.4 +0.6 = -139.5 kcal mol^{-1}

ΔS = ΔS(initiation)^{*1} + ΔS(terminal_AT)^{*2} + ΔS(AC/TG)^{*3} + ΔS(CT/GA)^{*3} + ΔS(TC/AG)^{*3} + ΔS(CC/GG)^{*3} + ΔS(CT/GA)^{*3} + ΔS(TC/AG)^{*3} + ΔS(CG/GC)^{*3} + ΔS(GA/CT)^{*3} + ΔS(AT/TA)^{*3} + ΔS(TC/AG)^{*3} + ΔS(CA/GT)^{*3} + ΔS(AG/TC)^{*3} + ΔS(GG/CC)^{*3} + ΔS(GT/CA)^{*3} + ΔS(TC/AG)^{*3} + ΔS(CG/GC)^{*3} + ΔS(GT/CA)^{*3} + ΔS(TX/AC)^{*3} = -5.7 +6.9 -22.4 -21 -22.2 -19.9 -21 -22.2 -27.2 -22.2 -20.4 -22.2 -22.7 -21 -19.9 -22.4 -22.2 -27.2 -22.4 +3.3 = -374 cal K^{-1} mol^{-1}

T$_m$(1 M Na$^+$) = 1000 x (-139.5/(-374 + 1.9872 x ln(0.000002/4))) = **346.3 K (73.1 °C)**

B REVERSE PRIMER

```
5'-CTGGACTCAACTGGTCAG—3' reverse primer
3'-GACCTGAGTTGACCAGTCG-5' consensus
```

ΔH = ΔH(initiation)^{*1} + ΔH(CT/GA)^{*3} + ΔH(TG/AC)^{*3} + ΔH(GG/CC)^{*3} + ΔH(GA/CT)^{*3} + ΔH(AC/TG)^{*3} + ΔH(CT/GA)^{*3} + ΔH(TC/AG)^{*3} + ΔH(CA/GT)^{*3} + ΔH(AA/TT)^{*3} + ΔH(AC/TG)^{*3} + ΔH(CT/GA)^{*3} + ΔH(TG/AC)^{*3} + ΔH(GG/CC)^{*3} + ΔH(GT/CA)^{*3} + ΔH(TC/AG)^{*3} + ΔH(CA/GT)^{*3} + ΔH(AG/TC)^{*3} + ΔH(GX/CG)^{*3} = +0.2 -7.8 -8.5 -8 -8.2 -8.4 -7.8 -8.2 -8.5 -7.6 -8.4 -7.8 -8.5 -8 -8.4 -8.2 -8.5 -7.8 -5.1 = -143.5 kcal mol^{-1}

ΔS = ΔS(initiation)^{*1} + ΔS(CT/GA)^{*3} + ΔS(TG/AC)^{*3} + ΔS(GG/CC)^{*3} + ΔS(GA/CT)^{*3} + ΔS(AC/TG)^{*3} + ΔS(CT/GA)^{*3} + ΔS(TC/AG)^{*3} + ΔS(CA/GT)^{*3} + ΔS(AA/TT)^{*3} + ΔS(AC/TG)^{*3} + ΔS(CT/GA)^{*3} + ΔS(TG/AC)^{*3} + ΔS(GG/CC)^{*3} + ΔS(GT/CA)^{*3} + ΔS(TC/AG)^{*3} + ΔS(CA/GT)^{*3} + ΔS(AG/TC)^{*3} + ΔS(GX/CG)^{*3} = -5.7 -21 -22.7 -19.9 -22.2 -22.4 -21 -22.2 -22.7 -21.3 -22.4 -21 -22.7 -19.9 -22.4 -22.2 -22.7 -21 -14 = -389.4 cal K^{-1} mol^{-1}

T$_m$(1 M Na$^+$) = 1000 x (-143.5/(-389.4 + 1.9872 x ln(0.000002/4))) = **343.1 K (69.95 °C)**

C PROBE

```
5'- TAGGCTAACGGATAGCATTG -3' anti-sense probe
3'-CATCCGATTGCCTATCGTAACG-5' consensus
```

ΔH = ΔH(initiation)^{*1} + ΔH(XT_CA)^{*3} + ΔH(TA/AT)^{*3} + ΔH(AG/TC)^{*3} + ΔH(GG/CC)^{*3} + ΔH(GC/CG)^{*3} + ΔH(CT/GA)^{*3} + ΔH(TA/AT)^{*3} + ΔH(AA/TT)^{*3} + ΔH(AC/TG)^{*3} + ΔH(CG/GC)^{*3} + ΔH(GG/CC)^{*3} + ΔH(GA/CT)^{*3} + ΔH(AT/TA)^{*3} + ΔH(TA/AT)^{*3} + ΔH(AG/TC)^{*3} + ΔH(GC/CG)^{*3} + ΔH(CA/GT)^{*3} + ΔH(AT/TA)^{*3} + ΔH(TT/AA)^{*3} + ΔH(TG/AC)^{*3} + ΔH(GX/CT)^{*3} = +0.2 +4.7 –7.2 -7.8 -8 -9.8 -7.8 -7.2 -7.6 -8.4 -10.6 -8 -8.2 -7.2 -7.2 -7.8 -9.8 -8.5 -7.2 -7.6 -8.5 -5.1 = -154.6 kcal mol^{-1}

ΔS = ΔS(initiation)^{*1} + ΔS(XT/CA)^{*3} + ΔS(TA/AT)^{*3} + ΔS(AG/TC)^{*3} + ΔS(GG/CC)^{*3} + ΔS(GC/CG)^{*3} + ΔS(CT/GA)^{*3} + ΔS(TA/AT)^{*3} + ΔS(AA/TT)^{*3} + ΔS(AC/TG)^{*3} + ΔS(CG/GC)^{*3} + ΔS(GG/CC)^{*3} + ΔS(GA/CT)^{*3} + ΔS(AT/TA)^{*3} + ΔS(TA/AT)^{*3} + ΔS(AG/TC)^{*3} + ΔS(GC/CG)^{*3} + ΔS(CA/GT)^{*3} + ΔS(AT/TA)^{*3} + ΔS(TT/AA)^{*3} + ΔS(TG/AC)^{*3} + ΔS(GX/CT)^{*3} = -5.7 +14.2 -21.3 -21 -19.9 -24.4 -21 -21.3 -21.3 -22.4 -27.2 -19.9 -22.2 -20.4 -21.3 -21 -24.4 -22.7 -20.4 -21.3 -22.7 -14 = -421.6 cal K^{-1} mol^{-1}

T$_m$(1 M Na$^+$) = 1000 x (-154.6/(-421.6 + 1.9872 x ln(0.000002/4))) = **343.2 K (70.05 °C)**

Figure A2.1: A schematic view of how the melting temperature is calculated for forward primers (A), reverse primers (B) and probes (C). ΔH^{*1} and ΔS^{*1} refer to duplex initiation values; ΔH^{*2} and ΔS^{*2} refer to A and T terminals; ΔH^{*3} and ΔS^{*3} refer to NN base pairs.

was used for the oligo design but the DNA template may vary during the PCR. In the Oligo Check mode PhyloPrimer considers all oligos to end with two blunt-ends as it does not know the DNA template. When degenerate bases are present in the DNA duplexes, the T_m is calculated for all the oligo variants and only the average T_m is reported.

2.A.3 ΔG calculation

The Gibbs free energy equation used to calculate ΔG is:

$$\Delta G^o_T = \Delta H^o - T\Delta S^o \tag{2.2}$$

where:

ΔG^o = Gibbs free energy (kcal mol^{-1})

T = temperature (K). Ideally this values should be set to the one that is used during the extension phase of the amplification. The two standard values used for the ΔG calculation are 298.15 K (25 °C) and 308.15 K (37 °C). However, PhyloPrimer sets 72 °C as default value as it corresponds to a common annealing temperature in the PCR.

For each secondary structure, the ΔG is calculated on the longest stretch of consecutive bases containing only single mismatches. If an oligo presents more than one stretch that reflects this property, the ΔG will be calculated for each stretch and the final ΔG will be attributed to the lowest ΔG values (Figure A2.2B-F). As seen for the T_m calculation, before applying the equation 2.2, ΔH and ΔS must be calculated first. These two values are obtained by the sum of different aspects depending which secondary structure is taken into analyses. Enthalpies and entropies for self-dimers (Figure A2.2A-B) and cross-dimers (Figure A2.2C) are calculated as follows:

- ΔH and ΔS initiation values which consider the fact that duplex initiation is not likely to happen, the ΔG is positive; marked as ΔH^{*1} and ΔS^{*1} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS terminal values in case the oligo duplex ends with an adenine or a thymine; marked as ΔH^{*2} and ΔS^{*2} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS values for all the nearest-neighbor base pairs; marked as ΔH^{*3} and ΔS^{*3} in Figure A2.2 (Table A2.1).
- ΔH and ΔS values that accounts for symmetry; marked as ΔH^{*4} and ΔS^{*4} in Figure A2.2 (Table A2.1A). They are applied only to self-dimers.

In the case of the ΔG calculation for hairpins loops there are different aspects to take in consideration which vary depending on the length of the hairpin loop. Hairpin loops of less than 3 bases are not considered thermodynamically stable and therefore are not taken in consideration. Furthermore, certain hairpin loops that are 3 or 4 bases are particularly stable and thermodynamics corrections are applied to them. In case the hairpin loop is 3-base long (Figure A2.2D), PhyloPrimer takes in consideration the following ΔH and ΔS values:

- ΔH and ΔS initiation values which consider the fact that duplex initiation is not likely to happen, the ΔG is positive; marked as ΔH^{*1} and ΔS^{*1} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS terminal values in case the oligo duplex ends with an adenine or a thymine at both the duplex extremities; marked as ΔH^{*2} and ΔS^{*2} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS values for all the nearest-neighbor base pairs; marked as ΔH^{*3} and ΔS^{*3} in Figure A2.2 (Table A2.1).
- ΔH and ΔS values for 3-base long loops; marked as ΔH^{*5} and ΔS^{*5} in Figure A2.2 (Table A2.2A).
- ΔH and ΔS bonus values for particular 3-base loop sequences; marked as ΔH^{*6} and ΔS^{*6} in Figure A2.2 (Table A2.2B).

In case the hairpin loop is 4-base long (Figure A2.2E), PhyloPrimer takes in consideration the following ΔH and ΔS values:

- ΔH and ΔS initiation values which consider the fact that duplex initiation is not likely to happen, the ΔG is positive; marked as ΔH^{*1} and ΔS^{*1} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS terminal values in case the oligo duplex ends with an adenine or a thymine not at the loop extremities; marked as ΔH^{*2} and ΔS^{*2} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS values for all the nearest-neighbor base pairs; marked as ΔH^{*3} and ΔS^{*3} in Figure A2.2 (Table A2.1).
- ΔH and ΔS values for 4-base long loops; marked as ΔH^{*5} and ΔS^{*5} in Figure A2.2 (Table A2.2A).
- ΔH and ΔS bonus values for particular 4-base loop sequences; marked as ΔH^{*6} and ΔS^{*6} in Figure A2.2 (Table A2.2C).
- ΔH and ΔS values on the terminal pair that closes to the loop; marked as ΔH^{*7} and ΔS^{*7} in Figure A2.2 (Table A2.3).

In case the hairpin loop is longer than 4 bases (Figure A2.2F), PhyloPrimer takes in consideration

A SELF-DIMERS – case 1

```
A C T C C T C G A T C A G G T C G T
            | | |   |   |   | | |
          T G C T G G A C T A G C T C C T C A
```

ΔH = ΔH(initiation)^{*1} + ΔH(CG_GC)^{*3} + ΔH(GA_CT)^{*3} + ΔH(AT_TG)^{*3} + ΔH(TC_GG)^{*3} + ΔH(CA_GA)^{*3} + ΔH(AG_AC)^{*3} + ΔH(GG_CT) ^{*3} + ΔH(GT_TA)^{*3} + ΔH(TC_AG)^{*3} + ΔH(CG_GC)^{*3} + ΔH(symmetry)^{*4} = +0.2 -10.6 -8.2 -2.5 +3.3 -0.9 -0.9 +3.3 -2.5 -8.2 -10.6 +0 = -37.6 kcal mol^{-1}

ΔS = ΔH(initiation)^{*1} + ΔS(CG_GC)^{*3} + ΔS(GA_CT)^{*3} + ΔS(AT_TG)^{*3} + ΔS(TC_GG)^{*3} + ΔS(CA_GA)^{*3} + ΔS(AG_AC)^{*3} + ΔS(GG_CT)^{*3} + ΔS(GT_TA)^{*3} + ΔS(TC_AG)^{*3} + ΔS(CG_GC)^{*3} + ΔS(symmetry)^{*4} = -5.7 -27.2 – 22.2 -8.3 +10.4 -4.2 -4.2 +10.4 -8.3 -22.2 -27.2 -1.4 = 110.1 cal K^{-1} mol^{-1}

ΔG = -37.6 –(298.15 x -110.1/1000) = **-4.77 kcal mol^{-1}**

B SELF-DIMERS – case 2

```
A C T C C T C G A T C A G G T C G T
¦ ¦     ¦     | | | |     ¦     ¦ ¦
T G C T G G A C T A G C T C C T C A
```

ΔH = ΔH = ΔH(initiation)^{*1} + ΔH(AC_TG)^{*3} + ΔH(terminal_AT)^{*3} + ΔH(symmetry)^{*4} = +0.2 -8.4 +2.2 +0 = -6 kcal mol^{-1}

ΔS = ΔS = ΔS(initiation)^{*1} + ΔS(AC_TG) ^{*3} + ΔS(terminal_AT)^{*3} + ΔS(symmetry)^{*4} = -5.7 -22.4 +6.9 -1.4 = -22.6 cal K^{-1} mol^{-1}

ΔG = ΔG = -6 –(298.15 x -22.6/1000) = +0.73 kcal x mol^{-1}

ΔH = ΔH(initiation)^{*1} + ΔH(GA_CT)^{*3} + ΔH(AT_TA)^{*3} + ΔH(TC_AG)^{*3} + ΔH(symmetry)^{*4} = +0.2 -8.2 -7.2 -8.2 +0 = -23.4 kcal mol^{-1}

ΔS = ΔS(initiation)^{*1} + ΔS(GA_CT)^{*3} + ΔS(AT_TA)^{*3} + ΔS(TC_AG)^{*3} + ΔS(symmetry)^{*4} = -5.7 -22.2 -20.4 -22.2 -1.4 = -71.9 cal K^{-1} mol^{-1}

ΔG = -23.4 –(298.15 x -71.9/1000) = -1.9 kcal mol^{-1}

ΔG = ΔG = **-1.9 kcal mol^{-1}**

C CROSS-DIMERS – case 1

```
  A C T C G A C G A T C C A G C A G A A C
      | | | | | |
G G T A G C T G C G G T A G T A G T C G T
```

ΔH = ΔH(initiation)^{*1} + ΔH(terminal_AT)^{*2} + ΔH(TC_AG)^{*3} + ΔH(CG_GC)^{*3} + ΔH(GA_CT)^{*3} + ΔH(AC_TG)^{*3} + ΔH(CG_GC)^{*3} = +0.2 +2.2 -8.2 -10.6 -8.2 -8.4 -10.6 = -43.6 kcal mol^{-1}

ΔS = ΔS(initiation)^{*1} + ΔS(terminal_AT)^{*2} + ΔS(TC_AG)^{*3} + ΔS(CG_GC)^{*3} + ΔS(GA_CT)^{*3} + ΔS(AC_TG)^{*3} + ΔS(CG_GC)^{*3} = -5.7 +6.9 -22.2 -27.2 -22.2 -22.4 -27.2 = -120 cal K^{-1} mol^{-1}

ΔG = -43.6 –(298.15 x -120/1000) = **-7.8 kcal mol^{-1}**

D HAIRPINS – case 1

```
C A C G A C G A G G \
  | |   | | |   |     A
  T G T T G C G C T /
```

ΔH = ΔH(CA_XT)^{*3} + ΔH(AC_TG)^{*3} + ΔH(CG_GT)^{*3} + ΔH(GA_TT)^{*3} + ΔH(AC_TG)^{*3} + ΔH(CG_GC)^{*3} + ΔH(loop_3)^{*5} + ΔH(GGATC)^{*3} = +0.6 -8.4 -4.1 -1.3 -8.4 -10.6 +0 +0 = -32.2 kcal mol^{-1}

ΔS = ΔS(CA_XT)^{*3} + ΔS(AC_TG)^{*3} + ΔS(CG_GT)^{*3} + ΔS(GA_TT)^{*3} + ΔS(AC_TG)^{*3} + ΔS(CG_GC)^{*3} + ΔS(loop_3)^{*5} + ΔS(GGATC)^{*6} = +3.3 -22.4 -11.7 -5.3 -22.4 -27.2 -11.3 +0 = -97 = -97 cal K^{-1} mol^{-1}

ΔG = -32.2 –(298.15*-97/1000) = **-3.279 kcal mol^{-1}**

E HAIRPINS – case 2

```
C G A T G C C G A A G A \
      | | |   | | |
      A C G T C T T A A /
```

ΔH = ΔH(AT_XA)^{*3} + ΔH(TG_AC)^{*3} + ΔH(GC_CG)^{*3} + ΔH(CC_GT)^{*3} + ΔH(CG_TC)^{*3} + ΔH(GA_CT)^{*3} + ΔH(AA_TT)^{*3} + ΔH(loop_4)^{*3} + ΔH(AG_TG_terminal)^{*7} + ΔH(AGAAAT)^{*6} = -2.9 -8.5 -9.8 -0.8 -1.5 -8.2 -7.6 +0 -1.1 -2.7 = -43.1 kcal x mol^{-1}

ΔS = ΔS(AT_XA)^{*3} + ΔS(TG_AC)^{*3} + ΔS(GC_CG)^{*3} + ΔS(CC_GT)^{*3} + ΔS(CG_TC)^{*3} + ΔS(GA_CT)^{*3} + ΔS(AA_TT)^{*3} + ΔS(loop_4)^{*3} + ΔS(AG_TG_terminal)^{*7} + ΔS(AGAAAT)^{*6} = -7.6 -22.7 -24.4 -4.5 -6.1 -22.2 -21.3 -11.3 -1.6 -6.7 = -128.4 cal K^{-1} mol^{-1}

ΔG = -43.1 –(298.15*-128.4/1000) = **-4.817 kcal mol^{-1}**

F HAIRPINS – case 3

```
G T C A T G G C G T C G \
¦ ¦     | | | | |
C A T G A C C G C G T A /
```

ΔH = ΔH(terminal_AT)^{*2} + ΔH(TG_AC)^{*3} + ΔH(GG_CC)^{*3} + ΔH(GC_CG)^{*3} + ΔH(CG_GC)^{*3} + ΔH(loop_6)^{*5} + ΔH(GT_CG_terminal)^{*7} = +2.2 -8.5 -8 -9.8 -10.6 +0 -4.5 = -39.2 kcal mol^{-1}

ΔS = ΔS(terminal_AT)^{*2} + ΔS(TG_AC)^{*3} + ΔS(GG_CC)^{*3} + ΔS(GC_CG)^{*3} + ΔS(CG_GC)^{*3} + ΔS(loop_6)^{*5} + ΔS(GT_CG_terminal)^{*7} = +6.9 -22.7 -19.9 -24.4 -27.2 -12.9 -11.6 = -111.8 = -111.8/1000 = -0.1118 cal K^{-1} mol^{-1}

ΔG = -39.2 –(298.15*-0.1118) = -5.86 kcal mol^{-1}

ΔH = ΔH(GT_CA)^{*3} + ΔH(loop_20)^{*5} + ΔH(TC_AT_terminal)^{*7} = -8.4 +0 -2.7 = -11.1 kcal mol^{-1}

ΔS = ΔS(GT_CA)^{*3} + ΔS(loop_20)^{*5} + ΔS(TC_AT_terminal)^{*7} = -22.4 -18.4 -7.1 = -47.9 cal K^{-1} mol^{-1}

ΔG = -11.1 –(298.15 x -47.9/1000) = +3.18 kcal mol^{-1}

ΔG = ΔG = **-5.86 kcal mol^{-1}**

Figure A2.2: A schematic view of how the ΔG is calculated for self-dimers (A and B), cross-dimers (C) and hairpin loops (D, E and F). ΔH^{*1} and ΔS^{*1} refer to duplex initiation values; ΔH^{*2} and ΔS^{*2} refer to A and T terminals; ΔH^{*3} and ΔS^{*3} refer to NN base pairs; ΔH^{*4} and ΔS^{*4} refer to symmetry bonus; ΔH^{*5} and ΔS^{*5} refer to loop length corrections; ΔH^{*6} and ΔS^{*6} refer to the 3- and 4- long loop bonus; ΔH^{*7} and ΔS^{*7} refer to loop terminal mismatches. The ΔG was calculated considering a temperature of 25 °C (298.15 K).

the following ΔH and ΔS values:

- ΔH and ΔS initiation values which consider the fact that duplex initiation is not likely to happen, the ΔG is positive; marked as ΔH^{*1} and ΔS^{*1} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS terminal values in case the oligo duplex ends with an adenine or a thymine not at the loop extremities; marked as ΔH^{*2} and ΔS^{*2} in Figure A2.2 (Table A2.1A).
- ΔH and ΔS values for all the nearest-neighbor base pairs; marked as ΔH^{*3} and ΔS^{*3} in Figure A2.2 (Table A2.1).
- ΔH and ΔS values for n-base long loops; marked as ΔH^{*5} and ΔS^{*5} in Figure A2.2 (Table A2.2A).
- ΔH and ΔS values on the terminal pair that closes to the loop; marked as ΔH^{*7} and ΔS^{*7} in Figure A2.2 (Table A2.3).

The hairpin-like duplex will have 2 ends: the one terminating into the loop and the other one (on the other extremity). As we saw above, the loop termination is treated differently depending on the loop size, the other end behaves as we saw for the self- and cross-dimers where it can be a blunt-end or a dangling end.

2.A.4 Formula corrections

2.A.4.1 Monovalent ions and Mg^{2+}

The most used monovalent ions in the PCR reactions are $Tris^{+}$, K^{+} and Na^{+}. It has been observed that all the monovalent ions act the same for the duplex stabilization. Therefore, they will be all considered as monovalent ions. The concentration of monovalent ions is usually 20-100 mM.

$$[Mon^{+}] = [K^{+}] + [Tris^{+}] + [Na^{+}] \tag{2.3}$$

where:

$[Mon^{+}]$ = monovalent ion concentration (M).
$[K^{+}]$ = K^{+} concentration (M).
$[Tris^{+}]$ = $Tris^{+}$ concentration (M).
$[Na^{+}]$ = Na^{+} concentration (M).

It has been observed that Mg^{2+} behavior is more complex than the one observed for monovalent ions. In fact, Mg^{2+} binds stoichiometrically to any of the dNTPs. The bound between Mg^{2+} and dNTP reduces the amount of free Mg^{2+} present for the stabilization of the PCR reaction.

Therefore, the first step for understanding the effects of Mg^{2+} is to calculate the concentration of free Mg^{2+}. The concentration of magnesium ions is usually 1.5–5 mM.

If $c_{\mathrm{dNTP}} < 0.8c_{\mathrm{Mg}}$, we assume that the amount of free Mg^{2+} equals the total amount of Mg^{2+} subtracted to the amount of dNTP:

$$[\mathrm{Mg}^{2+}] = c_{Mg} - c_{dNTP} \tag{2.4}$$

If $c_{\mathrm{dNTP}} \geq 0.8c_{\mathrm{Mg}}$, we assume that the amount of free Mg^{2+} is given from the ratio K_{a}:

$$K_a = \frac{c_{Mg} - [\mathrm{Mg}^{2+}]}{[\mathrm{Mg}^{2+}](c_{dNTP} - c_{Mg} + [\mathrm{Mg}^{2+}])},$$

$$K_a[\mathrm{Mg}^{2+}](c_{dNTP} - c_{Mg} + [\mathrm{Mg}^{2+}]) = c_{Mg} - [\mathrm{Mg}^{2+}],$$

$$K_a([\mathrm{Mg}^{2+}])^2 + (K_a c_{dNTP} - K_a c_{Mg} + 1)[\mathrm{Mg}^{2+}] - c_{Mg} = 0,$$

Resolving the quadratic formula,

$$[\mathrm{Mg}^{2+}] = \frac{-(K_a c_{dNTP} - K_a c_{Mg} + 1) \pm \sqrt{(K_a c_{dNTP} - K_a c_{Mg} + 1)^2 + 4K_a c_{Mg}}}{2K_a},$$

As the concentration must be a positive value the square root must be positive, then we obtain:

$$[\mathrm{Mg}^{2+}] = \frac{-(K_a c_{dNTP} - K_a c_{Mg} + 1) + \sqrt{(K_a c_{dNTP} - K_a c_{Mg} + 1)^2 + 4K_a c_{Mg}}}{2K_a}. \tag{2.5}$$

c_{dNTP} = total dNTP concentration (M)

c_{Mg} = total Mg^{2+} concentration (M)

$[\mathrm{Mg}^{2+}]$ = concentration of free Mg^{2+} (M)

K_a = Mg-dNTP association constant, $3 \times 10\ ^4$ for standard PCR buffers (50 mM KCl and 10 mM Tris)

In order to define which is the dominant ion effect (if the monovalent ions or Mg^{2+}) over the T_m and the secondary structure formations we need to calculate the ratio R:

$$R = \frac{\sqrt{[\mathrm{Mg}^{2+}]}}{[\mathrm{Mon}^{+}]} \tag{2.6}$$

2.A.4.2 Salt-corrected melting temperature

The equation 2.1 calculates the $\mathrm{T_m}$ at 1 M $\mathrm{Na^+}$ condition. In this section we report how to calculate $\mathrm{T_m}$ correcting the formula with salt-correction indexes.

If $R < 0.22$, the monovalent ions are considered having a dominant effect over the $\mathrm{T_m}$ and the formula for monovalent salt correction is applied:

$$\frac{1}{T_m(\mathrm{Mon^+})} = \frac{1}{T_m(1\,\mathrm{M\,Na^+})} + (4.29 f_{GC} - 3.95)10^{-5}\ln[\mathrm{Mon^+}] + 9.40 \times 10^{-6}(\ln[\mathrm{Mon^+}])^2 \tag{2.7}$$

where:

$T_m(\mathrm{Mon^+})$ = monovalent-corrected melting temperature (K).

f_{GC} = ratio of Gs/Cs in the oligo.

If $0.22 < R > 6.0$, $\mathrm{Mg^{2+}}$ is considered having a dominant effect over the $\mathrm{T_m}$ and the formula for $\mathrm{Mg^{2+}}$ correction is applied:

$$\frac{1}{T_m(\mathrm{Mg^{2+}})} = \frac{1}{T_m(1\,\mathrm{M\,Na^+})} + a + b\ln[\mathrm{Mg^{2+}}] + f_{GC}(c + d\ln[\mathrm{Mg^{2+}}]) + \\ + \frac{1}{2(N_{bp}-1)}[e + f\ln[\mathrm{Mg^{2+}}] + g(\ln[\mathrm{Mg^{2+}}])^2] \tag{2.8}$$

where:

$T_m(\mathrm{Mg^{2+}})$ = $\mathrm{Mg^{2+}}$-corrected melting temperature (K).

N_{bp} = number of bases in the oligo.

With the parameters *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h* and *i* being:

$$a = 3.92 \times 10^{-5} \tag{2.9}$$

$$b = -9.11 \times 10^{-6} \tag{2.10}$$

$$c = 6.26 \times 10^{-5} \tag{2.11}$$

$$d = 1.42 \times 10^{-5} \tag{2.12}$$

$$e = -4.82 \times 10^{-4} \tag{2.13}$$

$$f = 5.25 \times 10^{-4} \tag{2.14}$$

$$g = 8.31 \times 10^{-5} \tag{2.15}$$

If R > 6.0, the formula for Mg^{2+} correction (equation 2.8) is applied together with the parameters b, c, e and f which values are reported in the equation 2.10, 2.11, 2.13 and 2.14, respectively, and the corrected parameters a, d and g:

$$a = 3.92 \times 10^{-5}(0.843 - 0.352\sqrt{[\mathrm{Mon}^{+}]}\ln[\mathrm{Mon}^{+}]) \tag{2.16}$$

$$d = 1.42 \times 10^{-5}[1.279 - 4.03 \times 10^{-3}\ln[\mathrm{Mon}^{+}] - 8.03 \times 10^{-3}(\ln[\mathrm{Mon}^{+}])^{2}] \tag{2.17}$$

$$g = 8.31 \times 10^{-5}[0.486 - 0.258\ln[\mathrm{Mon}^{+}] + 5.25 \times 10^{-3}(\ln[\mathrm{Mon}^{+}])^{3}] \tag{2.18}$$

In Figure A2.3 we show an example of how the melting temperature was corrected for different PCR conditions. We can observe how the temperature changed drastically. The melting temperature calculated without any salt-correction was 73.1 °C (Figure A2.1A), when the monovalent ions had a dominant effect ($R < 0.22$; Figure A2.3A) the T_m was 36 °C, whereas when the R was in between 0.22 and 6, the T_m was 56.85 °C (Figure A2.3B) and with a R > 6.0 the T_m was 65.96 °C (Figure A2.3C).

2.A.4.3 Salt-corrected ΔG

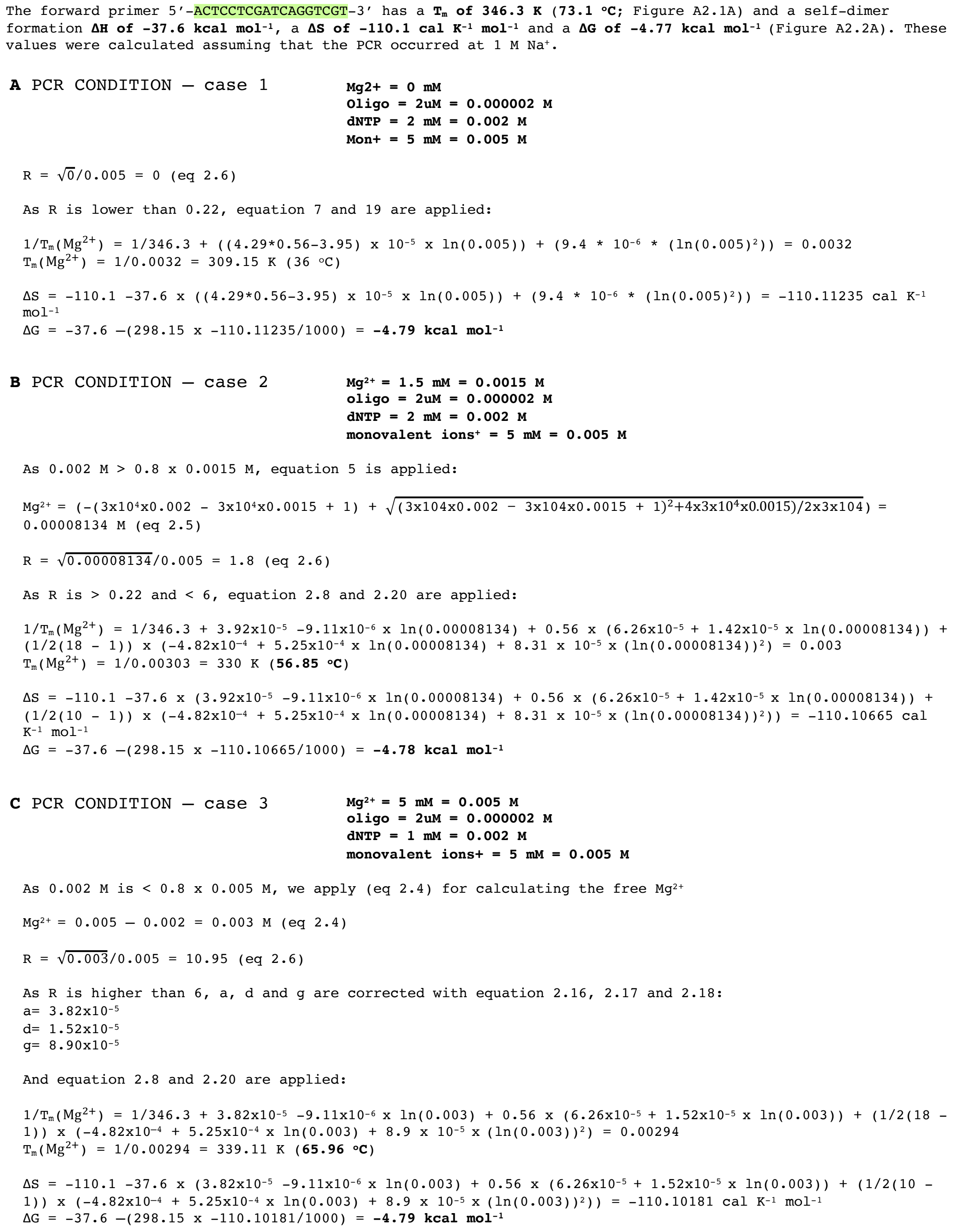
That equation 2.2 was for obtaining the ΔG in 1 M Na^+ conditions. Whereas ΔH is considered salt-independent, ΔS varies with different PCR conditions. We, therefore, need to correct ΔS in relation to the salt conditions.

$$\frac{\Delta S^{o}(\mathrm{Mon}^{+}) + Rln(\frac{C_T}{x})}{\Delta H^{o}} = \frac{\Delta S^{o}(1\,\mathrm{M\,Na}^{+}) + R\ln(\frac{C_T}{x})}{\Delta H^{o}} + (4.29f_{GC} - 3.95)10^{-5}\ln[\mathrm{Mon}^{+}]+ \\ + 9.40 \times 10^{-6}(\ln[\mathrm{Mon}^{+}])^{2},$$

$$\Delta S^{o}(\mathrm{Mon}^{+}) + R\ln(\frac{C_T}{x}) = \Delta S^{o}(1\mathrm{M\,Na}^{+}) + R\ln(\frac{C_T}{x}) + \Delta H^{o}\{(4.29f_{GC} - 3.95)10^{-5}\ln[\mathrm{Mon}^{+}]+ \\ + 9.4010^{-6}(\ln[\mathrm{Mon}^{+}])^{2}\},$$

$$\Delta S^{o}(\mathrm{Mon}^{+}) = \Delta S^{o}(1\,\mathrm{M\,Na}^{+}) + \Delta H^{o}\{(4.29f_{GC} - 3.95)10^{-5}\ln[\mathrm{Mon}^{+}]+ \\ + 9.40 \times 10^{-6}(\ln[\mathrm{Mon}^{+}])^{2}\}. \tag{2.19}$$

If $0.22 < R > 6.0$, Mg^{2+} is considered having a dominant effect over the ΔS and the formula for Mg^{2+} correction is applied. We can substitute the equation 2.8 with the the equation 2.2,



The forward primer 5′-ACTCCTCGATCAGGTCGT-3′ has a **T_m of 346.3 K (73.1 °C;** Figure A2.1A) and a self-dimer formation **ΔH of -37.6 kcal mol^{-1}**, a **ΔS of -110.1 cal K^{-1} mol^{-1}** and a **ΔG of -4.77 kcal mol^{-1}** (Figure A2.2A). These values were calculated assuming that the PCR occurred at 1 M Na^+.

A PCR CONDITION – case 1

Mg2+ = 0 mM
Oligo = 2uM = 0.000002 M
dNTP = 2 mM = 0.002 M
Mon+ = 5 mM = 0.005 M

R = $\sqrt{0}$/0.005 = 0 (eq 2.6)

As R is lower than 0.22, equation 7 and 19 are applied:

$1/T_m(Mg^{2+})$ = 1/346.3 + ((4.29*0.56–3.95) x 10^{-5} x ln(0.005)) + (9.4 * 10^{-6} * (ln$(0.005)^2$)) = 0.0032
$T_m(Mg^{2+})$ = 1/0.0032 = 309.15 K (36 °C)

ΔS = –110.1 –37.6 x ((4.29*0.56–3.95) x 10^{-5} x ln(0.005)) + (9.4 * 10^{-6} * (ln$(0.005)^2$)) = –110.11235 cal K^{-1} mol^{-1}
ΔG = –37.6 –(298.15 x –110.11235/1000) = **–4.79 kcal mol^{-1}**

B PCR CONDITION – case 2

Mg^{2+} = 1.5 mM = 0.0015 M
oligo = 2uM = 0.000002 M
dNTP = 2 mM = 0.002 M
monovalent ions$^+$ = 5 mM = 0.005 M

As 0.002 M > 0.8 x 0.0015 M, equation 5 is applied:

Mg^{2+} = (–(3x10^4x0.002 – 3x10^4x0.0015 + 1) + $\sqrt{(3x104x0.002 - 3x104x0.0015 + 1)^2+4x3x10^4x0.0015)/2x3x104}$) = 0.00008134 M (eq 2.5)

R = $\sqrt{0.00008134}$/0.005 = 1.8 (eq 2.6)

As R is > 0.22 and < 6, equation 2.8 and 2.20 are applied:

$1/T_m(Mg^{2+})$ = 1/346.3 + 3.92x10^{-5} –9.11x10^{-6} x ln(0.00008134) + 0.56 x (6.26x10^{-5} + 1.42x10^{-5} x ln(0.00008134)) + (1/2(18 – 1)) x (–4.82x10^{-4} + 5.25x10^{-4} x ln(0.00008134) + 8.31 x 10^{-5} x (ln$(0.00008134))^2$) = 0.003
$T_m(Mg^{2+})$ = 1/0.00303 = 330 K (**56.85 °C**)

ΔS = –110.1 –37.6 x (3.92x10^{-5} –9.11x10^{-6} x ln(0.00008134) + 0.56 x (6.26x10^{-5} + 1.42x10^{-5} x ln(0.00008134)) + (1/2(10 – 1)) x (–4.82x10^{-4} + 5.25x10^{-4} x ln(0.00008134) + 8.31 x 10^{-5} x (ln$(0.00008134))^2$)) = –110.10665 cal K^{-1} mol^{-1}
ΔG = –37.6 –(298.15 x –110.10665/1000) = **–4.78 kcal mol^{-1}**

C PCR CONDITION – case 3

Mg^{2+} = 5 mM = 0.005 M
oligo = 2uM = 0.000002 M
dNTP = 1 mM = 0.002 M
monovalent ions+ = 5 mM = 0.005 M

As 0.002 M is < 0.8 x 0.005 M, we apply (eq 2.4) for calculating the free Mg^{2+}

Mg^{2+} = 0.005 – 0.002 = 0.003 M (eq 2.4)

R = $\sqrt{0.003}$/0.005 = 10.95 (eq 2.6)

As R is higher than 6, a, d and g are corrected with equation 2.16, 2.17 and 2.18:
a= 3.82x10^{-5}
d= 1.52x10^{-5}
g= 8.90x10^{-5}

And equation 2.8 and 2.20 are applied:

$1/T_m(Mg^{2+})$ = 1/346.3 + 3.82x10^{-5} –9.11x10^{-6} x ln(0.003) + 0.56 x (6.26x10^{-5} + 1.52x10^{-5} x ln(0.003)) + (1/2(18 – 1)) x (–4.82x10^{-4} + 5.25x10^{-4} x ln(0.003) + 8.9 x 10^{-5} x (ln$(0.003))^2$) = 0.00294
$T_m(Mg^{2+})$ = 1/0.00294 = 339.11 K (**65.96 °C**)

ΔS = –110.1 –37.6 x (3.82x10^{-5} –9.11x10^{-6} x ln(0.003) + 0.56 x (6.26x10^{-5} + 1.52x10^{-5} x ln(0.003)) + (1/2(10 – 1)) x (–4.82x10^{-4} + 5.25x10^{-4} x ln(0.003) + 8.9 x 10^{-5} x (ln$(0.003))^2$)) = –110.10181 cal K^{-1} mol^{-1}
ΔG = –37.6 –(298.15 x –110.10181/1000) = **–4.79 kcal mol^{-1}**

Figure A2.3: A schematic view of how the T_m and ΔG values are calculated at different salt-conditions (A, B and C). The ΔG was calculated considering a temperature of 25 °C (298.15 K).

which describes how to calculate ΔG,

$$\frac{\Delta S^o(\mathrm{Mg}^{2+}) + R\ln(\frac{C_T}{x})}{\Delta H^o} = \frac{\Delta S^o(1\,\mathrm{M\,Na}^{+}) + R\ln(\frac{C_T}{x})}{\Delta H^o} + a + b\ln[\mathrm{Mg}^{2+}]+ \\ + f_{GC}(c + d\ln[\mathrm{Mg}^{2+}]) + \frac{1}{2(N_{bp}-1)}[e + f\ln[\mathrm{Mg}^{2+}] + g(\ln[\mathrm{Mg}^{2+}])^2],$$

$$\Delta S^o(\mathrm{Mg}^{2+}) + R\ln(\frac{C_T}{x}) = \Delta S^o(1\,\mathrm{M\,Na}^{+}) + R\ln(\frac{C_T}{x}) + \Delta H^o\{a + b\ln[\mathrm{Mg}^{2+}]+ \\ + f_{GC}(c + d\ln[\mathrm{Mg}^{2+}]) + \frac{1}{2(N_{bp}-1)}[e + f\ln[\mathrm{Mg}^{2+}] + g(\ln[\mathrm{Mg}^{2+}])^2]\},$$

$$\Delta S^o(\mathrm{Mg}^{2+}) = \Delta S^o(1\,\mathrm{M\,Na}^{+}) + \Delta H^o\{a + b\ln[\mathrm{Mg}^{2+}] + f_{GC}(c + d\ln[\mathrm{Mg}^{2+}])+ \\ + \frac{1}{2(N_{bp}-1)}[e + f\ln[\mathrm{Mg}^{2+}] + g(\ln[\mathrm{Mg}^{2+}])^2]\}. \quad (2.20)$$

If $R > 6.0$, the formula for Mg^{2+} correction (equation 2.20) is applied together with the parameters b, c, e and f reported in the equations 2.10, 2.11, 2.12 and 2.13, respectively, and the corrected parameters a, d and g from the equations 2.16, 2.17 and 2.18, respectively.

Once the salt-corrected ΔS is obtained, ΔG can be calculated with the equation 2.2. In the Figure A2.3 we can observe how the ΔG value is much less susceptible to change due to the PCR conditions than the $\mathrm{T_m}$ value. The ΔG calculated for the self-dimer secondary structure in Figure A2.2a was -4.77 kcal mol^{-1}, whereas with the different salt-correction formulas applied in Figure A2.3, that values ranged between -4.78 and -4.79 kcal mol^{-1}.

2.B PhyloPrimer links

Table A2.4: Links to the result pages for the six different sets of primer pairs designed with PhyloPrimer.

Primer pair	PhyloPrimer link
PP1	www.cerealsdb.uk.net/cerealgenomics/cgi-bin/phyloprimerResultsPrimer.cgi?defSet=mAcqoomtcDiwLYnJiqimaT
PP2	www.cerealsdb.uk.net/cerealgenomics/cgi-bin/phyloprimerResultsPrimer.cgi?defSet=BOgdmkNHFgNnqvizRNmIaT
PP3	www.cerealsdb.uk.net/cerealgenomics/cgi-bin/phyloprimerResultsPrimer.cgi?defSet=qSqbUdjiNoLlDWZkubeO_6aT
PP4	www.cerealsdb.uk.net/cerealgenomics/cgi-bin/phyloprimerResultsPrimer.cgi?defSet=qSqbUdjiNoLlDWZkubeO_3aT
PP5	www.cerealsdb.uk.net/cerealgenomics/cgi-bin/phyloprimerResultsPrimer.cgi?defSet=qSqbUdjiNoLlDWZkubeO_4aT
PP6	www.cerealsdb.uk.net/cerealgenomics/cgi-bin/phyloprimerResultsPrimer.cgi?defSet=jbndSGAEKDYTGkbhoSOFaT

Chapter 3

Glacier clear ice bands indicate englacial channel microbial distribution

Abstract

Distant glacial areas are interconnected by a complex system of fractures and water channels which run in the glacier interior and characterize the englacial realm. Even though the englacial environment is the biggest portion of a glacier, it is still widely unexplored due to technical sampling difficulties and to the assumption that its conditions are too challenging to life. However, several studies have now shown how this realm can sustain life in ice veins and englacial channels.

Englacial water can slowly freeze in the englacial channels where the slow freezing excludes air bubbles giving the ice a clear aspect. This ice is then uplifted to the surface ablation zone of the glacier by glacial ice movements and can therefore be observed on the glacier surface in the form of clear ice bands. We employed an indirect method to sample englacial water by coring these ice bands.

We were able for the first time to compare microbial communities sampled from clear (i.e. frozen englacial water bands) and cloudy ice (i.e. meteoric ice) through Illumina sequencing of the 16S rRNA gene from 62 ice samples. Although microbial communities were primarily shaped and structured by their spatial distribution on the glacier, ice type was a clear secondary factor explaining the microbial differences in the samples. One area of the glacier, in particular, presented significant microbial community differences between clear and cloudy ice. Whereas the clear ice communities presented typical cold-adapted glacial communities, the cloudy ice presented a less defined glacial community and more ubiquitous environmental organisms. These results suggest an important role of the englacial channels in the development of a glacial microbial community and in the microbial dispersion within the glacier.

3.1 Introduction

Widespread understanding of glaciers as biomes has only been achieved in the past few decades (Anesio and Laybourn-Parry, 2012). Glacial biomes are microbially dominated and are usually divided in three different environments: the glacial surface (supraglacial), within its interior (englacial), and at the base where the glacier is in contact with the bedrock (subglacial) (Anesio et al., 2017; Garcia-Lopez et al., 2019). Supraglacial studies have mainly focusing on cryoconite holes (Bagshaw et al., 2007; Cook et al., 2016; Edwards et al., 2011; Fountain et al., 2004; Musilova et al., 2015; Tranter et al., 2004; Uetake et al., 2019) and ice algal-associated communities (Yallop et al., 2012; Musilova et al., 2015; Lutz et al., 2017; Uetake et al., 2010), both of which are dominated by autotrophs. Here microbial communities are highly influenced by the surrounding environment from which wind and precipitation phenomena transport dust and particles and, consequently, nutrients and microorganisms to the glacier surface (Grzesiak et al., 2015). The subglacial environment, dark and oxygen depleted, is dominated by chemolithotrophs which are able to use the chemical compounds and H^+ released from the bedrock/ice grinding to produce energy (Boyd et al., 2014; Dieser et al., 2014; Kayani et al., 2018; Stibal et al., 2012b; Telling et al., 2015). Heterotrophic communities are also found in both supraglacial and subglacial environments, mainly utilizing organic carbon produced by other organisms (Anesio and Laybourn-Parry, 2012; Garcia-Lopez et al., 2019).

Potentially the largest glacial habitat is the englacial environment, that part of the glacier or ice sheet between the bottom and the surface. The englacial region of temperate ice contains pockets of water at all scales from microscopic veins formed at the junction of ice crystal boundaries to macroscopic water-filled crevasses (Bamber, 1988; Nye and Frank, 1973; Watts and England, 1976). Although some of these pockets may be isolated, others are interconnected pathways exchanging water between the surface and the bed (Catania et al., 2008; Fountain et al., 2005; Fountain and Walder, 1998). The englacial environment has not been widely studied due to the technical challenges associated with sampling this habitat, yet microbial metabolism has been observed within aqueous veins (Dani et al., 2012; Mader, 1992; Miteva, 2008; Price, 2000). Further, several studies of englacial ice cores have revealed microbial changes with the depth and age of ice and have successfully isolated microorganisms (An et al., 2010; Knowlton et al., 2013; Liu et al., 2019; Miteva, 2008; Singh et al., 2016), but the life challenging conditions (e.g. subzero temperatures and nutrient depletion) have cast doubts whether isolated englacial organisms can thrive. However, the water flowing within the englacial region has rarely been studied directly despite suggestions that englacial water may be the most metabolically active habitat within the englacial realm (Hotaling et al., 2017; Martinez-Alonso et al., 2019). Better definition of the microbial communities inhabiting englacial waters would help link biogeochemical processes connecting supraglacial and subglacial biomes and further refine the role of glaciers in carbon and nutrient cycling and how they could be influenced by glacier shrinking (Anesio et al., 2009; Hawkings et al., 2015; Kujawinski, 2017; Milner et al., 2017). More generally, the ecology of glacial environments provides an upstream boundary condition for downstream aquatic communities in streams, lakes, and tidal environments (Garcia-Lopez et al., 2019; Hood et al., 2015; O'Neel et al., 2015). As a first step in this effort, our study characterized and compared the diversity and structure of microbial communities between the water flowing through englacial fractures and the meteoric glacier ice formed in proximity of the fractures. I hypothesized that the microbial community within the englacial hydrological system would differ from that in the surrounding ice.

3.2 Materials and methods

The study site was Storglaciären, a small polythermal valley glacier in Arctic Sweden (67^o 54' 10" N 18^o 34' 00" E; Figure 3.1A). This relatively easily accessible glacier has been intensively studied hydrologically (Holmlund and Eriksson, 1989; Hooke and Pohjola, 1994; Jansson, 1996). To sample the englacial water, we employed an indirect method. Near-surface ice (i.e. maximum sampled depth of 131 cm) was sampled in the ablation zone (ice-exposed region on the lower third of the glacier). Fountain et al. (2005) showed that clear bands of ice, visible on the surface of Storglaciären, are the product of the slow freezing of englacial water within fractures deep in the glacier. The refrozen fractures are uplifted and exposed at the surface due to normal processes of glacier movement and ablation (melting) of the ice surface (Cuffey and Paterson, 2010; Pohjola, 1996). The slow freezing of water, particularly when flowing, favors the exclusion of air bubbles and formation of clear ice (Carte, 1961; Hubbard et al., 2000). Between the bands of refrozen englacial water, meteoric glacier ice formed from the compaction of snow, forming a dense matrix of air bubbles (i.e. cloudy ice).

3.2.1 Sample collection and ice classification

Three ice cores 45-131 cm deep were obtained from each location using a Kovacs 10 cm corer. Subsurface samples were collected to avoid potential contamination by surface weathering and melt water. Both clear ice bands (considered as frozen englacial fractures) and cloudy ice (considered here as meteoric glacier ice) were drilled in triplicate and processed in the same manner (Figure 3.1B-C). Each 10 cm diameter core was cut into two or three sections: the first 15-20 cm of ice from the surface were classified as surface samples (sux) and the rest of the core was classified as subsurface sample (sub). Occasionally, a core would include both a clear section

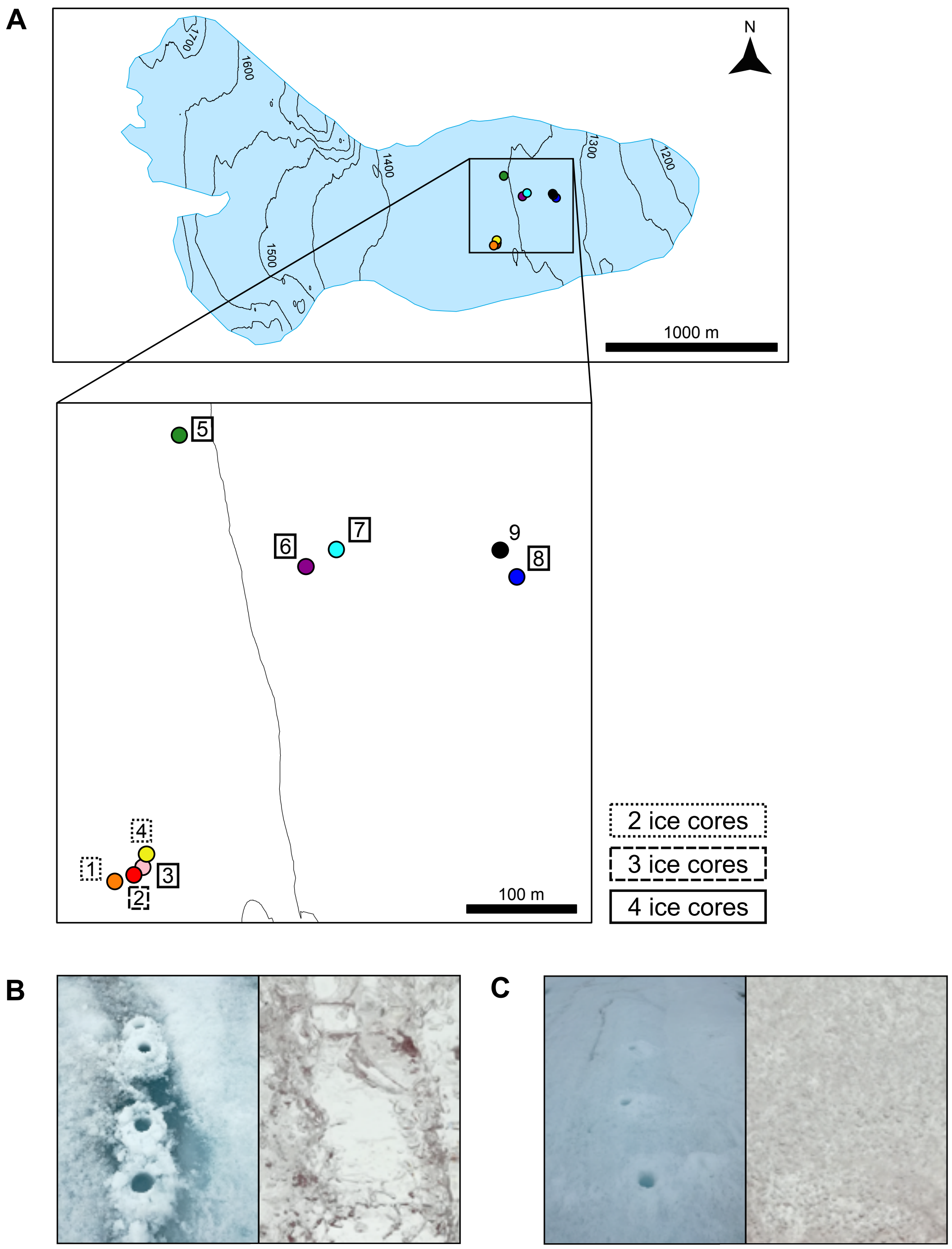


Figure 3.1: Map of the sampling location sited in the ablation zone of the Storglaciären and sampling site and ice type images. (A) Position of the 9 sampled sites on the glacier. Four ice cores were processed from sites 3, 5, 6, 7 and 8, three ice cores were processed from site 2, and two ice cores were processed from sites 1 and 4. Two surface algal samples were collected in site 9. (B) Clear ice band with an example of the clear ice matrix and (C) cloudy ice sampling site with an example of the cloudy ice matrix.

Table 3.1: Site, replicate, ice layer and type associated with each sample. Samples were classified by ice layer: surface (sux) and subsurface samples (sub); and by ice type: surface clear ice (SE), surface cloudy ice (SU), subsurface clear ice (E), subsurface cloudy ice (U) and subsurface mixed ice (when the ice showed a mixed ice matrix; M).

Site	Replicate	Depth (cm)	Layer	Ice type	Sample ID
1	A	0-30	sux	SE	1A-0-30
	A	30-55	sub	E	1A-30-55
	A	55-118	sub	U	1A-55-118
	B	0-25	sux	SE	1B-0-25
	B	25-113	sub	M	1B-25-113
2	A	0-30	sux	SE	2A-0-30
	A	30-83	sub	M	2A-30-83
	A	83-111	sub	E	2A-83-111
	B	0-20	sux	SE	2B-0-20
	B	20-125	sub	M	2B-20-125
	C	0-25	sux	SE	2C-0-25
	C	25-61	sub	E	2C-25-61
	C	61-119	sub	U	2C-61-119
3	A	0-33	sux	SE	3A-0-33
	A	33-102	sub	M	3A-33-102
	B	0-32	sux	SE	3B-0-32
	B	32-125	sub	M	3B-32-125
	C	0-32	sux	SE	3C-0-32
	C	32-111	sub	M	3C-32-111
	D	0-25	sux	SU	3D-0-25
	D	25-75	sub	U	3D-25-75
	D	75-122	sub	E	3D-75-122
4	A	0-15	sux	SU	4A-0-15
	A	15-75	sub	U	4A-15-75
	B	0-17	sux	SU	4B-0-17
	B	17-74	sub	U	4B-17-74
5	A	20-45	sub	U	5A-20-45
	B	0-20	sux	SE	5B-0-20
	C	0-25	sux	SU	5C-0-25
	C	25-80	sub	U	5C-25-80
	D	0-25	sux	SU	5D-0-25
	D	25-70	sub	U	5D-25-70
6	A	0-30	sux	SE	6A-0-30
	A	30-105	sub	E	6A-30-105
	B	0-25	sux	SE	6B-0-25
	B	25-75	sub	M	6B-25-75
	C	0-30	sux	SE	6C-0-30
	C	30-100	sub	U	6C-30-100
	D	0-20	sux	SE	6D-0-20
	D	20-63	sub	E	6D-20-63
7	A	0-30	sux	SE	7A-0-30
	A	30-55	sub	E	7A-30-55
	A	55-90	sub	U	7A-55-90
	B	0-25	sux	SE	7B-0-25
	B	25-71	sub	E	7B-25-71
	B	71-125	sub	U	7B-71-125
	C	0-30	sux	SE	7C-0-30
	C	30-65	sub	E	7C-30-65
	C	65-111	sub	U	7C-65-111
	D	0-16	sux	SU	7D-0-16
	D	16-79	sub	U	7D-16-79
8	A	0-20	sux	SE	8A-0-20
	A	20-87	sub	E	8A-20-87
	A	87-131	sub	U	8A-87-131
	B	0-20	sux	SE	8B-0-20
	B	20-77	sub	M	8B-20-77
	C	0-25	sux	SE	8C-0-25
	C	25-65	sub	U	8C-25-65
	D	0-20	sux	SE	8D-0-20
	D	20-100	sub	E	8D-20-100
9	A	0-2	algae	A	9A
	B	0-2	algae	A	9B

and cloudy section in its subsurface layer, perhaps due to the inclination of the refrozen fracture. In this circumstance the core was separated at the matrix interface. Each core section was conserved in a different sterile Whirl-Pak bag (Whirl-Pak, Nasco, USA) and classified as surface clear ice (SE), surface cloudy ice (SU), subsurface clear ice (E), subsurface cloudy ice (U) or subsurface mixed ice (when the ice showed a mixed ice matrix; M).

After moving to a new sampling site or switching to a different ice type, the core barrel was cleansed by coring into the ice of the new sampling location and discarding the core. Nine different sites were sampled in total. Sites 1, 2, 3 and 4 were within a 60 meter proximity, whereas site 5, the closest to those samples was 350 meters apart. Sites 6 and 7 were 100 meters distant from site 5 and 250 meters away from sites 8 and 9. Whereas at least three cores were taken per site, we processed two cores in sites 1 and 4, three cores at site 2, and four cores were taken from sites 3, 5, 6, 7 and 8 (Figure 3.1 and Table 3.1). At site 9, two surface samples were taken from the upper 2 cm where the ice was darker and visibly enriched with algae. These two samples (i.e. algae samples) were collected as microbial control and compared with the others.

The cores were temporarily stored on the glacier surface during the day, covered by snow, before being carried back to the station for processing. The cores were immediately melted at ambient temperature in the laboratory of the Tarfala Research Station. To avoid external contamination, we removed the outer layer of the ice core by discarding the first few milliliters of melted water (Christner et al., 2005). For major ion (i.e. nutrient) analysis 1.5 mL of the glacier melt water was filtered through a 0.22 µm cellulose nitrate inline syringe filter (25 mm diameter, Whatman^{TM}) and stored in a polypropylene autosampler vial at 3 °C; for cell counts 15 mL of water was stored with 2% of glutaraldehyde at 4 °C, and for DNA analyses the remaining water (1-3 liters) was filtered through sterile polycarbonate membrane filters (0.22 micron pores, 47 mm, Sigma-Aldrich) and stored at -20 °C. Additionally, 50 milliliters of Milli-Q water were filtered thought a filter with exactly the same procedure and stored for further analyses in order to assess any eventual lab contamination.

3.2.2 Geochemical analyses

Nutrients (Cl^-, $SO_4{}^{2-}$, $NO_3{}^-$, $PO_4{}^{3-}$, Mg^{2+}, Ca^{2+}, $NH_4{}^+$, $Na_{rock}{}^+$ and K^+) and DOC concentrations were quantified by Dr. Alexandra Holland at the University of Bristol glaciology laboratories. More details are reported in Appendix 3.A.

3.2.3 Cell enumeration and biovolume

Cell concentrations were determined for the prokaryotic and eukaryotic components after a thorough vortexing of the samples. All the samples were observed under a LEICA DM2000 LED microscope and imaged with Leica MC 120 HD camera connected to LAS v 4.12 software. Eukaryotic cells were counted under the visible light where each sample was first loaded on a Fuchs-Rosenthal haemocytometer and then two counting chambers were screened at a 400x magnification. The eukaryotic organisms were classified into four different types: *Ancylonema sp.*, *Mesotaenium sp.*, circular cells and oblong cylindrical cells (Figure 3.2A-D) and counted for each of the samples. For each cell type, thirty images were taken using a magnification of 400x for *Ancylonema sp.* and 100x magnification for the other cells. When filamentous cell chains (e.g. *Ancylonema* or Cyanobacteria) were observed, the single cells composing the chain were considered for cell counts and diameter/height measurements. Using Fiji software (Schindelin et al., 2012), the diameter and height of all the cells, where applicable, in order to calculate the average cell volume for these four ecotypes using formulae after Hillebrand (µm^3 cell^{-1}) (Hillebrand et al., 1999). In order to obtain the biovolume, the cell volumes were then multiplied by the cell counts (µm^3 mL^{-1}).

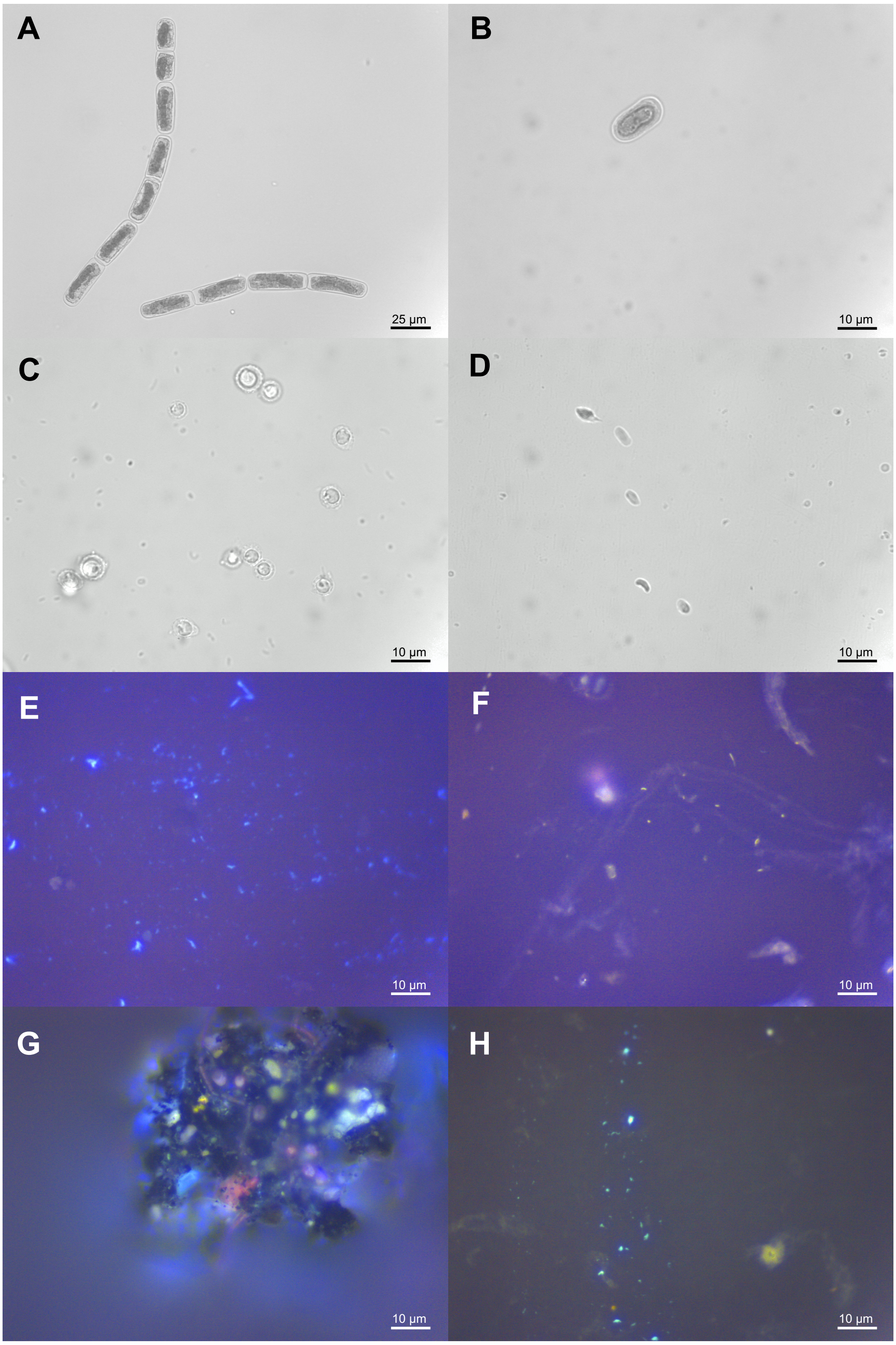


Figure 3.2: Microscopy pictures taken with LEICA DM2000 LED microscope and imaged with Leica MC 120 HD camera. The first four pictures (A-D) represent eukaryotic organisms where *Ancylonema sp.* (A), *Mesotanium sp.* (B), circular cells (C) and oblong cylindrical cells (D). These pictures were taken under visible light with 40x objective whereas a 100x objective was used for all the others. The last four pictures (E-H) were taken under fluorescent light and represent DAPI-stained bacteria (E), autofluorescence Cyanobacteria (F), particle-attached cells (G) and exopolysaccharide-attached cells (H).

Prokaryotic cells were counted under a fluorescent light following the protocol presented in Grzesiak et al. (2015) where 5 mL of melted glacier ice (0.5 mL for the algal samples of site 9) was incubated with 4',6-diamidino-2-phenylindole (DAPI; final concentration of 1%) in darkness for 10 minutes and then filtered through 0.2 µm pore size black polycarbonate filters (Millipore Isopore) for epifluorescence microscopy. For each sample, under 1000x magnification using an oil immersion objective, the prokaryotic cells emitting blue and yellow-orange fluorescent light over thirty camera field of views were counted (Figure 3.2E-F). The yellow-orange fluorescence was assumed to be the autofluorescence emitted by Cyanobacteria (Rassoulzadegan and Sheldon, 1986; Uetake et al., 2010). Although visible with this technique, no algal organisms were included in this count. Bacterial biovolumes were also approximated using formulae after Hillebrand (Hillebrand et al., 1999).

In our work, prokaryotic counts are presented as cell counts (cell mL^{-1}) whereas total counts of the prokaryotic and eukaryotic component were presented as biovolumes (µm^3 mL^{-1}).

3.2.4 DNA extraction and Illumina sequencing

The filters (0.22 micron pores, 47 mm, Sigma-Aldrich) were directly processed with the DNeasy PowerWater Kit (QIAGEN, Hilden, Germany) following the manufacturer protocol. DNA concentrations were measured with the Qubit® 1.0 Fluorometer and Qubit® dsDNA HS Assay Kits (Invitrogen, Carlsbad, CA, USA). Between 1 and 250 ng of DNA were obtained per sample. All the samples were then amplified with primers specific to the V3-V4 region (450-500 bp) of the 16S rRNA gene. The primers Pro341F and Pro805R target both the bacterial and archaeal organisms (Table A3.1) (Takahashi et al., 2014).

To account for low starting biomass and add on sequencing adapters, the first 25 PCR cycles were performed using the Pro341F and Pro805R primers and then a further 25 cycles were run with the same primers combined with the Illumina Nextera Transposase adapters (Table A3.1). PCR was run adding 12.5 µL of KAPA HiFi HotStart ReadyMix (Roche Applied Science), 1.5 µL of each 5 M primer (0.3 µM final concentration), between 5.50 to 10.5 µL of sample (5-30 ng of DNA) and nuclease free water up to a final volume of 25 µL PCR solution. PCR conditions were 3 minutes at 95 °C, 25 cycles of 20 seconds at 98 °C, 15 seconds at 65 °C and 15 seconds at 72 °C, and a final extension step of 5 minutes at 72 °C for the first step of the nested PCR. The second step consisted in 3 minutes at 95 °C, 25 cycles of 30 seconds at 98 °C, 30 seconds at 55 °C, 30 seconds at 72 °C and a final extension step of 5 minutes at 72 °C. All the PCR runs were checked on 1.5% w/v horizontal agarose gel (0.5 mg ethidium bromide ml^{-1}) in 1x TAE buffer (Tris acetate-EDTA) at 120 mV for 30 minutes (Bio-Rad PowerPac 300, Bio-Rad Laboratories). Negative controls did not show any band except than in one of the runs. That negative control sample was therefore sequenced. The amplicons were then indexed with the Nextera XT Index Kit, pooled together and sequenced in two lanes of the Illumina MiSeq using 600 cycle MiSeq reagent kit (version 2) obtaining paired 300 bp reads. Basecalling was done with Illumina Real Time Analysis (RTA) software version 1.18.54.0. The sequencing was performed by the University of Bristol Genomics Facility. The sequence data have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB40002.

3.2.5 Bioinformatics and statistical analyses

All the 62 glacial samples and the 2 negative controls were processed together following the same pipeline. All the analyses on the DNA reads were performed in the R environment v 3.6.1 (R Core Team 2019, 2019) except for the first step where primers and adapters were trimmed with the software CUTADAPT v 2.6 (Martin, 2011). The quality check and filtering of the amplicon sequences were performed using the R package DADA2 v 1.14.0 (Callahan et al., 2016) following these steps: read quality trimming, read dereplication, ASVs (amplicon sequence variants) inference, read merging, chimera detection and taxonomy assignment with the Silva database v 132 (Yilmaz et al., 2014). Using the R package decontam v 1.6.0 (Davis et al., 2018) we also removed the contaminant reads from all the samples. Contaminants were identified by the two negative controls (NC1 and NC2). NC1 was the DNA extracted from the filtered sterile Milli-Q water and treated with the exact same protocol as the other samples and NC2 was a PCR negative protocol and represented a faint line on one of the electrophoresis gels that was run during the amplicon preparation.

Two samples were overloaded during the Illumina sequencing giving an output of 690,952 and 1,278,606 sequences in 3D-75-122 and 9B, respectively (Table A3.2). The two overloaded samples were trimmed down to 263,233 sequences which corresponded to the number of reads in third sample ranked by read-count (8D-20-100).

I calculated sample rarefaction curves in order to check how the diversity was covered in all the samples with the R package iNEXT v 2.0.20 (Hsieh et al., 2016). Then we removed the singleton component from the dataset. Singletons were here defined as the ASVs represented by only one sequence in the entire dataset (Auer et al., 2017; Callahan et al., 2019) (Table A3.3). The ASV table was then transformed with the package DESeq2 v 1.26.0 (Love et al., 2014) and the cluster analysis was calculated on this dataset using Euclidean distances. In the heatmap, the samples were disposed following this sample clustering and only genera that represented more than 2% of the community in at least one sample of the dataset are reported. I also reported the 16S rRNA associated with Chloroplast (order-level) and WPS-2 (phylum-level) in order to give a better idea about the Unclassified component at genus-level. I investigated how each genus varied between different sites, ice types and ice layers with the Kruskal-Wallis test. This test was performed for each genus on the relative abundance dataset (no algal samples from site 9 were included). We considered the Kruskal-Wallis test to be significant when the p-value was lower than 0.05.

All the other statistical analyses such as permutational multivariate analysis of variance (PERMANOVA) and distance-based redundancy analysis (dbRDA) were performed on the dataset transformed with the Hellinger transformation (Legendre and Gallagher, 2001). PERMANOVA analyses were performed on Bray-Curtis dissimilarity matrices for ASV and microbial count data and Euclidean distance matrices for the geochemical dataset (9999 permutations). All the factorial analyses (e.g. PERMANOVA) were performed with three different factors: 'site' (9 levels as 1, 2, 3, 4, 5, 6, 7, 8 and 9), 'ice type' (5 levels as SE, SU, E, U and M) and 'layer' (2 levels as sux and sub). The factor 'layer' was divided only between surface and subsurface samples because the subsurface samples represented a high range of depths and this did not allow a more detailed division. The ice types E and U, which are the clear and cloudy subsurface ice, also represented samples from different ice depths. PERMANOVA performed on uniquely clear or cloudy dataset was performed on only the factor 'site' to check which of the two ice types showed a higher degree of differentiation between sites.

The algae samples (site 9) were excluded by the PERMANOVA and dbRDA analyses in order to not inflate the observed difference between samples. In the dbRDA analysis, sample 4B-17-74 represented an outlier (dbRDA1 = -2.83 and dbRDA2 = -4.00) was removed from the graph for visualization purposes.

PERMANOVA pairwise comparisons were performed with the R package pairwiseAdonis v 0.4 (Martinez Arbizu, 2020) and p-values were adjusted with the Bonferroni correction. I used the following R packages for data manipulation and graph plotting: ggplot2 v 3.2.1 (Wickham, 2016), gplots v 3.0.1.1 (Warnes et al., 2019), tidyr v 1.0.2 (Wickham et al., 2017), phyloseq v 1.30.0 (McMurdie and Holmes, 2013), vegan v 2.5.6 (Oksanen, 2017), viridis v 0.5.1 (Garnier, 2018), gridExtra v 2.3 (Auguie, 2017) and plyr v 1.8.5 (Wickham, 2011).

3.3 Results

3.3.1 Nutrient and DOC composition

Twenty-five percent of the geochemical variance was explained by the differences between each site location. Differences between clear and cloud ice was a secondary factor explaining 17% of the variance (p-value < 0.05; Table 3.2A).

PERMANOVA analysis performed on each variable showed that magnesium ($R^2 = 0.81$), potassium ($R^2 = 0.72$), calcium ($R^2 = 0.70$), sodium ($R^2 = 0.35$) and ammonium ($R^2 = 0.24$) showed significant values ($p\text{-value} < 0.05$) for the site locations. No significant results were obtained for the differences between ice types. In sites 1, 2, 3, 4 and 5 magnesium, calcium and potassium showed higher concentrations (17-19 ppb, 17-33 ppb and 26-123 ppb respectively) than in the other sites (6-13 ppb, 6-14 ppb and 0-18 ppb). Rock sodium was also lower in sites 6, 7 and 8 compared to the other sites with values of 5-25 ppb (8D-0-20 excluded) and 0-14 ppb. Ammonium values were higher in the sites 6, 7, 8 and 9 and especially in site 8 where the average value was 13 ± 6 ppb against 6 ± 5 ppb in all the other sites. The phosphate concentration was below the LOD in all of the samples. DOC values were much higher in the algae samples (site 9) with values of more than 1500 ppb while in the others all the values were below 500 ppb (Figure 3.3). Nutrient and DOC concentrations grouped by different ice types are reported in Figure A3.1.

3.3.2 Cell enumeration

The highest bacterial number was in site 9, where high content algal samples were collected, with 6x10^4 and 1x10^5 cell mL^{-1} in sample 9A and 9B, respectively. In all the other ice samples the bacterial concentration ranged between 2x10^3 and 3x10^4 cell mL^{-1}. The cell count was higher in the surface clear ice (SE) samples compared to the other ice types (Figure 3.4A) whereas fewer differences were observed among different sites (Figure 3.4B). Looking at the biovolume data, where also the algal component was taken in account, the algal samples presented, again, the

Table 3.2: Permutational multivariate analysis of variance (PERMANOVA) test performed on the ASV, geochemical, prokaryotic count and biovolume datasets for the model site x ice type (A) and site x layer (B). PERMANOVA was performed with 9999 permutations on Bray-Curtis dissimilarity matrices for all the datasets except from the geochemical dataset where a Euclidean distance matrix was used. The symbol '-' is reported for non-significant R^2 values where the statistic $p\text{-value} \geq 0.05$.

Datasets	**A Site x ice type**			**B Site x layer**		
	Factors	**R^2**	**p-value**	**Factors**	**R^2**	**p-value**
Nutrients and DOC	site	0.25	0.01	site	0.25	0.00
	ice type	0.17	0.00	layer	0.11	0.00
	site x ice type	-	-	site x layer	-	-
Prokaryotic count	site	-	-	site	0.20	0.04
	ice type	0.17	0.03	layer	0.08	0.01
	site x ice type	-	-	site x layer	-	-
Biovolume	site	-	-	site	-	-
	ice type	0.13	0.04	layer	0.08	0.02
	site x ice type	-	-	site x layer	-	-
ASV	site	0.21	0.00	site	0.21	0.00
	ice type	0.10	0.00	layer	0.05	0.00
	site x ice type	0.29	0.05	site x layer	-	-

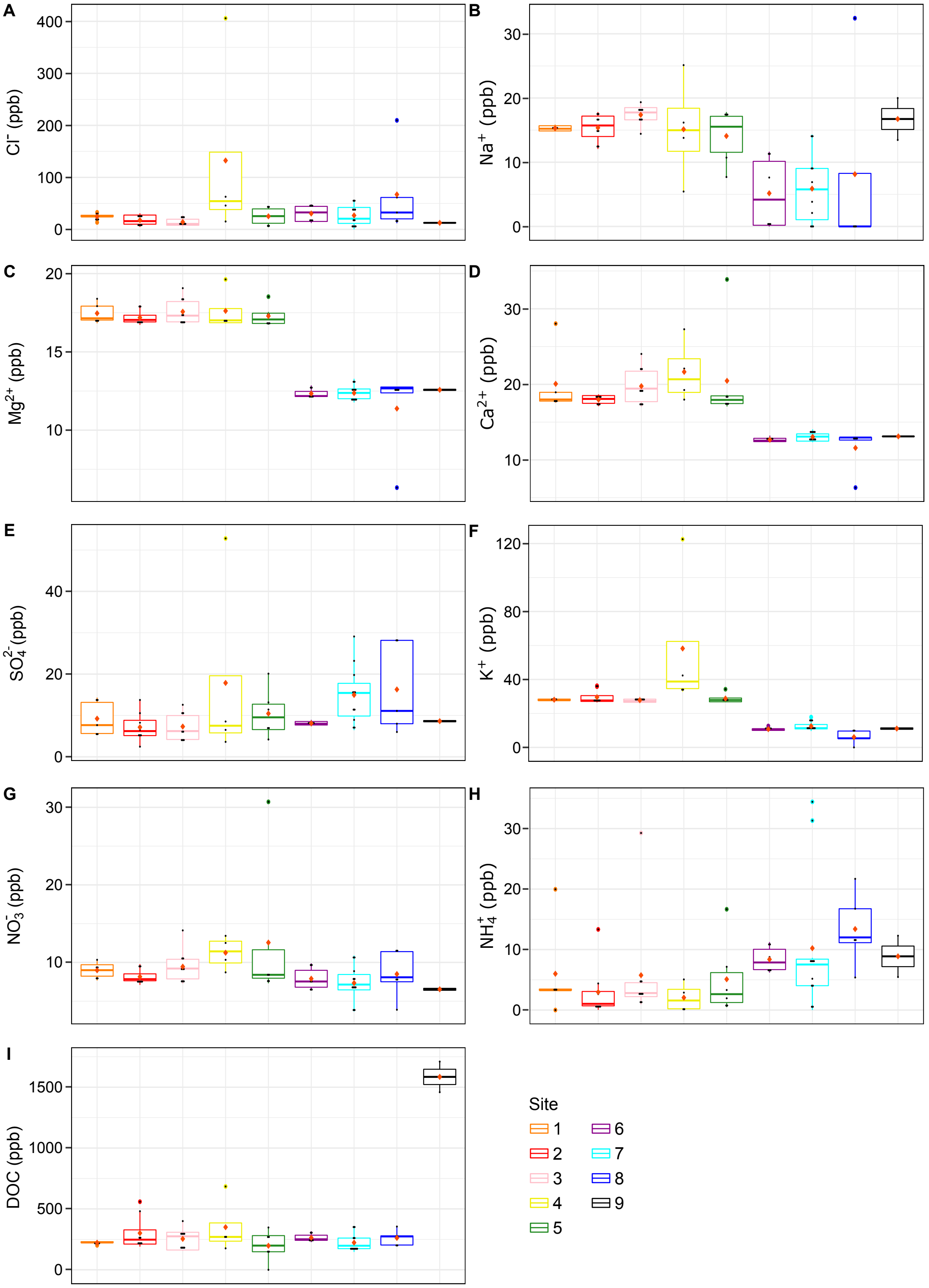


Figure 3.3: Geochemical data grouped by site for Cl⁻ (A), Na$^+$ (B), Mg^{2+} (C), Ca^{2+} (D), SO$_4{}^{2-}$ (E), K$^+$ (F), NO$_3{}^-$ (G), NH$_4{}^+$ (H) and DOC (I). All the values are reported in parts per billion (ppb).

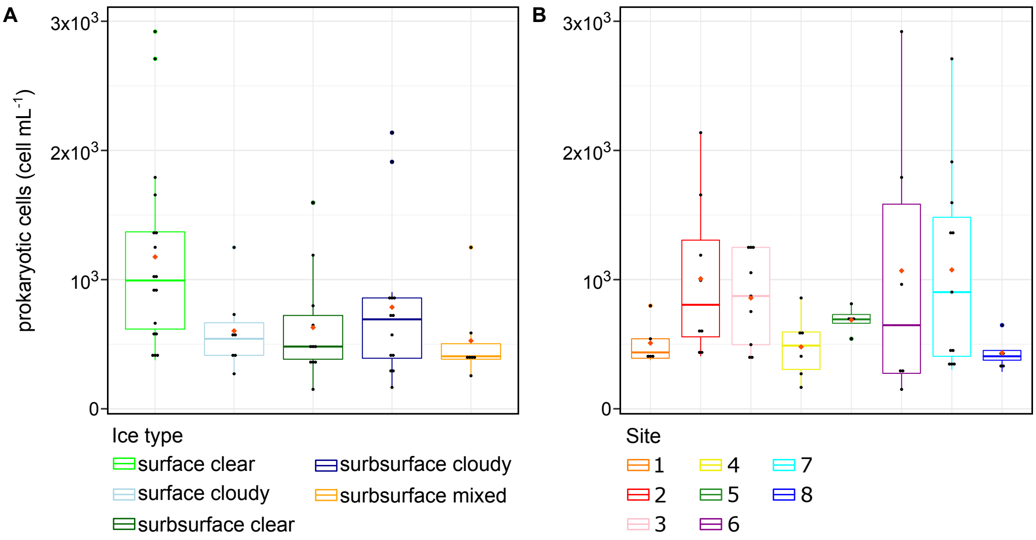


Figure 3.4: Prokaryotic cell counts for ice type-grouped samples (A) and site-grouped samples (B).

highest biovolume values with 5x10^7 and 9x10^7 µm^3 mL^{-1}. The other samples ranged between 3x10^4 and 6x10^6 (Figure A3.2A-B).

The only statistically significant factor (p-value $<$ 0.05) in the prokaryotic count and biovolume datasets was the ice type factor which explained the 17% and the 13% of the variance respectively (Table 3.2A). Less variance was explained by the model when the PERMANOVA analysis was run with ice layer (i.e. surface vs subsurface ice) as second factor (instead of the factor ice type; Table 3.2B). PERMANOVA pairwise comparisons showed that the only significant comparisons (p-value $<$ 0.05) were those between the SE samples and the other ice types. In particular, comparisons between SE and surface cloudy ice (SU), subsurface cloudy ice (U) and subsurface clear ice (E) explained 28%, 21% and 15% of the observed variance.

3.3.3 Microbial diversity

The two negative controls NC1 and NC2 resulted in 179 and 59,763 sequences, respectively (Table A3.2). NC2 sequences were represented by 219 ASVs and the most abundant ASVs (here defined as ASVs represented by more than 0.01% of the sample abundance) which represented the 97% of the sequences in this negative control, represented less than the 1% of all the sequences in all the other samples.

Between 15.2% and 73% of the sequences in all the samples were kept after the sequence clean-up and only 5 of the 62 glacier ice samples had fewer than 50,000 sequences (Table A3.2). The total number of ASVs present in the dataset was 20,509. The iNEXT diversity curves reached a plateau for the q1 (Shannon diversity) and q2 (Simpson diversity) indexes whereas they were still in an exponential phase for the q0 (ASV richness) index (Figure A3.3).

At high taxonomical (phyla-level) rank, all the ice samples presented similar communities dominated by Cyanobacteria (33.3% on average), Alphaproteobacteria (13.4%; Proteobacteria), Actinobacteria (11.7%), Bacteroidetes (11.3%), WPS-2 (10.5%), Firmicutes (5.4%), Acidobacteria (4.2%), Gammaproteobacteria (3.8%; Proteobacteria) and Armatimonadetes (2.2%). These phyla represented between the 79.8% and the 99.9% of all the taxa sampled. Phyla distribution across the different sites did not show any particular trend with the exception of a higher abundance of Armatimonadetes and Acidobacteria in sites 5, 6, 7 and 8 reaching the abundances of 13.4 and 13.5%, respectively; and Firmicutes in sites 7, 8 and 9 reaching 31.5%. In the subsurface ice, Armatimonadetes and Firmicutes had a higher ASV abundance reaching also 13.4 and 24.5% in these samples. Cyanobacteria was the most represented phylum in the dataset and reached 99.2% of bacteria in the algal samples collected from area 9, and ranged between 23% and 50% of the samples collected from SU, SE, E and M, but was constantly lower than 25% in cloudy ice.

At ASV-level the samples clustered in three main different groups where there was a first cluster composed by samples from site 1, 2, 3, 4 and 5, a second cluster with samples from sites 5, 6, 7 and 8 and a third cluster, more distantly related from the first two clusters, with samples from sites 5, 6, 7 and 8 with mainly subsurface cloudy samples (Figure 3.5A). Samples collected from site 5 clustered closed to samples from all the other sites. Additionally to the two algal samples (9A and 9B), the ice cores 6A and 8B also clustered independently from all the other samples. The Unclassified component at the genus level was between the 0.1% and the 49.9% in all the samples (9A and 3A-0-33, respectively) where most of the sequences were associated to the phyla Cyanobacteria, Proteobacteria and Armatimonadetes (Figure A3.4). Samples 9A and 9B had 99% of the sequences associated with chloroplast 16S rRNA. Clusters 1 and 2 had a consistent dominant community across all the samples where WPS-2, *Phormidesmis*, *Salinibacterium*, *Acidiphilium*, *Solitalea*, *Hymenobacter*, *Granulicella*, *Parafrigoribacterium* and *Polymorphobacter* constituted more than the 34% of the community in all the samples (except from samples 2A-83-111, 2C-25-61, 2C-61-119, 4B-17-74 and 5D-25-70). Cluster 2 and cluster 3 (site 6, 7 and 8) were also characterized by the genus *Clostridium*; and cluster 3 (subsurface cloudy samples in site 6, 7 and 8) alone was characterized by *Sediminibacterium* and *Bradyrhizobium* and a higher component of less abundant taxa (Figure 3.5B). Kruskal-Wallis tests, performed at specific genus relative distributions, showed that most of the genera varied between different sites, rather than among different ice types or layers. However, in particular, *Sediminibacterium* and *Bradyrhizobium* and *Clostridium* showed the highest Chi-squared values (Figure 3.5C). *Sediminibacterium*, *Bradyrhizobium* and *Pseudanabaena* also showed a distribution that also varied by ice type (Figure 3.5C).

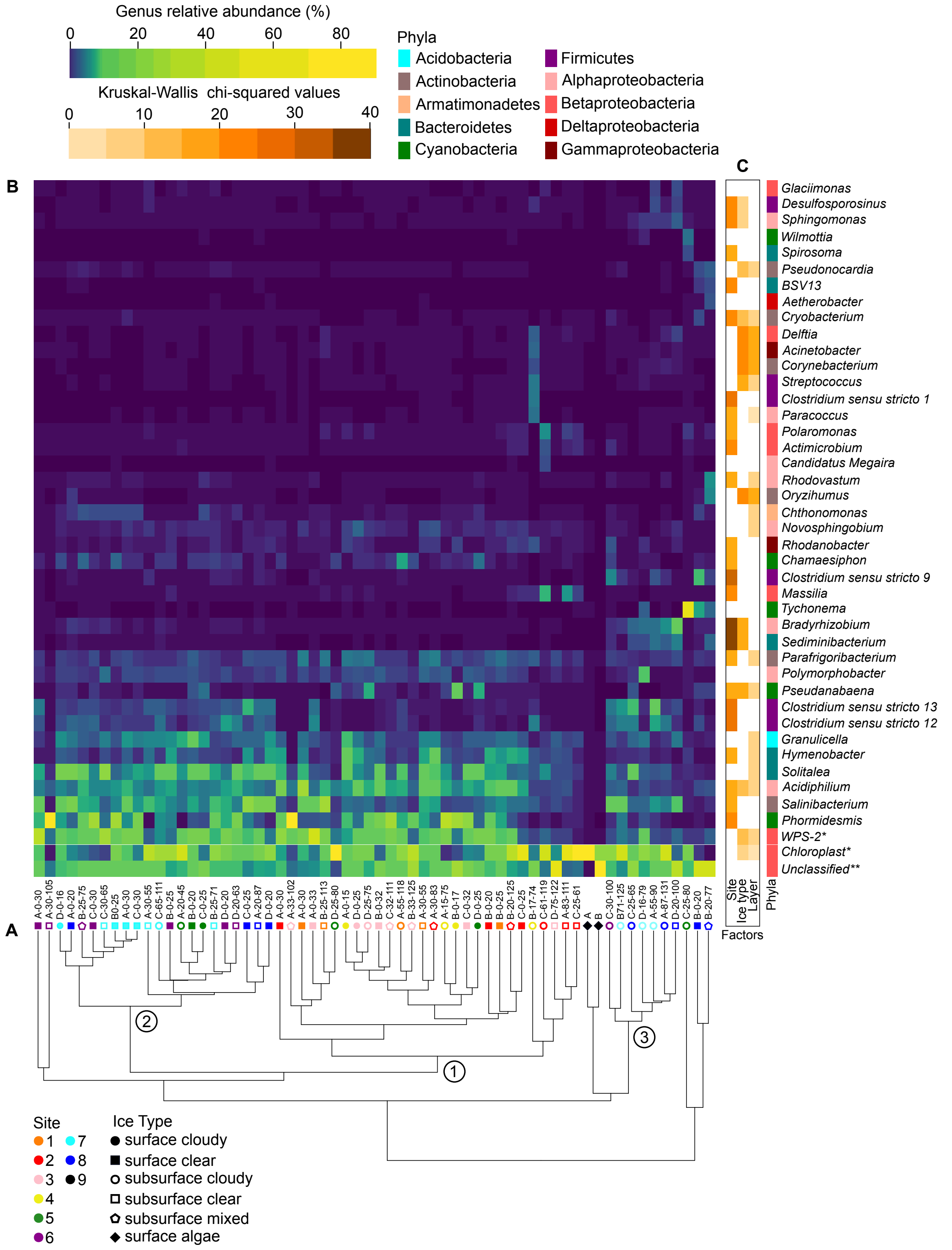


Figure 3.5: Genera abundance across the samples. (A) Cluster analysis performed on ASVs dataset transformed with the Deseq2 algorithm where the samples clustered in three main groups (1, 2 and 3). (B) Heatmap showing only the genera that represented more than 2% of the community in at least one sample of the dataset. (C) Heatmap reporting chi-squared values reported by Kruskal-Wallis tests performed on dataset without algal samples (site 9) for the factors site, ice type or layer; white boxes correspond to p-values < 0.05. *All the reported taxa are at the genus level with the exception of WPS-2 which is a phylum and Chloroplast which is an order. **The Unclassified component is explained in more details in Figure A3.4.

Table 3.3: Permutational multivariate analysis of variance (PERMANOVA) test performed on only the clear ice samples and only on the cloudy ice samples. PERMANOVA was performed with 9999 permutations on Bray-Curtis dissimilarity matrices for all the datasets except from the geochemical dataset where a Euclidean distance matrix was used. The symbol '-' is reported for non-significant R^2 values where the statistic p-value ≥ 0.05.

Datasets	**Factors**	**Clear ice**		**Cloudy ice**	
		R^2	**p-value**	**R^2**	**p-value**
Nutrients and DOC	site	0.28	0.02	-	-
	layer	0.30	0.00	-	-
	site x layer	-	-	-	-
Prokaryotic count	site	-	-	-	-
	layer	0.16	0.01	-	-
	site x layer	-	-	-	-
Biovolume	site	-	-	-	-
	layer	0.16	0.01	-	-
	site x layer	-	-	-	-
ASV	site	0.26	0.00	0.42	0.04
	layer	0.06	0.00	0.07	0.01
	site x layer	-	-	-	-

Only three ASVs, 63 sequences in total, were assigned to the Archaea component in all the dataset.

In the ASV dataset, 21% of the observed variance was explained by the factor site. The factor 'ice type' was a secondary factor explaining the 10% of the variance; 29% of the variance was explained by 'site x ice type' factor (p-value < 0.05, Table 3.2A). The 26% and the 42% of the variance was explained by the factor 'site' when the PERMANOVA was performed only on the clear and only the cloudy ASV dataset respectively (p-value < 0.05, Table 3.3).

3.3.4 Nutrients, DOC, site and taxa interactions

The clustering of subsurface cloudy ice samples of sites 6, 7 and 8 was correlated with an increase of the taxa *Clostridium*, *Bradyrhizobium*, *Salinibacterium*, *Sediminibacterium* and *Desulfosporosinus* (Figure 3.6). This is also supported by the Kruskal-Wallis test results (Figure 3.5C). This sample cluster was correlated with higher values of NH_4^+, SO_4^{2-} and Cl- and by lower concentrations of all the nutrients. All the other samples from site 6, 7 and 8 were correlated with higher values in NH_4^+. On the other hand, the group formed by the sites 1, 2, 3 and 4 was characterized by higher values of mainly K^+, Na^+, Ca^{2+} and Mg^{2+} and an increase in the genera *Corynebacterium*, *Streptococcus*, *Massilia*, *Hymenobacter*, *Pseudanabaena* and *Acidiphilium.* Ammonium had negative relation with all the other ions (Figure 3.6). The dbRDA clustering patterns corroborated those seen in cluster analysis (Figure 3.5A). DOC and NO_3^- were the geochemical variables that least affected the taxon and site distribution observed in the dbRDA

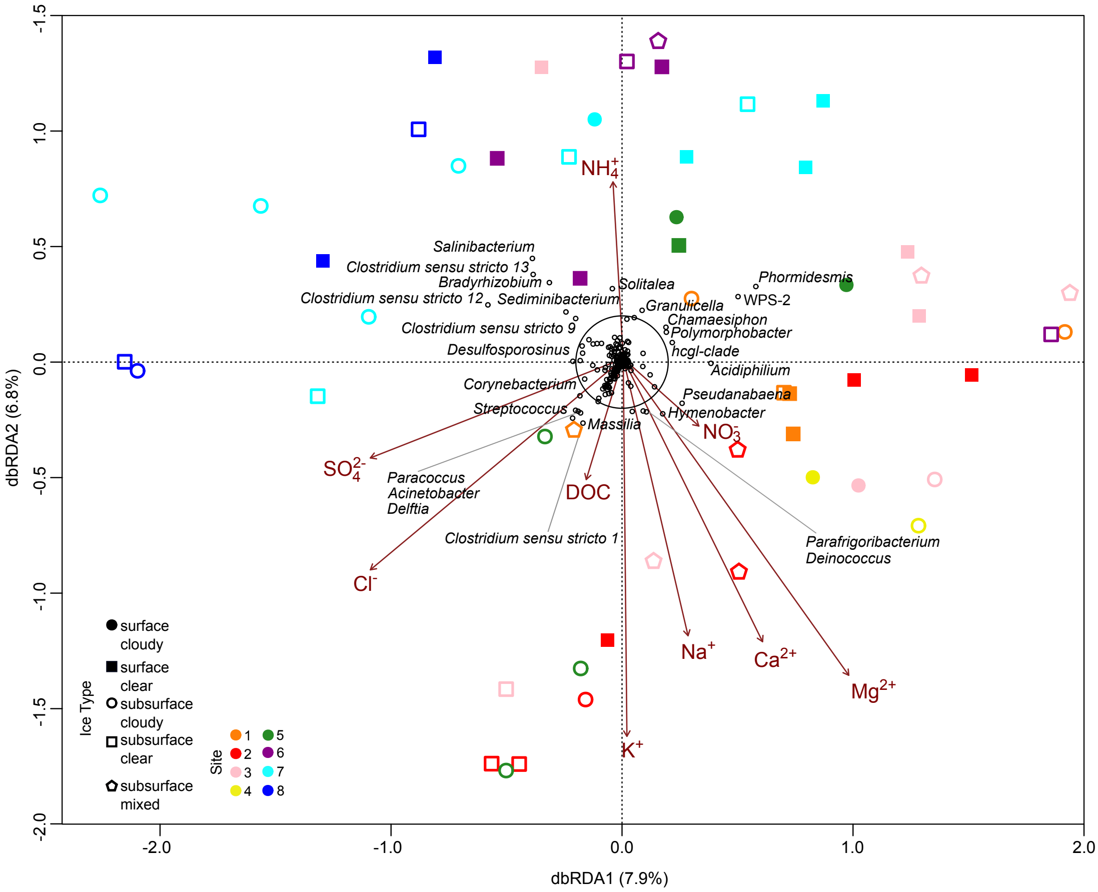


Figure 3.6: Distance-based redundancy analysis (dbRDA) bi-plot ordination performed on the Hellinger-transformed genus dataset and the geochemical dataset (Cl^-, Na^+, Mg^{2+}, Ca^{2+}, $SO_4{}^{2-}$, K^+, $NO_3{}^-$, $NH_4{}^+$ and DOC). Algae samples from site 9 were not included in the analysis. Only genera that had a dbRDA1 or dbRDA2 higher than 0.2 or lower than -0.2 were displayed in the plot. Vectors indicate direction of the geochemical variable effect in the bacterial community composition (Bray-Curtis similarity).

plot (shorter vectors).

3.4 Discussion

Common to all sampling sites were taxa previously found and isolated from other polar and cold environments, such as the genera *Phormidesmis* (Chrismas et al., 2016), *Salinibacterium* (Shin et al., 2012), *Solitalea* (Uetake et al., 2019), *Granulicella* (Oshkin et al., 2019) and *Hymenobacter* (Klassen and Foght, 2011). Most of these taxa have been described as being exopolysaccharide (EPS), ice-binding protein (IBP) and antifreeze protein (AFP) producers (Chrismas et al., 2018; Cid et al., 2016; Kielak et al., 2016). These substances have been shown to facilitate and protect cells from freeze/thaw cycles and to alter ice crystal formation therefore providing cryoprotection to advantage the survival in this challenging environment (Ali et al., 2020; Casillo et al., 2017; Deming and Young, 2017). The presence of these psychrophiles and their exudates was also supported by the microscopy (Figure 3.1G-H). The Cyanobacteria that were found in the dataset were also typical of the glacial environment (Lutz et al., 2017) and most of the genera (e.g. *Phormidesmis*, *Pseudanabaena*, *Chamaesiphon* and *Tychonema*) were filamentous cyanobacterial organisms adapted to cope with the stress imposed by challenging environments (Lan et al., 2010; Singh et al., 2010). Segawa et al. (2017) studied biogeographic patterns in cyanobacterial species colonizing glacial surfaces worldwide identifying both cosmopolitan (e.g *Phormidesmis sp.*, *Pseudanabaena sp.* and *Chamaesiphon sp.*) and local distributed species that differentiated due to site-specific conditions.

Although a distinction between clear and cloudy ice could be observed, microbial diversity and structure were mainly influenced by the location (Table 3.2A and Figure 3.5A). The main microbial differences across sites where observed between the area with sites 1, 2, 3 and 4 and the one with sites 6, 7 and 8 (Figure 3.5 and 3.6). The microbial community in sites 6, 7 and 8 differed from the others mainly because of the high abundance of the genus *Clostridium*, belonging to the phylum Firmicutes, which are spore-forming organisms (Ryall et al., 2012; Setlow, 2016). The ability to form spores gives these organisms an advantage in challenging environmental conditions, for this reason they have also been often observed as an essential part of the atmospheric microbial communities (Els et al., 2019b; Els et al., 2019a). These sites were also enriched with the phylum Acidobacteria, whose organisms are well-known acidophiles (Goltsman et al., 2015; Pankratov and Dedysh, 2010) and Armatimonadetes which is not well characterized, but it is often associated with Cyanobacteria (Woodhouse et al., 2017).

Differences between clear and cloudy ice were of secondary importance in explaining the observed variance in the taxonomy (Table 3.2A). Whereas a similar microbial community structure and diversity were shared between the clear ice across all sites and the cloudy ice across sites 1, 2, 3, 4 and 5. The microbial community of the subsurface cloudy ice of sites 6, 7 and 8 differed from the others and presented a community structure characterized by many medium-abundant taxa and a less defined dominant community (Figure 3.5B). Further, these subsurface cloudy ice samples had a lower abundance of those genera that constituted the dominant community in the other samples and that we defined above as commonly found in the polar region (e.g. *Phormidesmis* and *Hymenobacter*; Figure 3.5B). Other than *Clostridium* which had a similar abundance in all the samples of site 6, 7 and 8, the genera *Bradyrhizobium* and *Sediminibacterium* had a higher presence in the cloudy subsurface ice of these sites (Figure 3.5B-C). The genus *Bradyrhizobium* is mainly composed by plant symbiont nitrogen-fixers (Shah and Subramaniam, 2018) and *Sediminibacterium* is an ubiquitous genus often found in soil and fresh water environmental samples (Kang et al., 2014; Kim et al., 2013; Pinto et al., 2017). These two genera could have been transported to the glacier surface from the surrounding environment and trapped in the ice by the successive snow deposition and firn/ice formation. Microbes are indeed brought to the glacier mainly by weathering phenomena and aeolian transport (i.e. snowflakes and dust) and therefore, the glacial communities are strictly dependent and conditioned by the surrounding environment (Boetius et al., 2015; Hotaling et al., 2017). In the ablation zone of temperate glaciers, the ice is impermeable at depth (Fountain and Walder, 1998) and, the microorganisms, being trapped in the ice veins or crystals (Mader et al., 2006), cannot move within the ice. Therefore, englacial channels represent the only way to move within the ice. While the englacial water (i.e. clear ice) favors the development of a glacial microbial community dependent on water transport and oxygenation, the cloudy ice has never been melted/refrozen and represents a picture of the original microorganisms that were deposited on the glacier surface. These results showed how the microbial communities shared more differences between different cloudy ice samples, whereas communities in the clear ice were more homogeneous (Table 3.3).

While the contrast between clear and cloudy ice communities is clear in sites 6, 7 and 8, sites 1, 2, 3 and 4 did not show any pattern between different ice types (Figure 3.5B). Ice from different sites would have different source locations and different pathways through the glaciers before emerging in the ablation zone (Cuffey and Paterson, 2010; Hudleston, 2015). Therefore, they formed at different times and would reflect the environment at that time and source region on the glacier. Importantly, microbial communities are strongly correlated with variations in nutrient and particle concentration, which can vary dramatically spatially and temporally (Dieser et al., 2010; Lutz et al., 2016; Uetake et al., 2019).

Although Cyanobacteria are normally associated with the glacial surface because of their photosynthetic metabolism (Anesio and Laybourn-Parry, 2012; Stibal et al., 2006; Stibal et al., 2012b; Uetake et al., 2010), they have been previously found in meltwater streams (Makhalanyane et al., 2015) and in the englacial locations (Martinez-Alonso et al., 2019). Cyanobacteria was less abundant in the subsurface cloudy ice of sites 6, 7 and 8, where there is a notable absence of a microbial community typical of glacial environment. Other than this clustering, only the cyanobacterial genus *Pseudanabaena* showed a differential distribution between different ice types (Figure 3.5C). The broad presence of Cyanobacteria in the clear ice may suggest that the cyanobacterial organisms are washed into englacial systems from the glacier surface, which would be expected. The Cyanobacteria component of the dead or dormant in the englacial system may be an energy source for the heterotrophic englacial community. Similar relative abundances of Cyanobacteria between ice types may indicate a slow degradation rate of organic material in the englacial community. Remineralization of nutrients would be expected to be slow considering the relatively low abundance of bacteria in subsurface samples (average of $7x10^3$ cell mL^{-1}).

Nutrients show the same patterns as the ASV data; location represents the main explanatory factor (Table 3.2A). The concentrations of the ions Mg^{2+}, Ca^{2+}, Na^{+} and K^{+} were higher in sites 1, 2, 3, 4 and 5 compared to sites 6, 7 and 8. These ions are associated with the dissolution of soil and rock particles, therefore indicating a higher particle concentration in sites 1, 2, 3, 4 and 5 (Figure 3.3) (Li et al., 2007). Therefore, the higher presence of the spore-forming genus *Clostridium* in sites 6, 7 and 8 could be due to the low nutrient and low dust concentration in these sites. Low Ca^{2+} concentrations have previously been associated with low glacier pH (Li et al., 2007). Although we did not measure pH, sites 6, 7 and 8 showed a lower Ca^{2+} concentration compared to the other sites and were enriched with species belonging to the Firmicutes and Acidobacteria phyla, which are known acidophilic organisms (González-Toril et al., 2015). The only ion that showed a higher concentration in these sites was the ammonium which had a correlation with the presence of N_2-fixers (e.g. *Bradyrhizobium*) (Figure 3.6), suggesting a potential role of glacial organisms in producing bioavailable nitrogen.

Cell concentrations and biovolumes did not show the same patterns as those observed in the taxonomical and geochemical datasets. Instead, ice type was the major explanatory variable (Table 3.2A and Figure 3.4). Higher nutrient concentrations at a site did not correspond to a higher cell concentration which was also previously observed by Chen et al. (2016), leading to the conclusion that the nutrient presence shapes the microbial community structure, but not necessarily the microbial growth in this environment. An average of 10^4 cell mL^{-1} was observed by Grzesiak et al. (2015) in the surface glacier ice and a concentration of 10^2-10^3 cell mL^{-1} was observed in the subsurface ice (Mader et al., 2006). Surprisingly, the concentration did not change with depth suggesting that cell concentration was not influenced by sunlight intensity. This lack of microbial differences between surface and subsurface layers may be due to the typical unregular ice stratigraphy of the glacier ablation zone (Perolo et al., 2019). However, prokaryotic cell concentrations and biovolumes were higher in the surface clear ice compared to all the other ice types and largely explains the variance between ice types. Surface clear ice could represent a favorable environment for cell growth. The cells may be metabolically active during the water flow in the englacial channels and then, when exposed at the glacier surface, the organisms thrive under the new conditions of sunlight.

The glacier ice, although seemingly impermeable, is fractured by crevasses and perforated by an extensive network of englacial pathways (Fountain and Walder, 1998) which play a pivotal role in regulating flow of water and nutrients between the surface and bed. The indirect sampling approach we used to sample englacial and meteoric ice enabled us to successfully characterize microbial communities in englacial passages. Different microbial communities were found between clear and cloudy ice (sites 6, 7 and 8). The clear ice was populated by taxa typical of glacier environments whereas the cloudy ice was populated by taxa more typical of the landscape surrounding the glacier, suggesting aeolian transport to the glacier. This difference suggests that englacial channels play an important role in dispersing the microbial community inside the glacier (and presumably to the subglacial region) and in the development of a cold-adapted glacial microbial community. However, different clear/cloudy ice microbial clustering was observed in two different glacier areas (1, 2, 3 and 4 vs 6, 7 and 8), indicating that the role of the englacial channels could be highly impacted by the environment settings and weathering phenomena (e.g. particle transport) at the time of the ice formation.

3.5 Conclusion

In this study we utilized an indirect sampling method to characterize microbial communities found in englacial water. We were able to compare microbial communities from englacial water (clear ice) to meteoric glacier ice (bubble-rich cloudy ice). Although microbial communities were primarily shaped and structured by their spatial distribution on the glacier, ice type was an important secondary factor. A set of samples from one location on the glacier presented significant community differences between clear and cloudy ice. Whereas the clear ice communities presented typical cold-adapted glacial communities, the cloudy ice presented a less defined glacial community with more organisms from the surrounding non-glacial environment. The cloudy ice provides a picture of the original microbial community wind-transported to the glacier surface from the surroundings and then buried by subsequent snows, eventually compacted and turned into glacial ice. The clear ice captures that portion of the microbial community that survives in the glacial habitat. These results suggest the important role of the englacial hydrological system in the development of a glacial microbial community and its dispersion within the glacier. Metatranscriptomics studies of the englacial communities would help to further define community metabolism and the role of Cyanobacteria.

Acknowledgements

I thank Alex Anesio, Andrew Fountain and Martin Tranter to conceive this study and be so supportive. I thank Alex Holland to perform all the geochemical analyses reported in this chapter. And I thank Alex Anesio, Andrew Fountain, Gary Barker and Marian Yallop for their help in the data interpretation. I am grateful for all the assistance we received from the Tarfala Research Station crew which made our sample collection and processing smooth and efficient. I thank Jane Coghill and Christy Waterfall at the Bristol Genomics Facility for the Illumina sequencing and Miranda Nicholes for her help in the cell counting data generation. A modified

version of this chapter is in preparation to be submitted to a peer-reviewed journal: ‘Varliero, G., Holland, A., Barker, G. A. L., Yallop, M. L., Fountain, A. F., Anesio, A. M., Glacier clear ice bands indicate englacial channel microbial distribution’.

Appendix

3.A Geochemical data

Nutrients (Cl$^-$, $SO_4{}^{2-}$, $NO_3{}^-$, $PO_4{}^{3-}$, Mg^{2+}, Ca^{2+}, $NH_4{}^+$, $Na_{tot}{}^+$ and K^+) were quantified using capillary ion chromatography on a Thermo Scientific^{TM} Dionex^{TM} analytical ICS-5000, fitted with a simultaneous IonPac^{TM} AS11-HC 2 × 250 mm anion-exchange column and an IonPac^{TM} CS12 2 × 250 mm cation-exchange column. The limit of detections (LOD), determined by the mean concentration plus three times the standard deviation of procedural blanks ($n = 9$), were 8.1 ppb (Cl$^-$), 6.4 ppb ($SO_4{}^{2-}$), 8.6 ppb ($NO_3{}^-$), 16.5 ppb ($PO_4{}^{3-}$), 23 ppb (Mg^{2+}), 26 ppb (Ca^{2+}), 10 ppb ($NH_4{}^+$), 29 ppb ($Na_{tot}{}^+$) and 14 ppb (K^+). Accuracies were -0.1% (Cl$^-$), -3.4% ($SO_4{}^{2-}$), -0.5% ($NO_3{}^-$), -5.5% ($PO_4{}^{3-}$), -14% (Mg^{2+}), -6.5% (Ca^{2+}), -14% ($NH_4{}^+$), -14% ($Na_{tot}{}^+$) and -20% (K^+). Precisions were ± 0.47 (Cl$^-$), ± 2.0 ($SO_4{}^{2-}$), ± 1.0 ($NO_3{}^-$), ± 2.7 ($PO_4{}^{3-}$), ± 5.4 (Mg^{2+}), ± 3.7 (Ca^{2+}), ± 2.9 ($NH_4{}^+$), ± 3.7 ($Na_{tot}{}^+$) and ± 6.7 (K^+), as determined from comparison with a gravimetrically diluted single ion 1000 mg L^{-1} FlukaTM TraceCERT® ion chromatography standard to a concentration of 250 µg L^{-1} for each ion. In the chapter Na^+ values are reported as rock dissolved $Na_{rock}{}^+$. $Na_{rock}{}^+$ was calculated with the formula $Na_{sea}{}^+ = Cl_{tot}{}^-$ x (10760/19350) where $Na_{rock}{}^+ = Na_{tot}{}^+ - Na_{sea}{}^+$. The values 10760 and 19350 are respectively the concentrations of Na^+ and Cl$^-$ in sea water. Some of $Na_{rock}{}^+$ values that equal to 0 may be slightly negative values.

Dissolved Organic Carbon (DOC) concentration was quantified using a Shimadzu TOC-V_{WP} Organic Carbon Analyzer. Total carbon (TC) is the sum of inorganic carbon (IC) and DOC. TC was measured via the addition of phosphoric acid and persulfate to the sample, which was heated under UV radiation and converted to CO_2 where it was measured using non-dispersive infrared analysis (NDIR). IC was quantified by acidifying the sample with phosphoric acid and sparged to convert it to CO_2, where it was measured in the same way as TC. DOC was determined by subtracting the IC concentration from the TC concentration. The LoD was 28.1 ppb. Precision was ± 1.3 and accuracy was 2.3% as determined from comparison with a gravimetrically diluted 1000 ppm TOC certified stock standard to a concentration of 250 ppb (Sigma TraceCERT®).

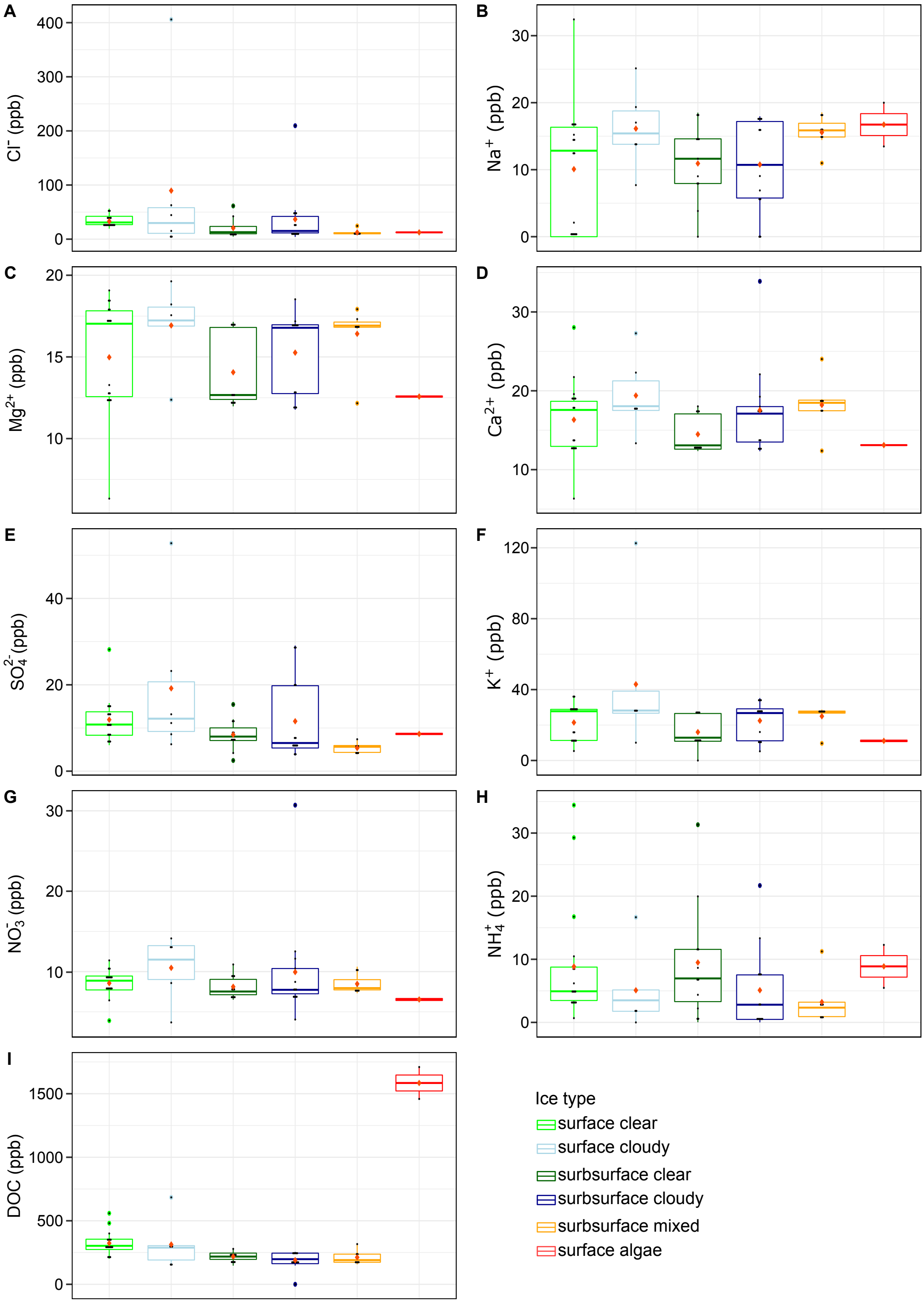


Figure A3.1: Geochemical data grouped by ice type for Cl- (A), Na^+ (B), Mg^{2+} (C), Ca^{2+} (D), $SO_4{}^{2-}$ (E), K^+ (F), $NO_3{}^-$ (G), $NH_4{}^+$ (H) and DOC (I). All the values are reported in parts per billion (ppb).

3.B Sequencing data

Table A3.1: Primer and Illumina Nextera Transposase adapter sequences.

Primer	**Sequence**
Pro341F	CCTACGGGNBGCASCAG
Pro805R	GACTACNVGGGTATCTAATCC
Pro341F + Illumina adapter	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNBGCASCAG
Pro805R + Illumina adapter	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACNVGGGTATCTAATCC

Table A3.2: Retrieved sequences after each step of the clean-up stage of the sequence ASV dataset.

Samples	Raw reads*	Primer removal*	Quality filter*	Forward ASV inference	Reverse ASV inference	Merged sequences	Non chimeric sequences	No negative	Final sequences (%)
1A-0-30	224564	219028	147351	145808	146201	139451	48648	48645	21.7
1A-30-55	203044	198742	143915	142100	143118	137634	74791	74789	36.8
1A-55-118	167352	162727	113510	113002	113242	111612	99616	99608	59.5
1B-0-25	179107	173887	120995	119210	120151	114182	45125	45122	25.2
1B-25-113	170690	156925	114383	114132	114274	113484	109964	109563	64.2
2A-0-30	254054	242719	176107	172617	174298	160215	44174	44167	17.4
2A-30-83	232178	226812	160031	158555	159226	154211	91455	91449	39.4
2A-83-111	166476	161533	117664	117137	117246	116136	108265	108253	65
2B-0-20	247998	242468	175622	172973	174184	164097	54306	54299	21.9
2B-20-125	252206	247178	177946	176212	176922	170389	99663	99653	39.5
2C-0-25	193916	189124	133404	132329	132786	129122	84636	84627	43.6
2C-25-61	192041	186185	137105	136430	136632	135366	127992	127934	66.6
2C-61-119	207339	201145	146786	145727	146168	142394	123708	123696	59.7
3A-0-33	35377	29553	10960	10885	10919	10662	7608	7607	21.5
3A-33-102	204826	182011	119906	118839	119411	115417	52589	52587	25.7
3B-0-32	154170	149194	100592	99969	100246	98188	76042	76039	49.3
3B-32-125	207275	203027	150866	150467	150579	149662	142298	142261	68.6
3C-0-32	320693	313216	228157	224974	226618	215319	114578	114555	35.7
3C-32-111	204325	198718	142253	141614	141752	139750	99070	99068	48.5
3D-0-25	257276	251180	185815	184666	185291	181892	139837	139832	54.4
3D-25-75	278733	273891	201909	201270	201404	199079	162891	162877	58.4
3D-75-122	1444473	1402145	751041	748378	749435	741485	691489	690952	47.8
4A-0-15	175536	170975	119295	118645	118918	116701	96471	96467	55
4A-15-75	207551	202957	144025	142734	143341	139103	78394	78389	37.8
4B-0-17	404452	264449	166465	165087	165742	161878	93358	93355	23.1
4B-17-74	223110	213245	157634	157053	157039	155614	153028	152784	68.5
5A-20-45	237112	232547	176737	176486	176402	175483	172027	172007	72.5
5B-0-20	178178	173539	118222	117473	117770	115019	90089	90085	50.6
5C-0-25	198069	193195	139064	138466	138684	137068	126716	126716	64
5C-25-80	231545	226883	168990	166940	167438	162968	153438	153087	66.1
5D-0-25	309798	304136	235492	231596	233954	222030	120326	120325	38.8
5D-25-70	236421	231121	177559	177139	177194	176294	163759	163730	69.3
6A-0-30	104562	100062	67929	66836	67314	62052	15848	15844	15.2
6A-30-105	264310	253674	186641	183195	185485	170988	56950	56949	21.5
6B-0-25	177917	172102	118512	117138	117744	113330	81841	81796	46
6B-25-75	148744	145580	97725	97368	97435	96301	88713	88710	59.6
6C-0-30	124675	121993	78662	78569	78535	78110	75995	75987	60.9
6C-30-100	149431	138242	93684	92045	93073	87341	66223	65296	43.7
6D-0-20	227031	222888	166380	165754	165971	163992	145493	145343	64
6D-20-63	189345	184637	130027	129503	129666	127749	112841	112831	59.6
7A-0-30	196250	192735	146898	146777	146769	146362	143291	143290	73
7A-30-55	158457	153128	106410	105451	106013	102577	86248	85188	53.8
7A-55-90	195491	173677	123858	122578	123226	118725	104033	102679	52.5
7B-0-25	256721	251884	194295	194071	194103	193282	185712	185712	72.3
7B-25-71	135460	132480	94667	94326	94437	93460	87727	87720	64.8
7B-71-125	308641	287586	205466	203689	204298	197776	119984	119911	38.9
7C-0-30	172561	169003	123767	123539	123618	122850	115339	115339	66.8
7C-30-65	245112	241578	179422	179240	179294	178643	176433	176406	72
7C-65-111	167840	161555	108498	107660	108080	105491	83264	83181	49.6
7D-0-16	264053	259587	187566	187341	187330	186247	183380	183358	69.4
7D-16-79	191445	179008	129421	127626	128717	123709	104803	103288	54
8A-0-20	257590	253744	188197	188027	188031	187333	184833	184821	71.8
8A-20-87	185394	181061	131355	130403	130812	127997	95255	95239	51.4
8A-87-131	270297	264641	201383	200754	201077	199987	197053	196480	72.7
8B-0-20	237481	233665	167847	167575	167598	166689	164433	164424	69.2
8B-20-77	238134	229540	156850	155934	156475	153944	147335	147330	61.9
8C-0-25	224394	218942	154811	153439	153986	148823	85633	85629	38.2
8C-25-65	178650	170465	120005	117818	119263	113112	91635	91063	51
8D-0-20	232212	226815	162010	161296	161566	159396	135607	135582	58.4
8D-20-100	366006	359788	270644	270316	269959	268983	265547	263233	71.9
9A	155250	151114	112921	112543	112783	111782	95444	95388	61.4
9B	2717700	2597221	1499585	1493666	1497640	1475254	1279538	1278606	47
NC1	2512	1274	238	236	235	213	179	-	7.1**
NC2	126406	118604	63685	63189	63408	61476	59763	-	47.3**

Table A3.3: Singleton percentage over each sample sequence and ASV number.

Samples	Singletons over sequence number (%)	Singletons over ASV number (%)
1A-0-30	0.95	51.05
1A-30-55	0.44	40.68
1A-55-118	0.05	16.39
1B-0-25	1.57	54.46
1B-25-113	0.02	8.26
2A-0-30	2.17	63.78
2A-30-83	0.39	41.17
2A-83-111	0.02	10.4
2B-0-20	1.36	54.97
2B-20-125	0.36	35.54
2C-0-25	0.44	40.87
2C-25-61	0.02	9.96
2C-61-119	0.07	19.53
3A-0-33	0.32	12.44
3A-33-102	0.59	53.62
3B-0-32	0.19	27.17
3B-32-125	0.01	6.94
3C-0-32	0.32	41.08
3C-32-111	0.18	31.88
3D-0-25	0.1	23.74
3D-25-75	0.05	20.33
3D-75-122	0.01	24.03
4A-0-15	0.16	31.29
4A-15-75	0.35	39.42
4B-0-17	0.18	33.33
4B-17-74	0.03	21.08
5A-20-45	0.01	3.21
5B-0-20	0.17	25.12
5C-0-25	0.03	9.33
5C-25-80	0.06	20.47
5D-0-25	0.27	40.12
5D-25-70	0.01	6.96
6A-0-30	6.63	72.18
6A-30-105	1.03	67.47
6B-0-25	0.29	33.77
6B-25-75	0.04	10.48
6C-0-30	0.01	3.45
6C-30-100	0.32	41.62
6D-0-20	0.05	18.8
6D-20-63	0.06	17.83
7A-0-30	0.01	4.12
7A-30-55	0.17	32.6
7A-55-90	0.06	19.93
7B-0-25	0.01	5.28
7B-25-71	0.04	8.61
7B-71-125	0.2	38.07
7C-0-30	0.03	9.77
7C-30-65	0.01	5.24
7C-65-111	0.18	33.64
7D-0-16	0.01	8.45
7D-16-79	0.1	31.83
8A-0-20	0.01	4.95
8A-20-87	0.17	28.25
8A-87-131	0.01	6.06
8B-0-20	0.02	11.63
8B-20-77	0.03	14.29
8C-0-25	0.42	39.65
8C-25-65	0.15	31.96
8D-0-20	0.06	16.09
8D-20-100	0.01	7.93
9A	0.12	45.35
9B	0.02	44.25

3.C Biovolume data

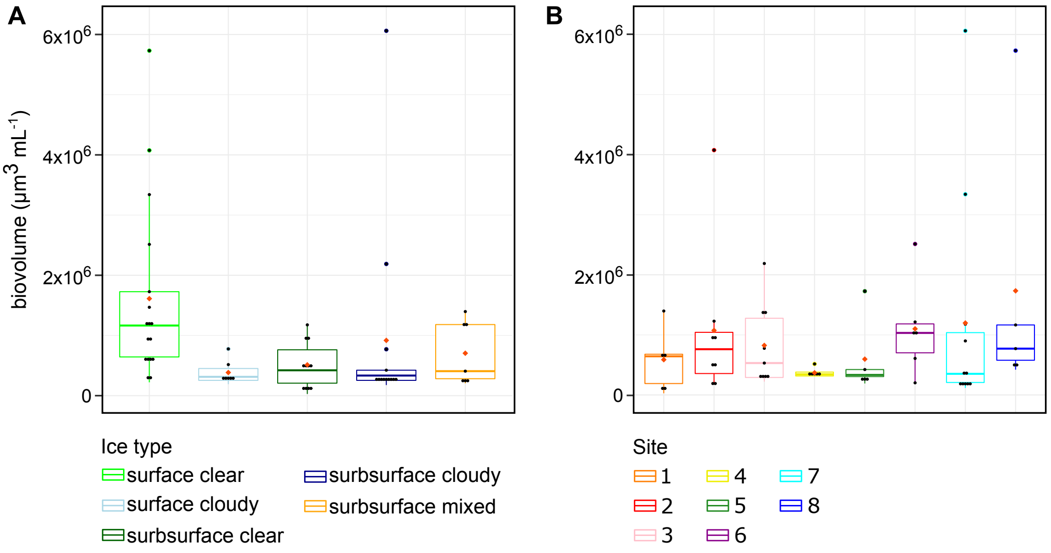


Figure A3.2: Biovolumes values for ice type-grouped samples (A) and site-grouped samples (B).

3.D Diversity data

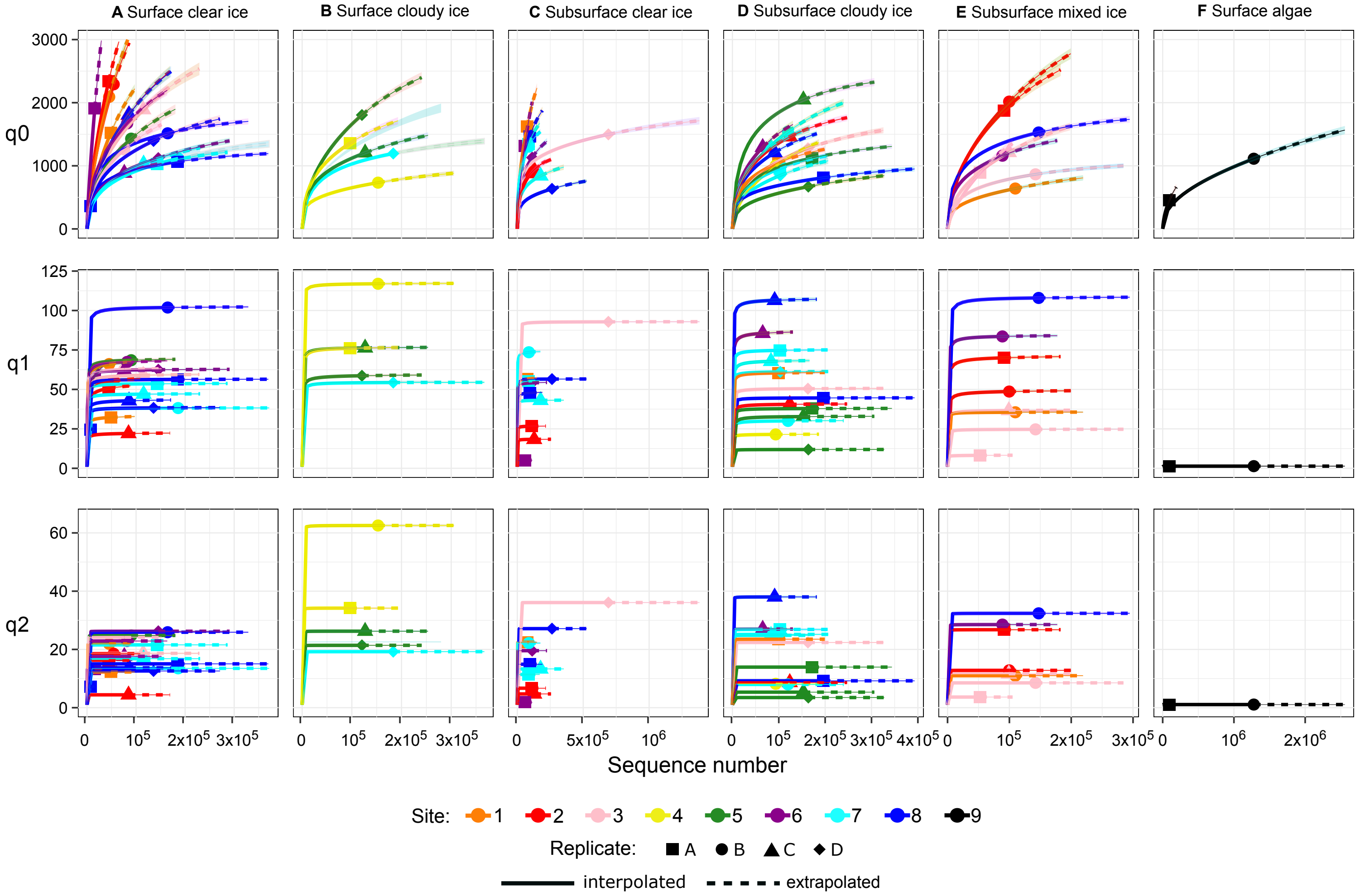


Figure A3.3: iNEXT rarefaction curves calculated on the q0 (species richness), q1 (Shannon's diversity index) and q2 (Simpson's diversity index) for surface clear samples (A), surface cloudy samples (B), subsurface clear samples (C), subsurface cloudy samples (D), mixed subsurface samples (E) and surface algal samples (F).

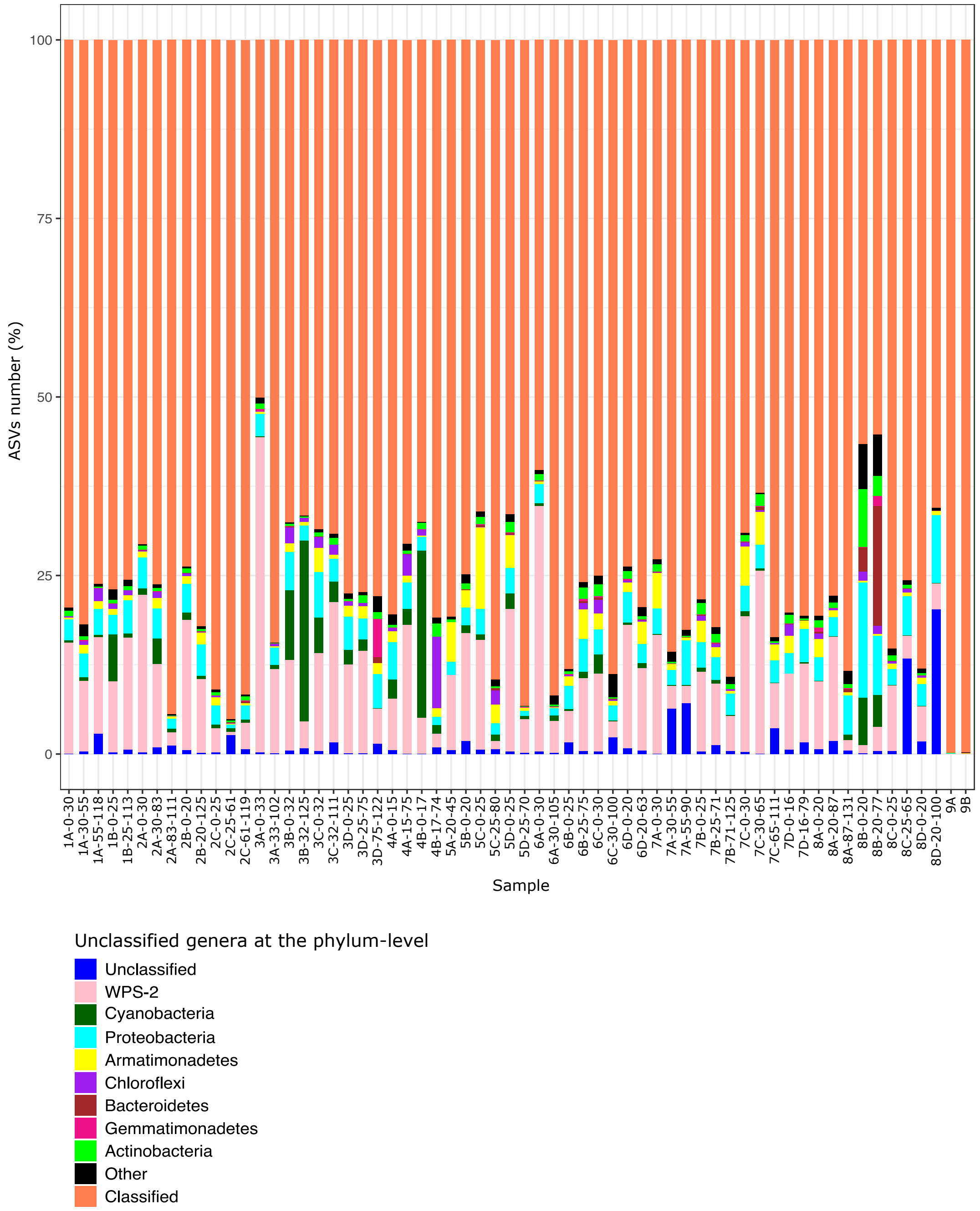


Figure A3.4: Phyla corresponding to the unclassified component at genus-level. Classified refers to the component that was classified at genus-level. Only phyla that represented more than 5% in at least one sample were reported.

Chapter 4

A taxon-wise insight into rock weathering and nitrogen fixation functional profiles of proglacial systems

Abstract

The Arctic environment is particularly affected by the global warming and a clear trend of the ice retreat is observed worldwide. In proglacial systems, newly exposed terrain represents different environmental and nutrient conditions compared to later soil stages. Therefore proglacial systems show several environmental gradients along the soil succession where microorganisms are active protagonists of the soil and carbon pool formation through nitrogen fixing and rock weathering phenomena.

In this chapter I studied the microbial successions of three different Arctic proglacial systems sited in Svalbard, Sweden and Greenland. From these systems, sixty-five surface soil samples were collected and whole shotgun sequenced obtaining more than 400 Gb of sequencing data. This dataset was then assembled together and taxonomic and functional annotated with Long-Meta. The latter is a new pipeline that I developed inside this study frame that focuses on the retrieval of specific functional annotation from taxa of interest, and vice-versa, giving a taxon*gene resolution.

Thanks to this pipeline, I was able to explore and compare microbial successions of the three different proglacial systems which showed common trends typical of these environments but also biogeographic taxonomic and functional diversity. Further, I explored genes related to nitrogen fixation and biotic rock weathering processes such as nitrogenase genes, obcA genes and genes involved in cyanide and siderophore synthesis and transport. Whereas I confirmed the presence of these genes in known nitrogen fixing and rock weathering organisms, in this study I also present organisms that, even if often found in soil and proglacial systems, have never been related to these processes before. The different genera showed different gene trends within and among the studied systems, indicating a community constituted by a plurality of organisms involved in these processes and where nitrogen fixation and biotic rock weathering were driven by different organisms at different soil succession stages.

4.1 Introduction

4.1.1 Glacier forefield processes

Due to the global warming, a clear trend of glacial ice retreat has been observed in recent decades worldwide (Maurer et al., 2019; Moon et al., 2018). This rapid loss of the cryosphere is impacting the ecosystems, increasing the glacial water discharge in the environment and leading to an expansion of the proglacial systems and to a consequent increase in the exposure of bedrocks that have been covered by the ice for thousands of years (Fountain et al., 2012; Heckmann et al., 2016). Chemical, physical and biological processes lead to the formation of soil from the mineralization of the bedrocks (Uroz et al., 2009; Uroz et al., 2015; Xi et al., 2018) in the process of weathering. Biological rock weathering is a key process in environments where the main role is played by soil microbiota and vegetation roots (Borin et al., 2010; Kelly et al., 1998; Porder, 2019) which help the release of nutrients and major ions into the soil with their metabolism and mechanical actions. The release of nutrients and major ions represent a source of enzyme cofactors and energy for the soil microbiota, especially in nutrient-poor soils, such as the early-stage forefield soils (Uroz et al., 2015), giving microbes a pivotal role in biogeochemical cycles (Koshila Ravi et al., 2019; Rousk and Bengtson, 2014). Proglacial systems are involved in another key environmental and ecological process: nitrogen fixation. This environment is a habitat of diverse diazotrophic communities that progressively enrich the soil with ammonia and bioavailable nitrogen sources to non-diazotrophic organisms (Bradley et al., 2014; Nash et al., 2018).

Rock weathering and nitrogen fixation create several gradient conditions in the proglacial environment, going further away from the ice front. Whereas the ground is dominated by rocks close to the ice edge, the soil increases and deepens going further away from it, with an associated increase in the vegetation. The bioavailability of nutrients, such as organic carbon and nitrogen, also increases with the formation of the soil. The presence of such gradients and the progressive ground exposure to the atmosphere makes proglacial systems very suitable for the study of the microbial successions (Edwards and Cook, 2015).

Microbial communities show trends along chronosequences (Bajerski and Wagner, 2013; Jiang et al., 2019; Schmalenberger and Noll, 2010; Zumsteg et al., 2012). Previous work has shown that the microbial communities close to the ice edge are usually dominated by autotrophs and chemolithotrophs which are able to use soil minerals and sunlight as energy source and enrich the soil with biological available organic carbon and nutrients (Fernández-Martínez et al., 2017; Liu et al., 2012; Schmidt et al., 2008). These first stages of the succession are also the most influenced by the glacier inputs and discharges in the environment (Hotaling et al., 2017). Going further from the ice, different studies have observed a decrease of the autotroph component and an increase in the heterotroph microbial component able to take advantage of the progressive organic-enriched soil (Bradley et al., 2016). These trends are also accompanied by an increase of the vegetation complexity with the distance from the ice edge, establishing also symbiotic and mutualistic relationships with the soil microbiome (Knelman et al., 2012; Rime et al., 2016).

In spite of the pivotal role of the microbial communities in the rock weathering, the protagonists and the mechanisms of these processes are not very clear. Different mechanisms and rock weathering enhancing organisms have been discovered and found thanks to the study of soil isolates of both bacteria (Frey et al., 2010; Lepleux et al., 2012; Liu et al., 2012; Olsson-Francis et al., 2015; Wang et al., 2019; Xi et al., 2018; Wongfun et al., 2014) and fungi (Brunner et al., 2011), observing the production of organic acids (e.g. oxalate) and hydrogen cyanide (HCN) to mobilize the nutrients such as iron, sulfur and phosphorus (El-Tarabily et al., 2008), and an increase in siderophore production to import iron into the cell. Different functionality trends have been identified in forefields with GeoChip data (Fernández-Martínez et al., 2016) and also specific environmental enzymes (Jiang et al., 2019; Li et al., 2020), focusing on enzymes involved in S and Fe metabolisms (Mitchell et al., 2013) or enzymes involved in polymer lysis and uptake which increased in the later stages of the succession (Liu et al., 2012).

Compared to rock weathering processes, diazotrophic organisms are better understood and characterized. Organisms spanning more than 13 phyla have been identified as nitrogen fixers (Addo and Dos Santos, 2020). However, even if this process is essential to the forefield dynamics, there is a lack of understanding of the diazotrophs organism variation along the proglacial microbial succession (Brankatschk et al., 2011; Nash et al., 2018).

4.1.2 Case studies

Whereas previous studies have analyzed Arctic forefield microbial communities with several approaches, none has studied taxonomy and functional succession profiles with a shotgun metagenomics approach. In order to have a comprehensive picture of gene function over glacial chronosequences, we analyzed 65 different metagenomes from three different proglacial systems: two forefields from two small glacier valleys, the Midtre Lovénbreen in Svalbard and the Storglaciären in Sweden, and a proglacial field of the Greenland ice sheet (GrIS) in proximity of point 601.

This dataset was already object of a publication to which I contributed (Nash et al., 2018). In that manuscript, however, we uniquely explored the diazotrophic communities and how they varied among the three proglacial systems and in relation to the measured Total Nitrogen (TN) and Total Organic Carbon (TOC) concentrations for each of the systems. Furthermore, we did not investigate any microbial succession distribution in the forefield systems.

In this chapter I have extended the previously published analysis of the data by investigating i) how taxonomical groups varied along the different proglacial successions and ii) which organisms were involved in different proglacial processes. I focused on the two processes that shape forefield dynamics and nutrient bioavailability the most: nitrogen fixation, exploring nitrogenase genes, and rock weathering processes, looking at the obcA genes which are involved in the first step of the oxalate biosynthesis (Nakata, 2011), genes involved in cyanide synthesis and genes involved into siderophore synthesis and transport.

4.1.3 LongMeta, a new pipeline for sequencing analysis

There are many tools and approaches for assigning taxonomy and genes to individual sequencing reads (Oulas et al., 2015; Roumpeka et al., 2017; Thomas et al., 2012). However, none of the published pipelines and software allow us to analyze and explore the functional profiles at specific taxon-levels and to answer questions such as ‘Which genes a specific taxonomic group could potentially express?’. For this reason, I have developed a similarity-based pipeline that integrates both taxonomy and gene-level functionality assignment giving an insight into gene composition at taxon level.

Gene assignment is more accurate when performed on long rather than short sequences (Tamames et al., 2019). The main reason why I did not use an unsupervised approach (e.g. binning and ORF prediction) to assign gene and taxonomy to this dataset is because of the poor assembly quality that is usually obtained from complex low-coverage environmental datasets. In this case, the suboptimal contig quality, specifically indels and truncated sequences, challenges ORF-prediction algorithms. The studied datasets are constituted by soil communities which are generally recognized as the most complex and diverse, posing even more challenges to read assembling and characterization (Howe et al., 2014; Schloss and Handelsman, 2006). I implemented my approach with the use of publicly available software and algorithms, such as Diamond (Buchfink et al., 2015) and bowtie2 (Langmead and Salzberg, 2012), and with custom Perl scripts available on GitHub (https://github.com/gvMicroarctic/LongMeta).

4.2 Materials and methods

4.2.1 Sample collection and sampling site

Samples were collected from the Midtre Lovénbreen forefield during summer 2013 and, from Storglaciären forefield and the proglacial field of the Greenland ice sheet in proximity of point 601, during summer 2014. The sampling was conducted with the same approach in all the systems: only the first 10 cm of soil was collected in a sterile Whirl-pack bag and then frozen at -20 °C till the processing. The samples were collected from the soil along transects going from the proximity of the ice edge, going further away. Samples were not collected close to vegetation patches, rivers or discontinuous soil patches to avoid site specific effects as much as possible. The three different systems are sited in the Arctic circle and are above the 67° N parallel (Figure 4.1A). These systems present different morphologies and have a different deglaciation rate, the ice sheet being much slower than the glaciers in the small valleys and the collected soil was also exposed to the atmosphere at different timescale. The sampling size and the geographical characteristics in the three different systems was also considerably different. The samples in Greenland were taken up to 10 km away from the closest ice point (Figure 4.1B). The samples in Svalbard were taken up to 1600 meters from the glacier toe and, as the forefield faces on a fjord, this last point of the succession is sited close to the sea water (Figure 4.1C). Regarding Sweden, samples were taken up to a river that delimits the end of the small forefield area, at 350 meters (Figure 4.1D). This diversity and the geographical dispersion of this sites allowed to compare functional and taxonomical trends between different systems. In this chapter the Greenland proglacial system will be named G, the one in Svalbard will be called SV and the one in Sweden SW. The samples were classified by distance from the glacier toe and the distance was calculated as the distance between the sampling site and the closest ice edge point.

4.2.2 Geochemical data

Soil total organic carbon (TOC) and total nitrogen (TN) were analyzed by Dr. Maisie Nash at the University of Bristol glaciology laboratory as described in Nash et al. (2018).

4.2.3 DNA extraction and sequencing

Genomic DNA was extracted with PowerMax soil DNA isolation kit (MO BIO Laboratories, Carlsbad, CA, USA) according to the manufacturer's instructions from 5-10 grams of soil per sample. Sixty-five different samples were selected for Illumina sequencing. All the sequencing libraries were prepared with TruSeq Nano DNA kit (Illumina, San Diego, CA, USA) and then whole shotgun-metagenome sequencing was performed with HiSeq 2500 Illumina platform for SV

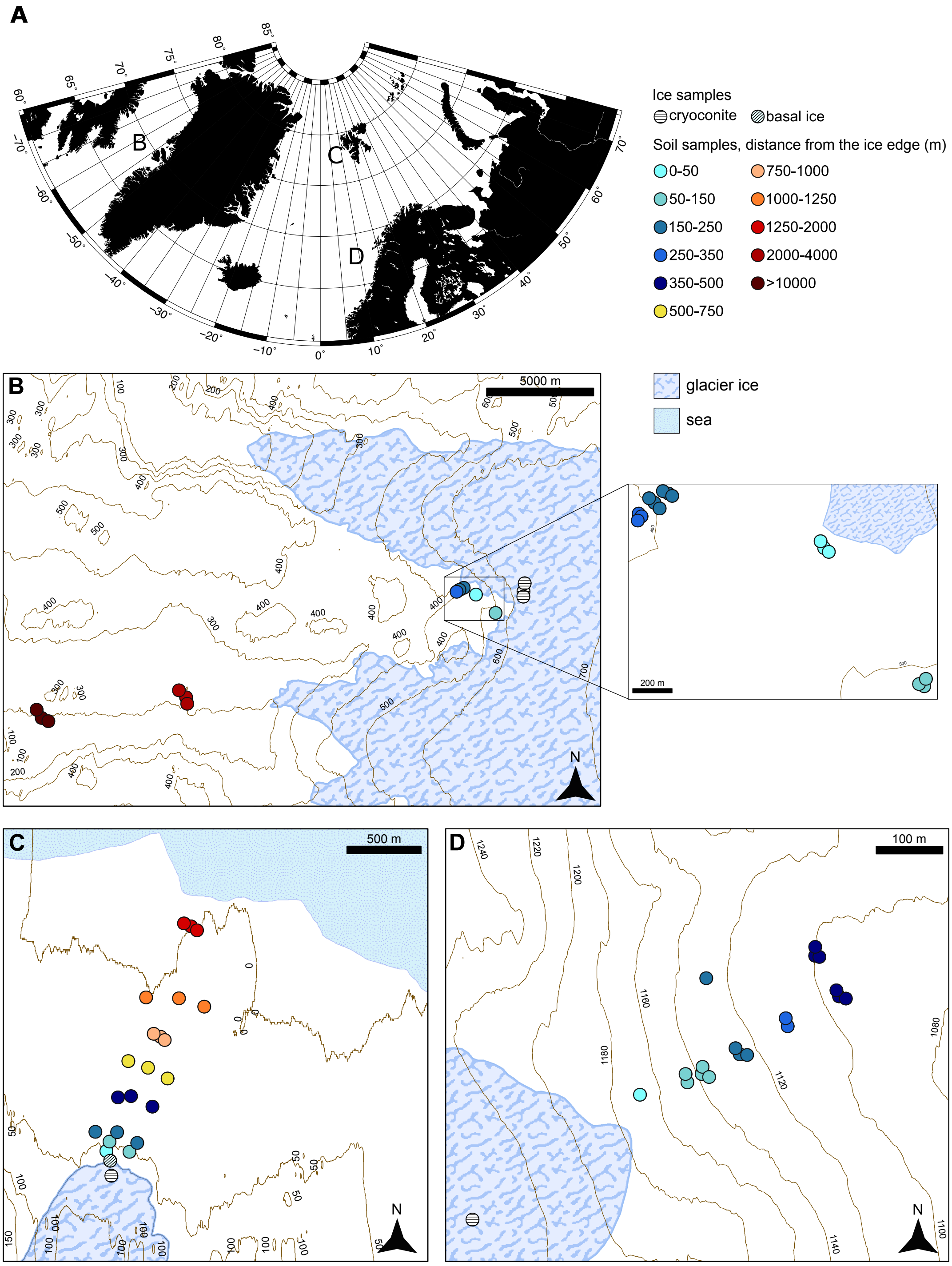


Figure 4.1: Sampling site locations. (A) Overview of the sampling site. (B) Greenland forefield sampling site. (C) Svalbard forefield sampling site. (D) Sweden forefield sampling site.

samples (PE 2x100), and with NextSeq 500 platform for from G and SW datasets (PE 2x150) by the Bristol Genomics facilities. Illumina sequence basecalling was performed with the Real-Time Analysis (RTA) software v 2.4.6.

4.2.4 LongMeta implementation

LongMeta relies on an assembly approach in which all the Illumina reads are assembled together. Successively, the taxonomy and gene coding regions are assigned to the assembly contigs by aligning a database of known sequences to it. Single sample taxonomy and gene abundances are then obtained by mapping the reads back to the assembly and calculating the contig and coding region coverage on each taxonomic assigned contig and coding region (Figure 4.2).

Before running LongMeta, the Illumina reads were checked, and quality trimmed with FastQC v 0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic v 0.36 (Bolger et al., 2014). The latter was run with the command options `ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 MINLEN:26`, and the additional options `TRAILING:10 for SV, SLIDINGWINDOW:5:20 CROP:149` for SW, and `SLIDINGWINDOW:5:24 MINLEN:26 CROP:149`

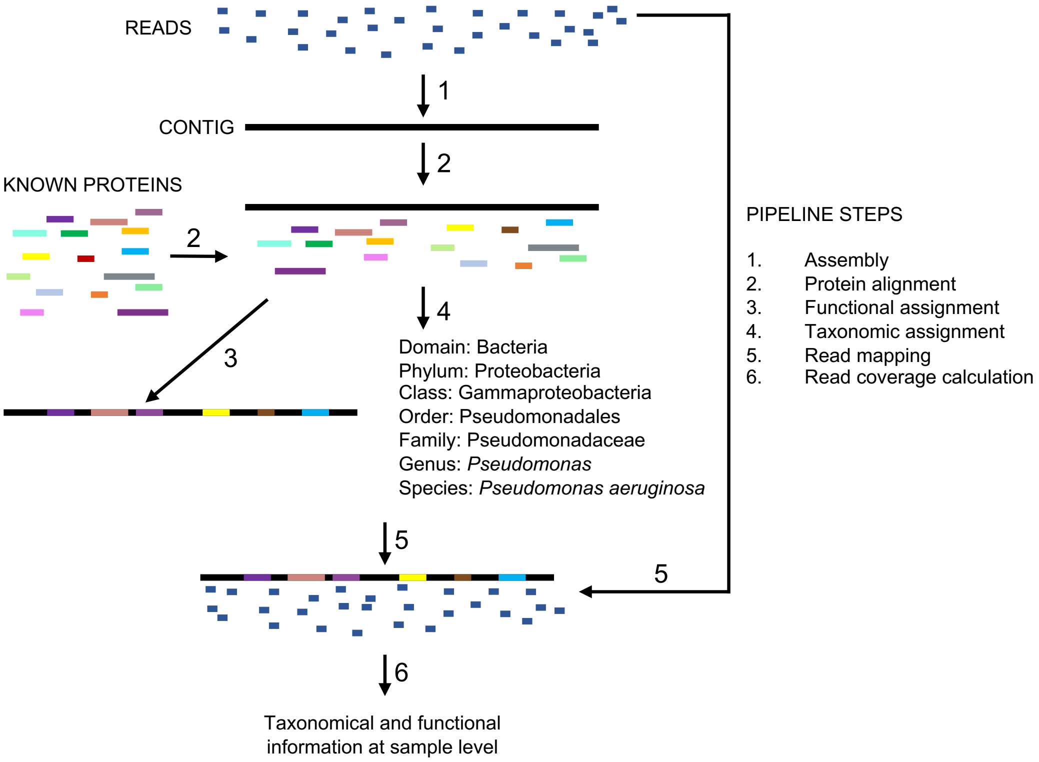


Figure 4.2: LongMeta pipeline. All the reads are first assembled together, then all the contigs are assigned to the rightful taxonomy and coding region by mapping known protein sequences to them. Finally, the reads are mapped back in order to obtain contig and coding region coverage information and calculate taxonomic and gene relative abundances at sample level.

for G. The sequences were then co-assembled from the 3 different datasets with the software MEGAHIT v 1.1.3 (Li et al., 2015). The assembly was performed on the Illumina trimmed paired reads with the parameters (`--k-min 25 --k-max 145 --k-step 8 -t 64`).

4.2.4.1 Assembly quality check and Diamond blast

The script `longMeta-summary` was used to check and remove contigs shorter than 300 bases from the assembly. All the contigs were then aligned against the NCBI non-redundant (nr) database v5 (Sayers et al., 2020), a comprehensive protein database comprising 257,100,652 proteins, with Diamond 0.9.22 (Buchfink et al., 2015) using the command line options `-e 0.000001 -F 15 --range-culling --range-cover 20 --id 50 --top 10 -f 6 -c1 -b4.0`.

4.2.4.2 Functional and taxonomic characterization

Gene and taxonomy were assigned to the contigs through the screening of the Diamond file (tabular format). The assignment was performed by a script called `longMeta-assignment` which assigns non-overlapping coding region to the sequences. Non-overlapping genes are defined as those genes that overlap for at most three bases independently on their strand sense. Most of the microbial genes, in fact, overlap only with few bases (Cock and Whitworth, 2007; Huvet and Stumpf, 2014). This criterion may be too stringent if the study targets are, for example, viruses where the gene overlap has been shown to be higher (Pavesi et al., 2013). The number of overlapping bases allowed can be changed with the command option `--overlap`. The potential protein coding regions are assigned by selecting the best Diamond protein alignments. First, the algorithm orders the alignments from the highest to the lowest identity score (IS) and in case of more alignments having the same IS, it orders the alignments from the highest to the lowest bit-score (BS). The algorithm assigns the best protein alignment and saves the alignment coordinates of the protein. This section of the genome is now 'blocked', and no other protein can be assigned to it. The protein parsing restarts and every time the algorithm finds a protein that does not overlap to any previously assigned protein, it assigns the new protein to the DNA sequence. The process goes on until all the protein-matches and the DNA sequences have been processed (Figure 4.3A and Figure A4.1).

If a section of the genome was aligned to multiple database proteins with the same IS and BS, only one random protein will be assigned but all the proteins will be kept for the taxonomy assignment (default, `--max-equal`). Once the gene coding sequences are assigned, the pipeline screens only assigned proteins that have an identity score higher than 80 (`--cutoff-ID-best`). If more of one such matches are present and they all have a concordant taxonomy at genus level

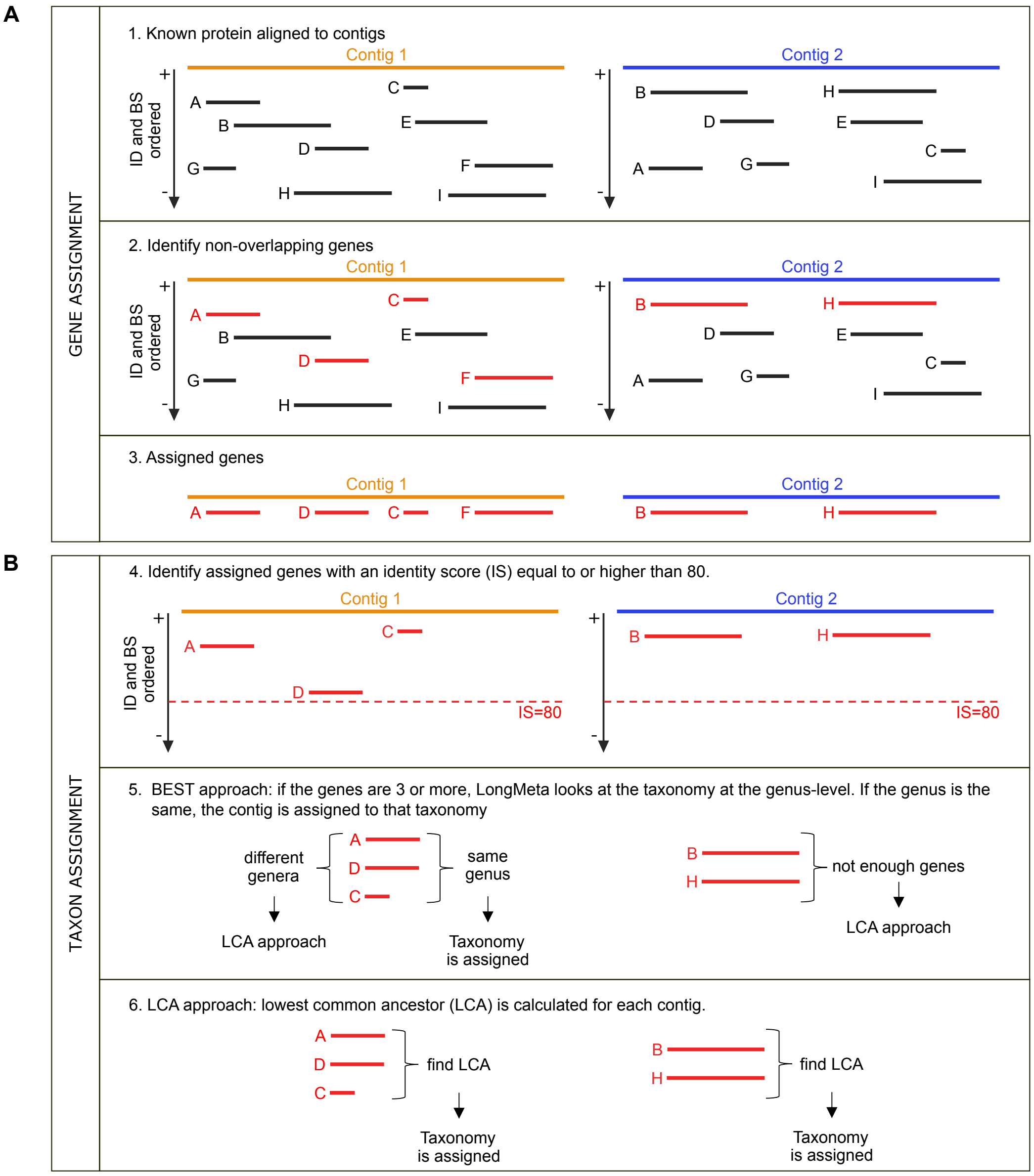

A
GENE ASSIGNMENT
1. Known protein aligned to contigs
Contig 1
Contig 2
ID and BS ordered
+
-
A
B
C
D
E
F
G
H
I
2. Identify non-overlapping genes
3. Assigned genes
B
TAXON ASSIGNMENT
4. Identify assigned genes with an identity score (IS) equal to or higher than 80.
IS=80
5. BEST approach: if the genes are 3 or more, LongMeta looks at the taxonomy at the genus-level. If the genus is the same, the contig is assigned to that taxonomy
different genera
same genus
LCA approach
Taxonomy is assigned
not enough genes
LCA approach
6. LCA approach: lowest common ancestor (LCA) is calculated for each contig.
find LCA
Taxonomy is assigned
find LCA
Taxonomy is assigned

Figure 4.3: LongMeta pipeline where gene (A) and taxonomy (B) is assigned to assembly contigs.

(`--min-best`), the sequence is assigned to that taxonomy. We define this as the BEST approach. If these two conditions were not met, the pipeline runs a weighted top-down lowest common ancestor (LCA) algorithm in order to assign the taxonomy to the remaining sequences. This LCA algorithm considers the IS and an averaged bit-score (ABS). The BS is highly influenced by the alignment length and therefore the pipeline divides it by the alignment length giving an indication of the BS per base (ABS) and providing a result which is not skewed towards long genes (Figure 4.3 and Figure A4.2). The alignment taxonomy is analyzed with a top-down approach, screening the matches from the rank domain down to the genus level. Double weight is given to the alignments that have an IS higher than 80 (`--cutoff-ID-lca`). Proteins that show a high IS when aligned to the studied sequence are more likely to belong to organisms highly related to it and therefore are given more weight in the LCA calculation.

The LCA algorithm is used in many pipelines and usually priorities accuracy over precision as when in doubt between different low-level taxonomies, it assigns the sequence to a higher-up taxonomy. On the contrary, the BEST approach increases precision over the accuracy directly assigning taxonomy when it detects a clear taxonomy signal. This two-step taxonomy assignment allows us to identify the taxonomy at low-level in sequences. The use of this hybrid approach is advised when working with contigs. In case the user works with reads, the LCA approach is advisable. The user can also select the use of the so-called BEST approach to select only the contigs where there is a clear signal of the taxonomy.

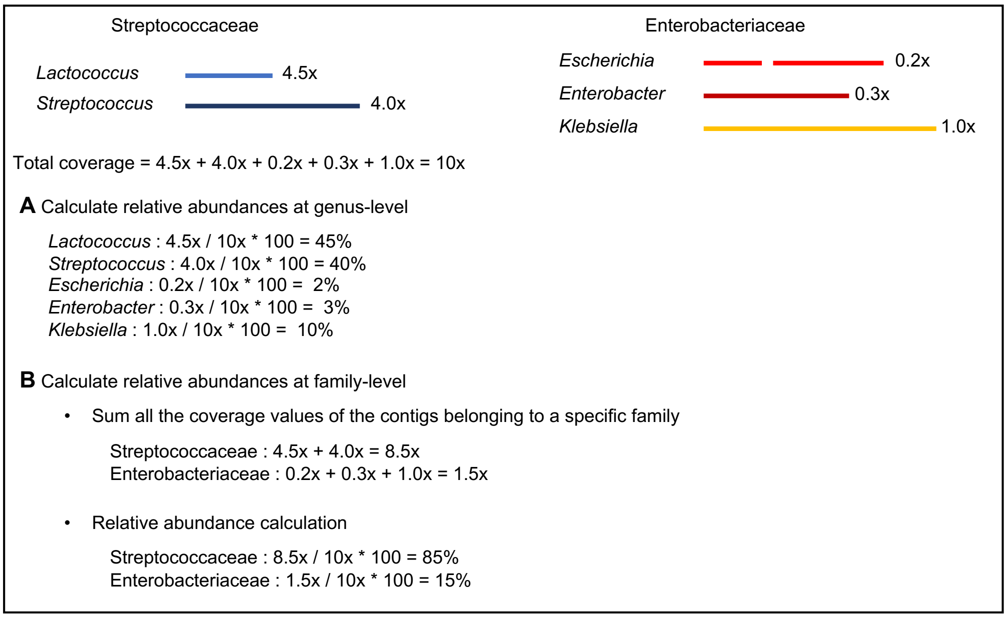
4.2.4.3 Chimera detection

The assembly can be screened for chimeric contigs. The script `longMeta-chimera` checks if a contig contains one or more taxonomically differing high identity clusters (HICs). The latter is defined as a cluster of three or more consecutive high identity genes (IS $\geq$ 80) in a row. If more than one HIC is found in the contig sequence, the latter will be divided into multiple sub-contigs. Each sub-contig will be assigned to the taxonomy specific of its HIC. The rank level, the identity score threshold and the number of genes used to define an HIC can be changed by the user (`--taxon-level`, `--ID-limit` and `--cluster-limit`).

4.2.4.4 Relative abundance estimations

In order to assign specific taxonomy and functions to each sample, reads must be mapped back to the assembly, for which we used bowtie v 2 2.3.4.3 (Langmead and Salzberg, 2012) with the parameters `--phred33 --local -I 100 -X 800 --no-hd --no-unal -D 30 -R 3 -N 0 -L 20 -i S,1,0.25 --non-deterministic -p 20`.

The script `longMeta-coverage` was used to assign taxonomy and gene content to each sample. Taxonomy coverages are reported as the bowtie-mapped base coverage in all the contigs assigned to a specific taxonomy. Similarly, gene coverages are reported as the base coverage in all the coding region assigned to a specific gene. This script excludes contigs assigned to Metazoa or Streptophyta (by default, `--ignore-uncl`) and also taxonomies that are represented by less than 1% of the contigs and represent less than 1% of the bases compared to the entire assembly (`--perc-limit`). The minimum taxonomic level that LongMeta works with is the genus rank. Taxonomical relative abundance is calculated by the script `longMeta-relative`. When working at the genus-level, the relative abundance is simply calculated by dividing the genus coverage by the total coverage in a sample (Figure 4.4A). When working at higher taxonomic



Total coverage = 4.5x + 4.0x + 0.2x + 0.3x + 1.0x = 10x

A Calculate relative abundances at genus-level

Lactococcus : 4.5x / 10x * 100 = 45%
Streptococcus : 4.0x / 10x * 100 = 40%
Escherichia : 0.2x / 10x * 100 = 2%
Enterobacter : 0.3x / 10x * 100 = 3%
Klebsiella : 1.0x / 10x * 100 = 10%

B Calculate relative abundances at family-level

- Sum all the coverage values of the contigs belonging to a specific family

  Streptococcaceae : 4.5x + 4.0x = 8.5x
  Enterobacteriaceae : 0.2x + 0.3x + 1.0x = 1.5x

- Relative abundance calculation

  Streptococcaceae : 8.5x / 10x * 100 = 85%
  Enterobacteriaceae : 1.5x / 10x * 100 = 15%

Figure 4.4: Taxonomic relative abundance calculation at genus level (A) and family level (B).

ranks, the script calculates the average coverage between all the genera belonging to the same taxonomic rank and then divides the averaged coverage by the total coverage in sample (Figure 4.4B).

4.2.4.5 Databases

The databases used by LongMeta can be found and downloaded from https://github.com/gvMicroarctic/LongMeta. The GitHub page also explains how to set the LongMeta pipeline and the required databases up. The databases are:

- Protein database. I have used the NCBI-nr database v5 (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz).
- Protein accession number correspondences with taxid numbers which are unique numbers, each corresponding to a different taxon (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz).
- Taxid number correspondences with taxonomy paths (ftp://ftp.ncbi.nih.gov/pub/taxonomy/new_taxdump/new_taxdump.tar.gz).
- Protein name correspondences to the protein accession numbers. Protein nomenclature information was retrieved from the nr database sequence headers and formatted with a custom script where protein names were trimmed and curated in order to have a more unified protein nomenclature.

- Gene Ontology (GO) number correspondences to protein accession number. The information was retrieved from http://current.geneontology.org/annotations/goa_uniprot_all.gaf.gz and ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz (Ashburner et al., 2000; Carbon et al., 2019).
- Gene Ontology (GO) number correspondences to GO category name. The information was found at http://purl.obolibrary.org/obo/go.obo.

All the files were formatted with custom scripts. LongMeta can also be used with user's customized databases.

4.2.4.6 Data exploration

This pipeline outputs many different files. I created `longMeta-explore`, to process these outputs into usable formats. The user can easily retrieve only the gene information for specific taxonomy also relying on GO classification if wanted. In this study, for example, I was able to easily retrieve gene coverage information at genus level for the gene nitrogenase, obcA, cyanide synthase and siderophore related genes.

4.2.5 Statistical analyses

The Geochemical dataset and LongMeta abundance and coverage datasets were imported to the R environment (R Core Team 2019, 2019) where all the statistical analyses were performed. In all the graphical visualization and the principal component analysis (PCA), the samples were categorized in relation to their distances from the ice edge and divided into different groups: 0-50, 50-150, 150-250, 250-350, 350-500, 500-750, 750-1000, 1000-1250, 1250-2000, 2000-4000 and >10000 meters (Table A4.1).

Taxonomic data was reported and analyzed as taxonomic relative abundance calculated as explained in Section 4.2.4.4. Diversity indices were calculated on the genus relative coverage values with Shannon's (H) and Inverse Simpson's (1/D) diversity indices. The diversity values were fitted with lowess curves in order to detect diversity trends in relation to the site distance from the ice edge.

Functional data was reported and analyzed as weighted gene coverage. The coverage of a specific gene in a specific sample was divided by the base coverage of the sample. The coverage scaling was necessary because different samples assembled with different efficiencies depending on their sequencing depth and community complexity. This led to a differential base coverage in different samples and therefore the gene coverages, if not scaled, could reflect the sample coverage and not the real gene coverage in the microbial community. In this chapter, further to the gene coverage, I also present functional profiles as Gene Ontology (GO) biological categories.

Permutational multivariate analysis of variance (PERMANOVA) was performed with 9999 permutations on Bray-Curtis dissimilarity matrices for the taxonomy, diversity and functional datasets, and Euclidean distance matrices for the geochemical dataset. The statistical tests were performed against the fixed factor ‘forefield‘ which has three levels: Greenland, Svalbard and Sweden.

In order to show how the three different datasets varied with the sample distance from the ice edge and in relation to the geochemical dataset (i.e. TN and TOC), I calculated Mantel test statistics r using 9999 permutations and the Spearman’s rank correlation coefficient. Mantel tests were conducted on either Bray-Curtis or Euclidean matrices as for the PERMANOVA.

Whereas the Mantel test was used to compare two different datasets (i.e. two different sets of variables), I calculated directly the Spearman’s rank correlation coefficient r_s when I wanted to compare only two different variables (e.g. a specific phylum vs distance from the ice edge). I have used the linear model calculation only when comparing two variables with the same unit (i.e. Proteobacteria vs Actinobacteria relative coverage).

All the statistical analyses reported in the main text were performed without the inclusion of ice samples, unless otherwise specified. Statistical tests were interpreted as significant if p-value < 0.05.

I used several R packages to perform statistical analyses and plot the graphs: vegan v 2.5.6 (Oksanen, 2017), ggpubr v 0.3.0 (Kassambara, 2018), ggplot2 v 3.3.0 (Wickham, 2016), gplots v 3.0.3 (Warnes, 2012), reshape v 0.8.8 (Wickham, 2007) and gridExtra v 2.3 (Auguie, 2017).

4.3 Results

4.3.1 Geochemical data

Total nitrogen (TN) and total organic carbon (TOC) showed different trends among the three proglacial systems (Figure 4.5). The Greenland (G) dataset showed low values in the first stages of the succession, maximum values at 150-250 meters and a gradual decrease going further away from the glacier. The Svalbard (SV) and Swedish (SW) datasets showed an increase from the ice edge to more distant samples. SW showed the lowest TN and TOC concentrations with values 0-2 and 0-7 mg g^{-1} in all the soil samples. The SV had a similar value range with 0-1 and 0-6 mg g^{-1}, respectively, in all the soil samples except from the SV samples collected at 1650 meters from the ice edge (the closest samples to the sea; Figure 4.1C) where the TN and TOC mean values were 4 and 55 mg g^{-1}. Compared to the SV and SW systems, G showed the highest

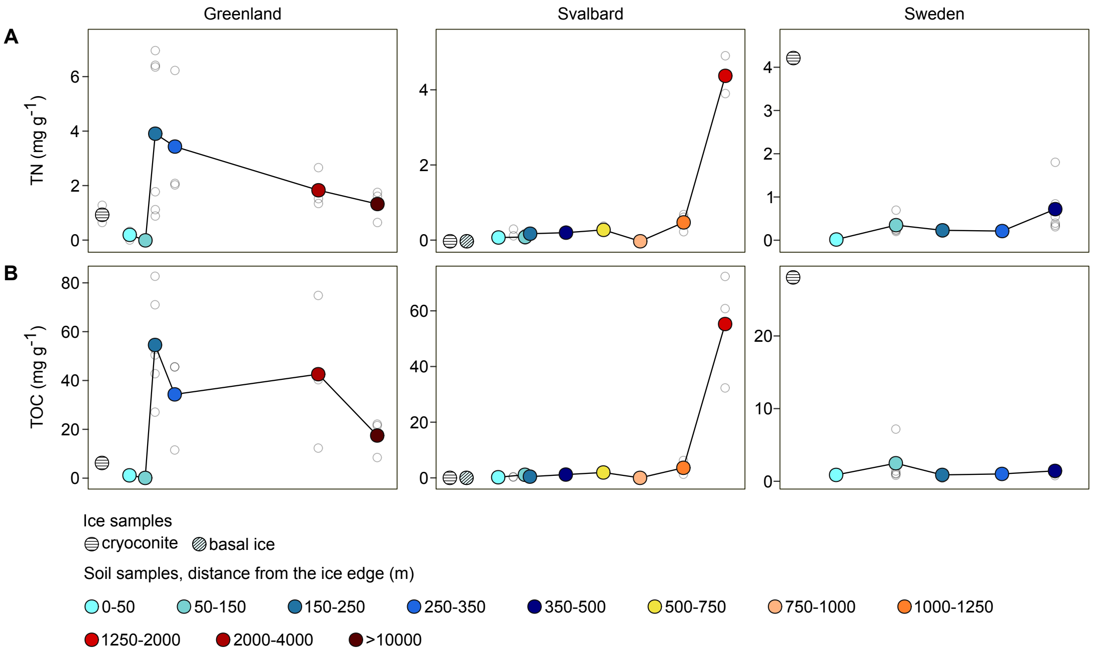


Figure 4.5: Total Nitrogen (TN) (A) and Total Organic Carbon (TOC) (B) trends across the three different proglacial systems. Samples were grouped into different categories and colored in relation to sample distance from the ice edge. Colored dots indicate the average values for each distance whereas white dots are the values of the individual samples.

TN and TOC values ranging 0-7 and 0-82 mg g^{-1}.

The two variables, TN and TOC, showed a positive Spearman's correlation across all the systems ($r_s = 0.89$; p-value < 0.05). Sample separation among the different forefield systems (factor 'forefield') explained the 26% of the observed variance in this dataset (p-value < 0.05). The correlation between these two variables and the distance from the ice edge was then explored with a Mantel test statistic r which was equal to 0.22 (p-value < 0.05). Different r values were obtained when Mantel tests were performed on the different forefield datasets: correlation with the sample distance from the ice edge was significant only in the SV and SW forefields with an r of 0.39 and 0.33 (p-value < 0.05), respectively.

4.3.2 Assembly specifics

The quality checked reads that were used to perform assembly were between 4 and 119 million per sample (Table A4.1) for a total of 433 billions of bases (433 Gb) used. The assembly was 30 Gb long, the N50 was of 841 bases and the minimum contig length was 300, whereas the maximum was 561,967. Ice sample reads mapped back to the contigs with a higher percentage compared to the soil samples where 93-94%, 75-93% and 92% of the reads mapped back for the G, SV and SW forefields. Whereas the percentages were 50-78%, 34-64% and 65-79% for the soil samples (Table A4.1). The average base coverage of the assembly ranged between 0.3x and

1.5x in the samples from the SV dataset, 0.7x and 3.7x in the SW dataset and for 0.7x and 3.4x in the G dataset. In each dataset, the highest coverages were observed in the ice samples. In the forefield assembly, 1% of the contigs were defined as chimeric and therefore split into shorter contigs.

4.3.3 Taxonomy diversity and trends

The Inverse Simpson's diversity index (1/D) ranged between 25 and 128 across all the samples, whereas the Shannon's diversity index (H) ranged between 4 and 5. In all the proglacial systems, both indices were lower in the ice samples where the 1/D reached a maximum of 54 and H a maximum of 4.4. Both the diversity indices had overall lower values in the Svalbard forefield. Fitted lowess lines showed increasing diversity trends along all the forefield (Figure 4.6).

PERMANOVA performed on the diversity dataset (i.e. 1/D and H) showed that only the 20% of the variance was explained by the factor 'forefield' (p-value $<$ 0.05) (Table 4.1A). Diversity indices did not show a correlation with the distance from the ice edge, nor with TOC or TN concentration values. However, when performed on the different forefield datasets, the Mantel test statistic r was significant (p-value $<$ 0.05) for the SV dataset (r = 0.36) (Table 4.1E). When it was performed on soil and ice samples it was significant for all the datasets, where the r for

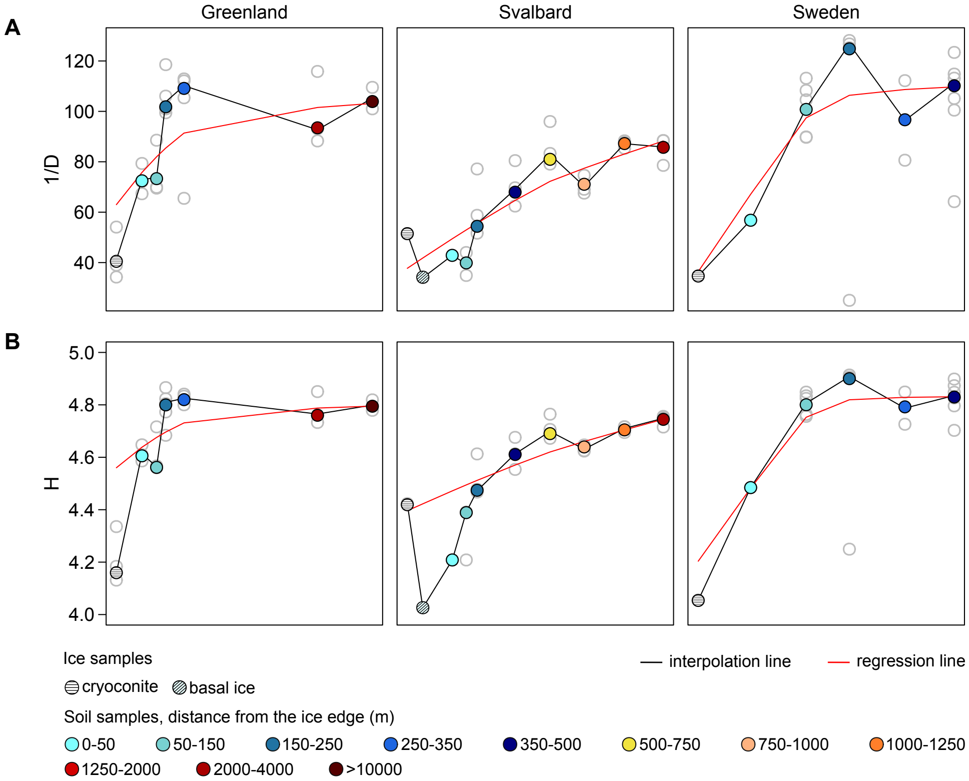


Figure 4.6: Inverse Simpson's (A) and Shannon's (B) diversity indices (1/D and H) calculated on the genus-level taxonomic dataset.

G was 0.20, for SV was 0.45 and for SW was 0.18 (Table A4.2E).

Microbial communities showed a minor sample clustering distribution across different forefields (Figure 4.7A). Ice and ice-edge samples (< 50 m from the glacier) clearly separated from the soil samples in the PCA representation, whereas soil samples collected at different distances from the ice edge did not show clear distinct clusters between each other (Figure 4.7B).

PERMANOVA performed on the taxonomy dataset was significant (p-value < 0.05) for the factor 'forefield' which explained the 13, 22 and 23% of the variance at phylum-, order- and genus-level (Table 4.1B). Less variance was explained when the PERMANOVA was performed including the ice samples (Table A4.2B). The Mantel test performed between distance and taxonomical dataset (genus-level) was not significant. When the Mantel test statistic r was calculated on the separate forefield datasets (i.e. G, SV and SW), the variable distance showed a significant (p-value < 0.05) r of 0.19, 0.61 and 0.24 for G, SV and SW, respectively (Table 4.1F). Only the G dataset showed a significant correlation to the geochemical dataset (i.e. TN and TOC) with r equal to 0.42 (Table 4.1F).
Figure A4.3 reports how the taxonomic dataset (genus-level) is influenced by the sample distance from the ice edge and the variables TN and TOC. Ice communities and early-stage soil communities were mainly influenced by the genera *Thiobacillus*, *Purpureocillum*, *Methylotenera*, *Cryobacterium* whereas samples sited more distant in the microbial succession were conditioned by *Solirubribacter*, *Hyphomicrobium*, *Chthoniobacter* and *Mycolicibacterium.* TOC and TN shaped the distribution of several genera, such as *Pseudolabrys*, *Bradyrhizobium* and *Rhodoplanes.*

At the phylum level, the two most abundant phyla, representing between 58 and 84% in the soil and 47 and 62% in the ice samples, were Proteobacteria and Actinobacteria. In all the proglacial system communities, these two phyla showed opposite trends in their abundance, especially in the later stages of the succession (Figure 4.8A). The linear model showed a significant negative correlation ($R^2 = 0.62$) between Proteobacteria and Actinobacteria (Figure 4.8B). The highest correlation being between Alphaproteobacteria and Actinobacteria at class-level (Table A4.3). The phyla Acidobacteria, *Candidatus* Rokubacteria, Elusimicrobia, Nitrospirae and Planctomycetes showed a significant positive Spearman's rank correlation coefficient (r_s) in at least one of the three forefield systems, showing an increased in the phyla relative abundance going further away from the glacier toe. Whereas Ascomycota, candidate division AD3 and Firmicute showed both positive and negative r_s in different proglacial systems. The phyla Armatimonadetes, Bacteroidetes, Cyanobacteria, Firmicutes showed a high abundance in the ice samples and the sites proximal the ice edge and then decreased in more distant soil (Figure 4.8C).

Table 4.1: Permutational multivariate analysis of variance (PERMANOVA) performed between the distance from the ice edge and the diversity index dataset (A), taxonomy dataset at the phylum-, order- and genus- level (B), gene dataset (C) and the GO (Gene Ontology) dataset (D). Mantel test performed to calculate the correlation between the distance from the ice edge and the geochemical dataset (i.e. TN + TOC) with the diversity index dataset (E), taxonomy dataset at the genus-level (F), gene dataset (G) and the GO (Gene Ontology) dataset (H). Each of these four datasets were tested with all the samples from the three different proglacial systems (G + SV + SW), only the G system, only the SV system and only the SW system. The symbol '-' is reported for Mantel test statistics were p-value ≥ 0.05.

A

Diversity indices	
R^2	**p-value**
0.20	0.00

B

Taxonomy		
Rank	**R^2**	**p-value**
phylum	0.13	0.00
order	0.22	0.00
genus	0.23	0.00

C

Gene	
R^2	**p-value**
0.26	0.00

D

GO categories	
R^2	**p-value**
0.10	0.00

E

Diversity indices				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	-	-	-	-
G	-	-	-	-
SV	0.36	0.00	-	-
SW	-	-	-	-

F

Taxonomy				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	-	-	-	-
G	0.19	0.04	0.42	0.00
SV	0.61	0.00	-	-
SW	0.24	0.01	-	-

G

Gene				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	0.20	0.00	-	-
G	-	-	0.29	0.01
SV	0.47	0.00	-	-
SW	0.37	0.00	-	-

H

GO categories				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	-	-	-	-
G	0.50	0.00	-	-
SV	0.31	0.00	-	-
SW	-	-	-	-

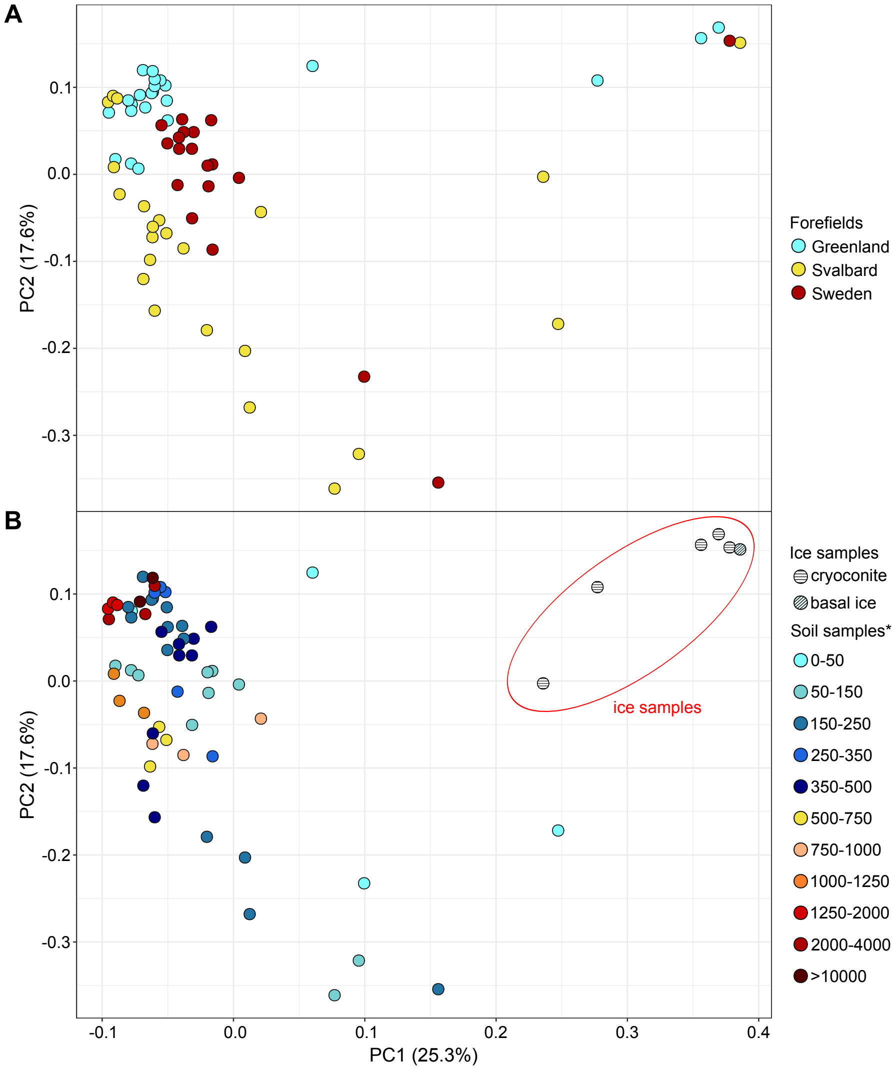


Figure 4.7: Principal component analysis (PCA) showing the sample distribution based on the genus taxonomic dataset. Samples are colored based on (A) which forefield and (B) which soil succession stage they were collected from. *soil samples are categorized in relation to their distance from the ice edge (m).

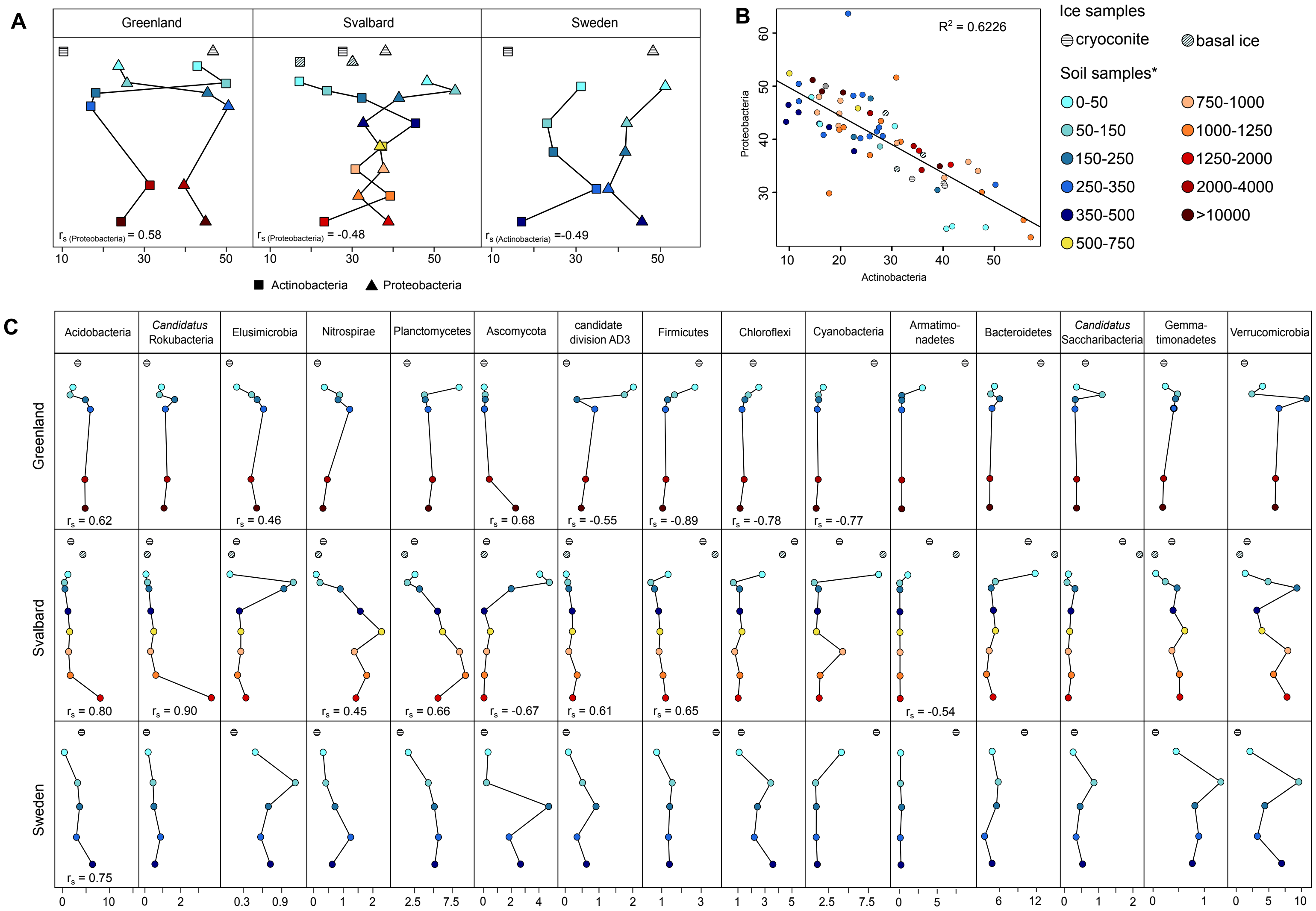


Figure 4.8: Taxonomy classification at the phylum-level. (A) Actinobacteria and Proteobacteria trends in the dataset. (B) Linear model relation between Actinobacteria and Proteobacteria relative abundances (p-value = 3 x 10^{-15}). (C) Remaining phylum trends in the datasets. Spearman's rank correlation coefficient (r_s) was calculated between the sample distance and the relative abundance data for each phylum. r_s is reported below the appropriate line plot only when significant (p-value < 0.05). *soil samples are categorized in relation to their distance from the ice edge (m).

4.3.4 Gene trends

The PERMANOVA performed on the gene dataset (Table 4.1C) explained the 26% of the observed variance of the samples across different forefields (p-value < 0.05) but only 10% of the variance in the Gene Ontology (GO) dataset (Table 4.1D). Both values decreased when I looked at the datasets without ice (Table A4.2C-D). The Mantel test statistic performed between the distance from the ice edge and the entire dataset showed a correlation of 0.20. The gene dataset was then significant (p-value < 0.05) for SV and SW when correlated to the distance from the ice edge, with r equal to 0.47 and 0.37, respectively. In Greenland the geochemical datasets (TN and TOC) correlated with an r of 0.29 (Table 4.1G).

The 2422 GO categories found in the proglacial dataset were then checked for statistical correlations with the distances from the ice edge, TN and TOC. Out of a total of more than 2400 GO categories, 431 and 110 were positively and negatively correlated to distance, 233 and 374 to TN and 202 and 328 to TOC. In Appendix 4.D, I report all the GO classes that showed a significant correlation (p-value < 0.05) higher than 0.4 or lower than -0.4. For example, the distance from the ice edge had a significant positive correlation with oxalate metabolic processes ($r = 0.61$) which I will investigate further in next sections. Also, genes involved in starvation responses (cellular response to amino acid starvation), or RNA and DNA repair decreased with the increase of the soil distance from the ice front. Furthermore, more genes indicating photosynthetic metabolism were present in early stages of the succession. Finally, distance and both the geochemical (i.e. TN and TOC) positively correlated with the distribution of genes involved to the response to drug and antimicrobial compounds.

4.3.4.1 Nitrogen fixation

No common and clear trends of the nitrogenase gene coverages were observed across the three forefields, where the coverage values peaked at different forefield stages. In the G forefield the highest coverage of nitrogenase genes was observed at 150-250 m distance, SV showed higher coverage values in the medium soil stages (i.e. 250-750 m) and SW showed a gradual coverage increase with the site distance from the ice edge. The number of genera associated to nitrogenase genes followed the coverage trends (Figure 4.9A). Whereas Spearman's correlation coefficient for gene coverage versus distance was not significant for any of the datasets, it was significant (p-value < 0.05) for the G, SV and SW datasets when performed against the TN data (r equals to 0.50, 0.30 and 0.54, respectively).

Thirty genera were found to have at least one assigned region of the assembly with a nitrogenase coding region. The distance where the majority of taxa had a peak in the nitrogenase coverage

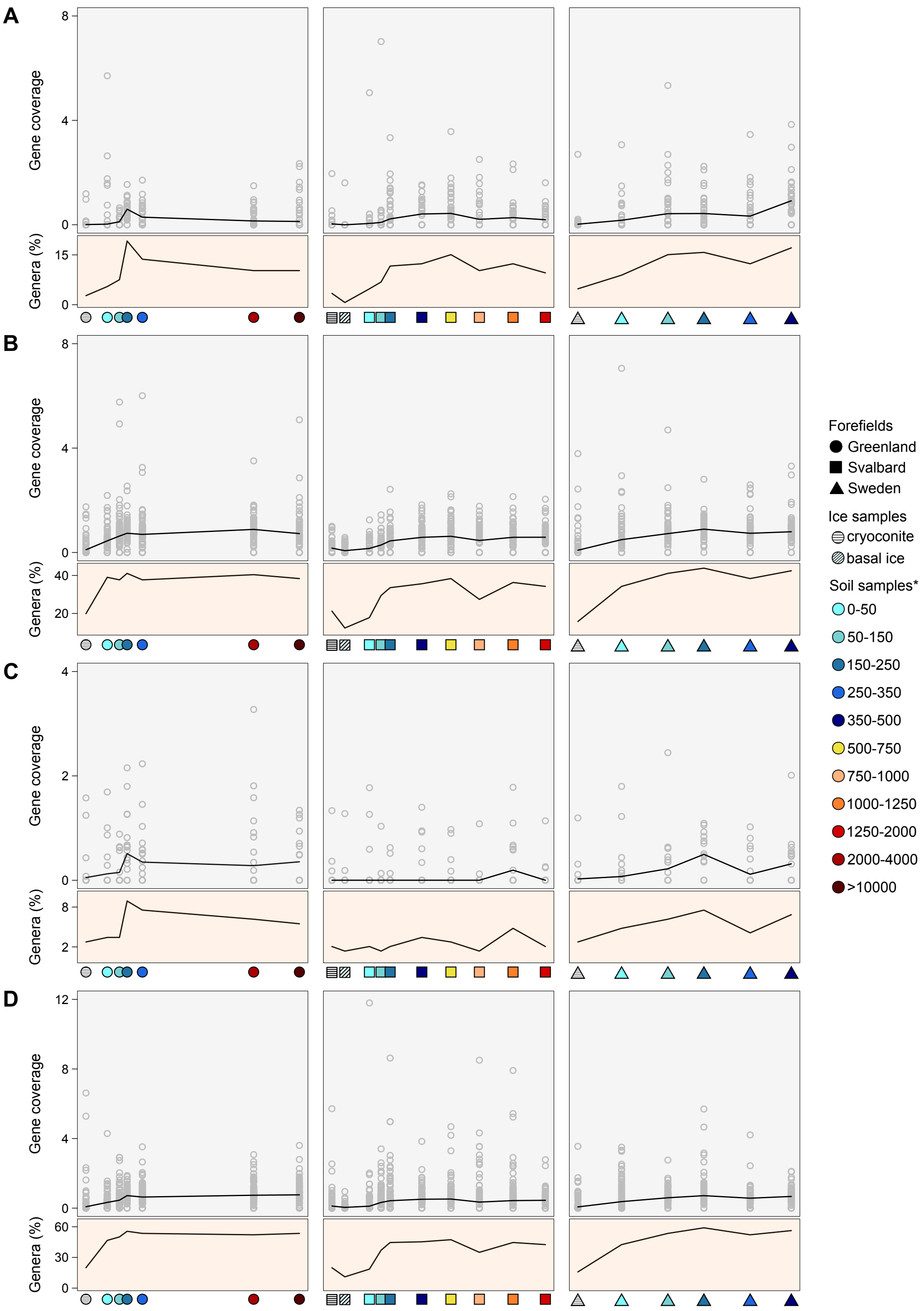


Figure 4.9: Gene coverage and number of genera trends along the microbial succession for (A) nitrogenase genes, (B) obcA genes which are involved in the oxalate biosynthesis, (C) cyanide synthase genes and (D) siderophore-related genes. The number of genera is represented as the percentage calculated as the number of genera that possess the studied gene divided by the total number of genera at a certain distance. *soil samples are categorized in relation to their distance from the ice edge (m).

was 150-250 meters in the G succession, 500–750 meters for the SV and 50-150 meters in the SW datasets.

Geobacter, *Bradyrhizobium*, *Nostoc* and *Paraburkholderia* had the highest number of genes associated with nitrogenase activity, 156, 101, 53 and 49, respectively (Figure 4.10). *Nostoc* was the most abundant genus in the ice samples and in the samples closer to the glacier edge whereas the other three genera showed similar coverage across the soil succession. *Frankia* had 17 contig regions associated to nitrogenase, it was not present in the ice samples and showed the highest coverage in the early and medium soil stages of the SV and SW soil successions. Other taxa that presented high coverages of the nitrogenase genes at early and medium stages of the microbial succession were *Phycicoccus*, *Variovorax*, *Capsulimonas* and *Paenibacillus*. *Corallococcus*, *Micromonospora* and *Pseudomonas* showed a higher abundant in the later stages of the SW succession.

Two hundred and two assembled coding regions associated with nitrogenases were also associated with unclassified contigs at the genus-level. Whereas 50% of the genes were assigned to unknown organisms at phylum-level, 40% of these nitrogenase coding regions were assigned to Proteobacteria and another 5% to Verrucomicrobia (Figure A4.4A).

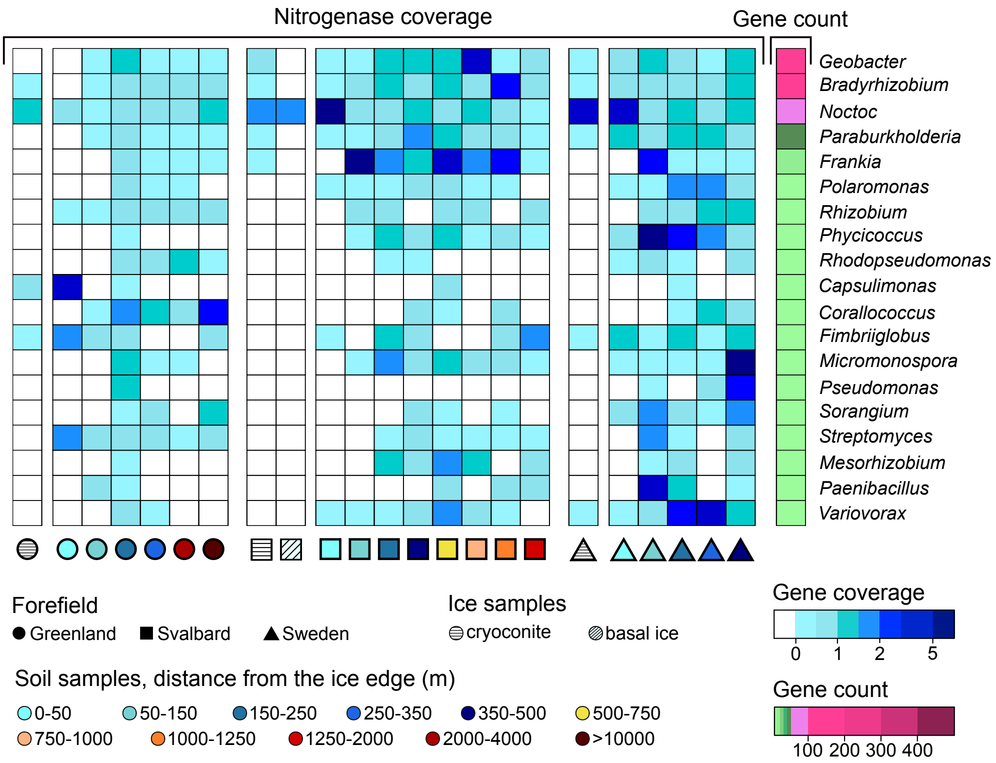


Figure 4.10: Nitrogenase gene trends at genus-level along the microbial succession. Only genera with more than 1 coding region associated to a nitrogenase are reported.

4.3.4.2 Rock weathering

Genes related to rock weathering processes (i.e. obcA, cyanide synthase and genes involved in siderophore synthesis and transport) showed a general lower coverage in the ice and early soil samples and an increase with the microbial succession. As for the number of genera related to the nitrogenase genes, the number of genera containing these genes followed the coverage trends, increasing with the distance from the ice edge (Figure 4.9B-D).

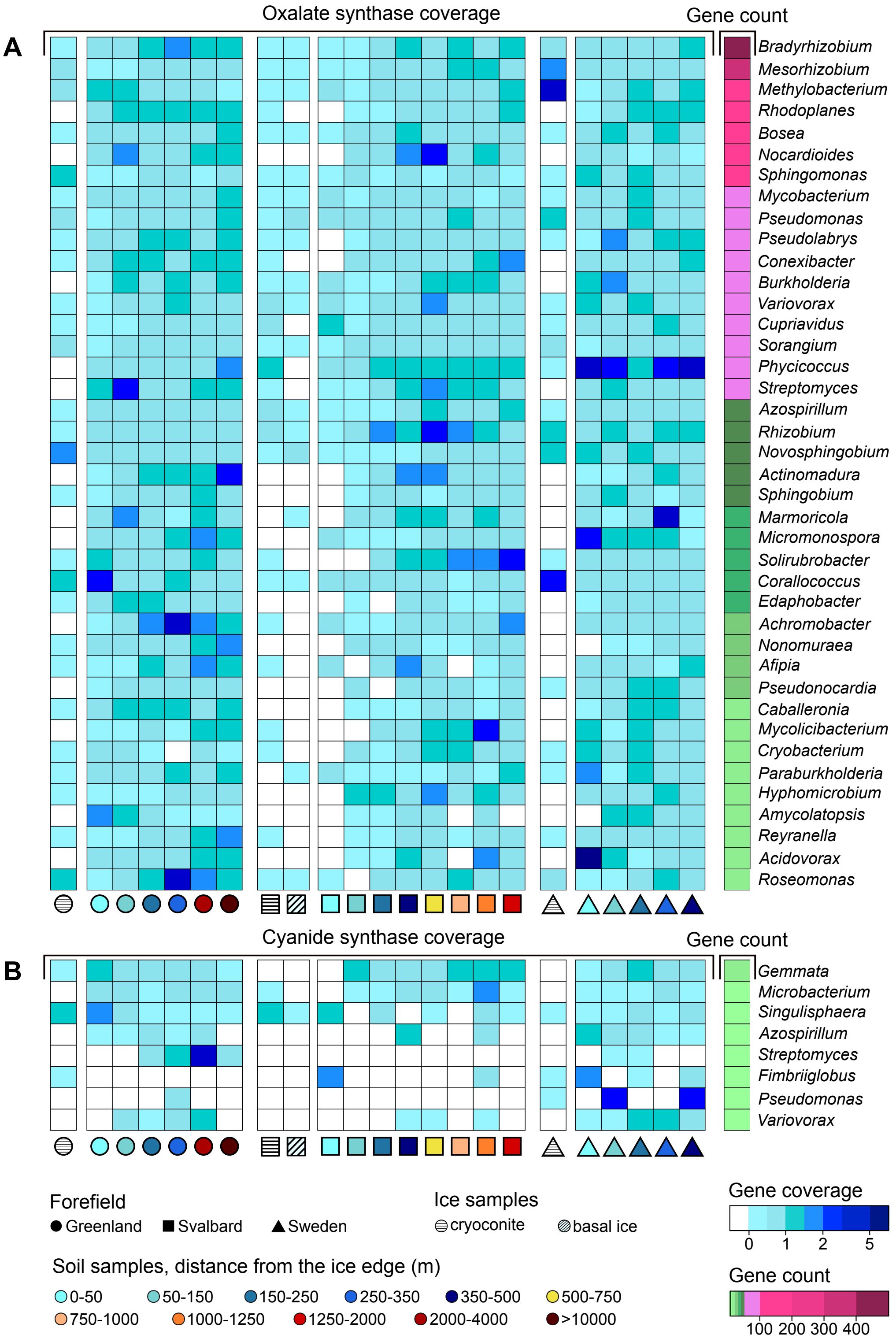
When looking at the gene distribution at the genus level (Figure 4.11) I did not observe a common trend of gene distribution. For all the three rock weathering genes, different taxa showed different trends in the same forefield and the same taxon showed different trends in different forefields.

Sixty-five genera had at least one obcA gene assigned to their contigs. The genera *Bradyrhizobium*, *Mesorhizobium*, *Methylobacterium*, *Rhodoplanes*, *Bosea*, *Nocardiales* and *Sphingomonas* had the highest content with 456, 320, 157, 127, 117, 117 and 105 genes respectively (Figure 4.11A). Some of the other genera showed fewer associated contig areas (i.e. fewer genes), but a high coverage of this gene involved in the oxalate biosynthesis (i.e. obcA) at different soil stages.

Cyanide synthase genes were less abundant in the three datasets and only 16 genera were associated to these. The genus *Gemmata* had most of the genes with 18, followed by *Microbacterium* (7) and *Singulisphaera* (5). These genera where also the most ubiquitous in the successions compared to the other cyanide synthase-related genera (Figure 4.11B). The three genera that, in particular, showed a higher base coverage in some sites of the microbial succession stages were *Streptomyces*, *Pseudomonas* and *Variovorax.*

More genera were associated with genes involved in siderophore synthesis and transport (81). *Bradyrhizobium*, *Sphingomonas*, *Streptomyces* and *Variovorax* had most of the genes and, they had overall a higher abundance in the early stages of the SV and SW successions (Figure 4.11C).

The obcA gene was also associated with 2234 unclassified contigs at genus level, the cyanide synthase gene to 49 and the siderophore genes to 1919 contig regions. These unclassified contigs mainly belonged to the phyla Proteobacteria, Acidobacteria, Actinobacteria and Planctomycetes (Figure A4.4B-D).


A
Oxalate synthase coverage
Gene count
Bradyrhizobium
Mesorhizobium
Methylobacterium
Rhodoplanes
Bosea
Nocardioides
Sphingomonas
Mycobacterium
Pseudomonas
Pseudolabrys
Conexibacter
Burkholderia
Variovorax
Cupriavidus
Sorangium
Phycicoccus
Streptomyces
Azospirillum
Rhizobium
Novosphingobium
Actinomadura
Sphingobium
Marmoricola
Micromonospora
Solirubrobacter
Corallococcus
Edaphobacter
Achromobacter
Nonomuraea
Afipia
Pseudonocardia
Caballeronia
Mycolicibacterium
Cryobacterium
Paraburkholderia
Hyphomicrobium
Amycolatopsis
Reyranella
Acidovorax
Roseomonas
B
Cyanide synthase coverage
Gene count
Gemmata
Microbacterium
Singulisphaera
Azospirillum
Streptomyces
Fimbriiglobus
Pseudomonas
Variovorax
Forefield
● Greenland
■ Svalbard
▲ Sweden
Ice samples
cryoconite
basal ice
Soil samples, distance from the ice edge (m)
0-50
50-150
150-250
250-350
350-500
500-750
750-1000
1000-1250
1250-2000
2000-4000
>10000
Gene coverage
0
1
2
5
Gene count
100
200
300
400

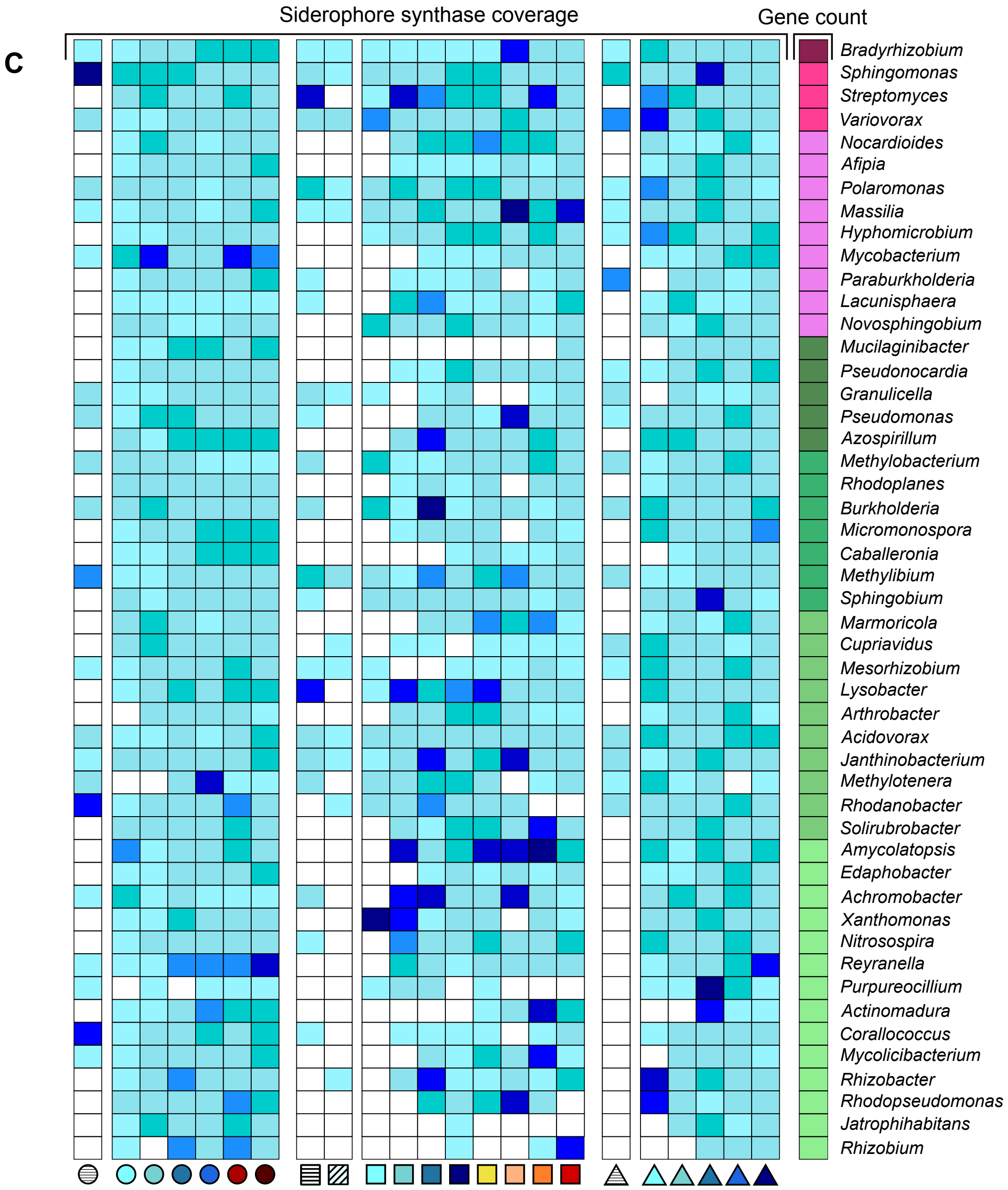


Figure 4.11: Rock weathering gene trends at genus-level along the microbial succession for (A) obcA genes which are involved in the oxalate biosynthesis, (B) cyanide synthase genes and (C) siderophore-related genes. Only genera with more than 10 coding regions associated to (A) and (C), and 1 to (B), are reported.

4.4 Discussion

Microbial successions have been widely studied with the aim to understand microbial-driven environmental processes and how communities are shaped by environmental factors (Fierer et al., 2010). Forefields constitute ideal systems to study microbial successions as they are characterized by different environmental gradients (e.g. TOC and TN increase) from the bedrocks close to the ice edge to the more developed soil. These gradients condition the microbial diversity and structure and, at the same time, are conditioned by microbial communities (Uroz et al., 2009; Uroz et al., 2015; Wang et al., 2020). The two main biological-driven processes, that create these gradients, are nitrogen fixation and rock weathering. Thanks to the development of LongMeta, a new pipeline to analyze metagenomes, I have been able to investigate microbial successions from three different proglacial systems with the aim of exploring forefield microbial trends, investigating nitrogen fixing and rock weathering processes, and exploring the taxonomy associated with these processes.

Along the microbial successions, microbial communities increased in diversity between ice and soil samples and also with the soil complexity, where the biggest increase was at the early soil stages of the succession (Figure 4.6). This was previously observed in Jiang et al. (2018) where bacterial complexity was higher in later soil stages compared to early stages. The first succession stages are characterized by challenging environmental conditions and are deeply influenced by the glacier environment (characterized by a lower microbial diversity). For these reasons, this soil has a lower community diversity compared to the more developed soil where there is the formation of differentiated ecological niches due to the nutrient increase and plant development (Dumbrell et al., 2010; Reynolds et al., 2003).

In all three proglacial systems, the ice microbial communities comprised organisms typically found in the cryosphere (Figure 4.8C). Cyanobacteria are widely found in the glacial environment thanks to their ability to use sunlight as energy source (Segawa et al., 2017). Firmicutes are spore-forming and therefore adapted to the challenging glacier conditions (Galperin, 2016). Organisms belonging to the Bacteroidetes phylum have been found in many glacial habitats (e.g. Smith et al., 2016; Wilhelm et al., 2013; Zhang et al., 2009). The diffused presence of these nonspore-forming heterophic organisms has been proposed to related to their ability to assimilate and use recalcitrant substances (Zeng et al., 2013) abundant in glacier environment (Kmezik et al., 2020). Organisms belonging to Armatimonadetes have been found in this environment (e.g. Zhang et al., 2015; Bajerski and Wagner, 2013; Gokul et al., 2016). This phylum is not well characterized but its few isolates have been characterized as chemoheterotrophs (Tamaki et al., 2011; Lee et al., 2014). Cyanobacteria, Firmicutes, Bacteroidetes and Armatimonadetes organisms also showed a high abundance in the first stages of the succession where the soil is highly affected by glacial water discharges (Boyd et al., 2014; Dieser et al., 2014; Bradley et al., 2014) and environmental conditions favor the development of communities adapted to nutrient depleted conditions.

Whereas, in the first soil stages of the succession, microbial diversity showed a presence of autotrophs (i.e. Cyanobacteria), spore-forming organisms (i.e. Firmicutes) and organisms specialized to use recalcitrant compounds (i.e. Bacteroidetes), heterotrophic trends were shown to increase in later stages of the microbial succession. In the three proglacial systems, the microbial succession was characterized by the increase of Acidobacteria (Figure 4.8C). Organisms belonging to this phylum are heterotrophs and adapted to use a variety of organic carbon sources. Furthermore, many of these organisms are acidophiles (Kielak et al., 2016) and therefore more adapted to live in the later stages of the succession where the soil is enriched of organic carbon and pH is lower (Zumsteg et al., 2012). Other phyla such as *Candidatus* Rokubacteria, Elusimicrobia, Nitrospirae and Planctomycetes showed an increased abundance in the more developed soil of the later stages. Elusimicrobia have been isolated mainly from soil and are principally insect symbionts (Méheust et al., 2019). *Candidatus* Rokubacteria and Verrucomicrobia are characterized by organisms able to perform sulfate reduction (Anantharaman et al., 2018). Nitrospirae metabolism relies on nitrite oxidation (Daims et al., 2016). These microorganisms are favored in the later stages of the succession where nitrogen and sulphur stocks have been built by diazotrophs and rock weathering organisms.

Actinobacteria and Proteobacteria are two ubiquitous phyla and are both characterized by a wide metabolic range (Barka et al., 2016; Taylor et al., 2019). Actinobacteria and Alphaproteobacteria, in particular, showed oscillating trends along the three proglacial environments (Figure 4.8A and Table A4.3). This may be due to their competition over the same energy sources (i.e. organic carbon and nitrogen). Their opposite trend is indeed less marked in samples closer to the ice edge where the organic carbon is less abundant. This trend could be also due to the different soil matrix and minerology differences that can be encountered at small-scale distances in the soil. Some studies have proposed that dominant community switches in soil could be due to these factors (Mitchell et al., 2013; Uroz et al., 2015).

Although ice-related/soil-related organism trends were observed in all the proglacial systems, the latter showed differences in relative abundances of several taxa (e.g. Nitrospirae, *Candidatus* division AD3, Chloroflexi and Gammatimonadetes) showing how local patterns and characteristics could influence the microbial distribution and the biogeography of different soil microbial communities. Environmental selection to specific conditions (e.g. DOC, TN and pH) have previously been observed to strongly shape microbial communities at large biogeographic spatial scales (Hanson et al., 2012; Malard and Pearce, 2018). Highlighting this aspect, no correlation between distance from the ice edge and the entire forefield dataset was observed, a high correlation was obtained only when looking at the microbial distribution across the each separate forefield (Table 4.1F), showing how the forefield local factors had a role in shaping the microbial communities. Compared to the taxon dataset, the variance observed in gene dataset was better explained by the forefield differentiation (factor 'forefield'; Table 4.1) where this indicates how gene and trait-based data could be used to biogeographic studies instead of the more widely used taxonomy data (Green et al., 2008).

Whereas the Svalbard and Sweden communities were mainly shaped by their distance from the ice edge, the Greenland communities were more shaped by TN and TOC gradients. This could be due by the nature of its complex dynamics where the Greenland ice sheet does not retreat linearly and it retreats slower compared to the other systems, leading to more complex dynamics for the microbial development (Nienow et al., 2017). The same result was observed when looking at the functional dataset (Table 4.1G). However, this dataset (not divided into different forefields) also showed a significant correlation with the distance from the ice edge, suggesting that, in this case, certain functional profiles such as photosynthesis, response to starvation or to antimicrobial compounds were more conserved across the different forefield gradients (Table A4.4).

In this chapter I have investigated the taxonomy of the organisms performing two important microbially-driven environmental processes: nitrogen fixation and rock weathering. To do so, I focused on 4 different set of genes: i) nitrogenase genes, ii) obcA genes involved in the first step of the oxalate biosynthesis, iii) genes involved in cyanide synthesis and iv) genes involved in siderophore synthesis and transport. Whereas the first one is involved in the nitrogen fixation, the other three are involved in rock weathering processes where oxalate and cyanide have been observed to increase rock and mineral dissolution and siderophore are molecules used to facilitate the iron uptake inside the cells.

These processes have previously been shown to be prevalent closer to the ice edge where the microorganisms are able to produce metabolic energy from atmospheric nitrogen and inorganic compounds (e.g. minerals) are favored (Brankatschk et al., 2011; Fernández-Martínez et al., 2016). In all the three systems, genes associated with these processes were shared by a lowest percentage of taxa in the early stages and by more in later succession stages (Figure 4.9). These trends did not show linear increase along the proglacial successions but, however, they showed a general increase from the ice samples to the more developed soil. Interestingly, nitrogenase genes showed a correlation with TN concentration values, highlighting the pivotal role of diazotrophs in the creation of nitrogen pool in proglacial systems (Nash et al., 2018).

Gene trends observed at specific genus level showed a variety of trends and distributions suggesting that different organisms drive the same ecological process (i.e. nitrogen fixation and rock weathering) but at different stages of the microbial succession.

Nitrogenases are key enzymes in the fixation of atmospheric nitrogen to ammonium. These enzymes are coded by a wide gene class (e.g. nifK, nifH and nifD genes) and have been widely used to detect biological nitrogen fixation potential in a variety of environmental settings (Hoffman et al., 2014) and diazotrophic organisms have been found spanning a wide range of different bacterial phyla (Addo and Dos Santos, 2020).

Most of the 30 genera associated with nitrogenase genes in our dataset belonged to the phylum Proteobacteria (17) and Actinobacteria (6). The only cyanobacterial organism to which nitrogenase was assigned in our dataset was *Nostoc* which also showed a higher abundance in the ice realm compared to the proglacial soil and has previously been detected by Fernandez et al. in the Arctic region (2020). The other cyanobacterial organisms present in our dataset were present in a lower percentage (1%) compared to *Nostoc* (4%); between them there are, for example, *Phormidesmis* and *Chamaesiphon.* These organisms are typical of the cold environments and do not perform nitrogen fixation (Jungblut and Vincent, 2017; Uetake et al., 2019; Segawa et al., 2017).

In the soil, diazotrophic organisms belonging to Proteobacteria and Actinobacteria were more abundant (Figure 4.10). These phyla are usually associated with soil environment and root symbiosis (Gtari et al., 2012; Rudnick et al., 1997). Other than *Nostoc*, also *Bradyrhizobium*, *Geobacter*, *Paraburkholderia* and *Frankia*, known nitrogen fixing organisms (Calderoli et al., 2017; Choi and Im, 2018; Hara et al., 2019; Sellstedt and Richau, 2013; Thangaraj et al., 2017) were abundant in all our studied systems (Figure 4.10).

Streptomyces has only recently been shown to perform nitrogen fixation by isolation and sequencing from environmental soil (Dahal et al., 2017). This genus was also associated with nitrogenase in our study. This result represents the first confirmation of the presence of nitrogenase genes in *Streptomyces.* Furthermore, the genera *Cellulomonas*, *Phycicoccus*, *Fimbriiglobus* and *Variovorax* were also associated to these genes but have never showed this characteristic before. The distribution of *Cellulomonas* and *Phycicoccus* also showed to be conditioned by low TN concentration suggesting they may contribute to the build-up of nitrogen stocks in the early stages of the succession (Figure A4.3). Genomes belonging to these genera and present in the comprehensive NCBI online databases do not present any nitrogenase coding region in their sequence. This discrepancy between the nitrogenase absence in the *Cellulomonas*, *Phycicoccus*, *Fimbriiglobus* and *Variovorax* genomes and the fact that they did show nitrogenase genes in our dataset may have different origins. It could be that these genes were incorrectly assigned in our dataset, due to the possible presence of chimeric contigs. Or, it could be due to gene lateral transfer from diazotrophic organisms, phenomena that is common in environmental samples (Hall et al., 2017; Gillings, 2017). Also, it could just be that no diazotrophic organisms belonging to these genera have ever been sequenced, but, as in the *Streptomyces* genus, some organisms belonging to the same genus are diazotrophs and some are not.

Compared to the biological nitrogen fixation, less is known about biological rock weathering. Whereas some microbial processes involved in the rock weathering have already been identified, the protagonists of these processes still remain unclear (Samuels et al., 2020). In this chapter, I explored the taxonomy associated to three different rock weathering associated genes: the obcA gene, cyanide synthase genes and genes involved in siderophore transport and synthesis (Figure 4.11). Whereas oxalate, cyanide and siderophore syntheses have been shown to be correlated with rock weathering (Ferreira et al., 2019; Frey et al., 2010; Welch et al., 1999), these three molecules are involved in other processes. For instance, the release of oxalate has been shown to play an important role in pathogenicity and metal tolerance (Palmieri et al., 2019) and cyanide release plays a role in microbial competition (Blumer and Haas, 2000). And finally, siderophores are molecules commonly used by microorganisms for the uptake of environmental iron but also to sequester iron from hosts (Braun and Hantke, 2011; Krewulak and Vogel, 2008). Even if all of these processes are present in organisms not involved in rock weathering, we expect a higher abundance of these organisms in early succession samples where the rock weathering is predominant.

The genus that showed most of the genes involved in oxalate and siderophore biosynthesis was *Bradyrhizobium.* Organisms belonging to this genus are mainly recognized as plant symbionts but can also be present as free-living organisms (Čuklina et al., 2016; Kulkarni et al., 2015). The high content and diversity of genes involved in siderophore metabolism relates to the fact that high production of siderophore is needed to uptake iron released from the rock dissolution performed by oxalate release. These organisms are also diazotrophs and require iron as nitrogenase cofactor (Rubio and Ludden, 2008). *Mesorhizobium* and *Paraburkholderia* are also common diazotrophic component of the rhizosphere but have been detected as free-living organisms (Ahmad et al., 2008) and they showed high presence of both these genes. *Paraburkholderia* was found to have chemotactic sensitivity for oxalate (Haq et al., 2018) even if it was not directly showed that it can produce it. *Micromonospora* also had high oxalate-related gene abundance especially in the first stages of the SW dataset. Organisms related to this genus have been found in soil by different studies (Malisorn et al., 2020; Thawai et al., 2016). And have previously been found to increase phosphate and iron solubilization in soil thanks to organic acid release, such as oxalate (El-Tarabily et al., 2008). Overall *Burkholderia* had a higher presence of both siderophore and oxalate-related genes in early stages. These organisms have been identified by many different studies as oxalate producers (Nakata, 2011; Oh et al., 2014) and have been often found in proglacial soil (Frey et al., 2010). *Pseudomonas* was shown to produce oxalic acids (Hamel et al., 1999) which concords with the trends in our datasets. Additionally, contigs associated to *Pseudomonas* were found to contain cyanide synthase coding regions. *Streptomyces* was previously identified as phosphate solubilizer, but mechanisms were not identified (Liu et al., 2016). Other than the oxalate genes, this genus was also shown to have cyanide synthase genes in our dataset.

Finally, *Fimbriiglobus*, *Corallococcus*, *Variovorax* and *Phycicoccus* are organisms typically found in soil (Singh et al., 2015; Yoon et al., 2008; Zhang et al., 2011) and were associated to oxalate, cyanide and siderophore genes for the first time in this study.

Thanks to LongMeta, I was able to explore environmental microbial successions and to confirm and find new organisms associated to specific environmental processes such as rock-weathering and nitrogenase genes. The pipeline is reproducible, easy to use and generalizable to microbial dataset exploration. However, it also presents some cons. This pipeline relies on known sequence databases and its detection level is limited to known organisms and genes, even if, we saw how this approach can still allow the detection of known genes in organisms where that specific genes have never been observed before. Furthermore, the pipeline relies on an assembly approach and is therefore skewed towards organisms that are more abundant in the dataset and that assembled better. The pipeline was designed with low-coverage environmental datasets in mind and it is not recommended for high-coverage sequencing datasets where a more focused metagenome-assembled genome (MAGs) approach is advisable (Chapter 5).

4.5 Conclusion

Microbial successions in the three proglacial systems showed early soil stages enriched with autotrophic, spore-forming and recalcitrant compound degraders showing a community influenced from the common ice microbiome, whereas later soil stages showed a higher heterotrophic microbial component. Although there were common taxonomic trends among the proglacial systems, taxa contribution to the different proglacial microbial communities showed differences suggesting the presence of biogeographic differences between proglacial systems. Further, rock weathering and nitrogenase gene distributions peaked at different succession soil stages in different proglacial systems. Different genera showed differential gene coverage at different stages of the microbial successions, indicating a community constituted by a plurality of organisms involved in these processes but where the latter were driven by different organisms in different soil succession stages. Whereas I confirmed the presence of nitrogenase and rock weathering genes in known nitrogen fixing and rock weathering organisms, in this study I also present organisms that, even if often found in soil and proglacial systems, have never been related to these processes before. The involvement of these organisms in nitrogen fixation and rock weathering processes should be confirmed with a deeper sequencing dataset and a MAG approach.

These analyses were performed with the LongMeta pipeline, a new reproducible pipeline to obtain taxon*gene information from shotgun metagenomic information and publicly available at https://github.com/gvMicroarctic/LongMeta.

Acknowledgements

I thank Gary Barker, Alex Anesio, Liane Benning, Maisie Nash, Marian Yallop, Katherine Wright, Joshua Blacker and James Bradley for conducting the sample collection. I also thank Jane Coghill and Christy Waterfall at the Bristol Genomics Facility for Illumina DNA sequencing.

A modified version of this chapter is in preparation to be submitted to a peer-reviewed journal: 'Varliero, G., Nash, M., Anesio, A. M., Barker, G. A. L., A taxon-wise insight into rock weathering and nitrogen fixation functional profiles of proglacial systems'.

Appendix

4.A LongMeta algorithms

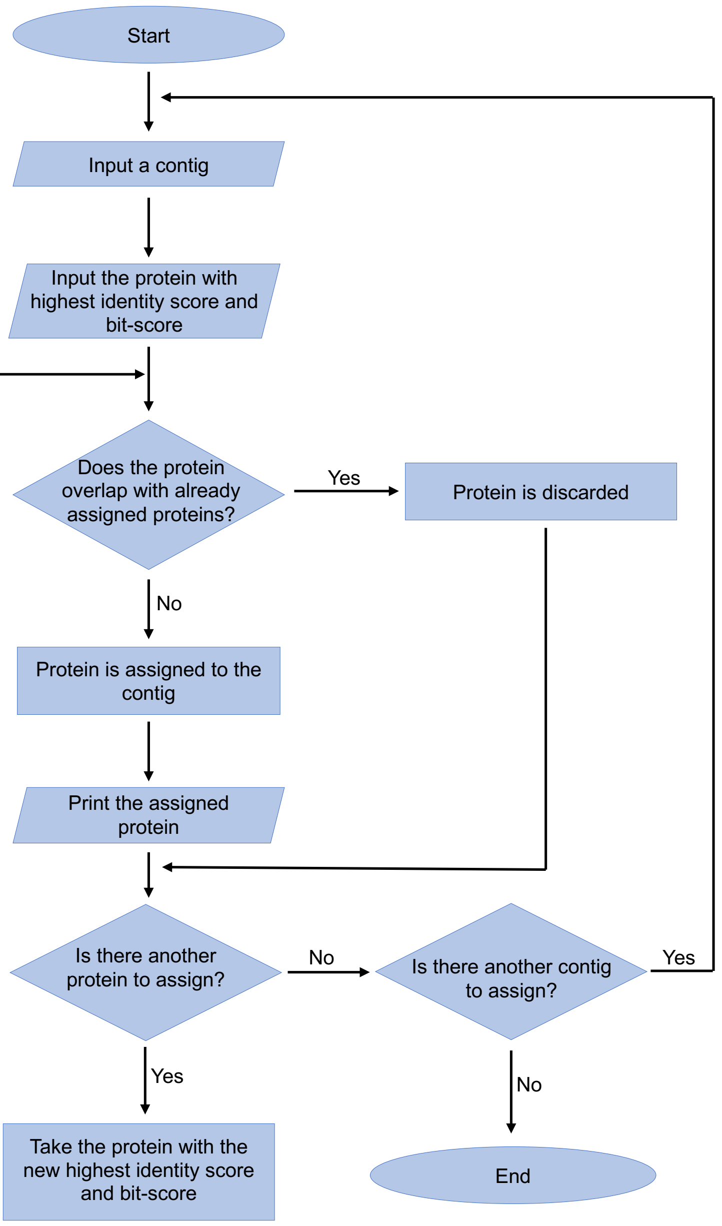
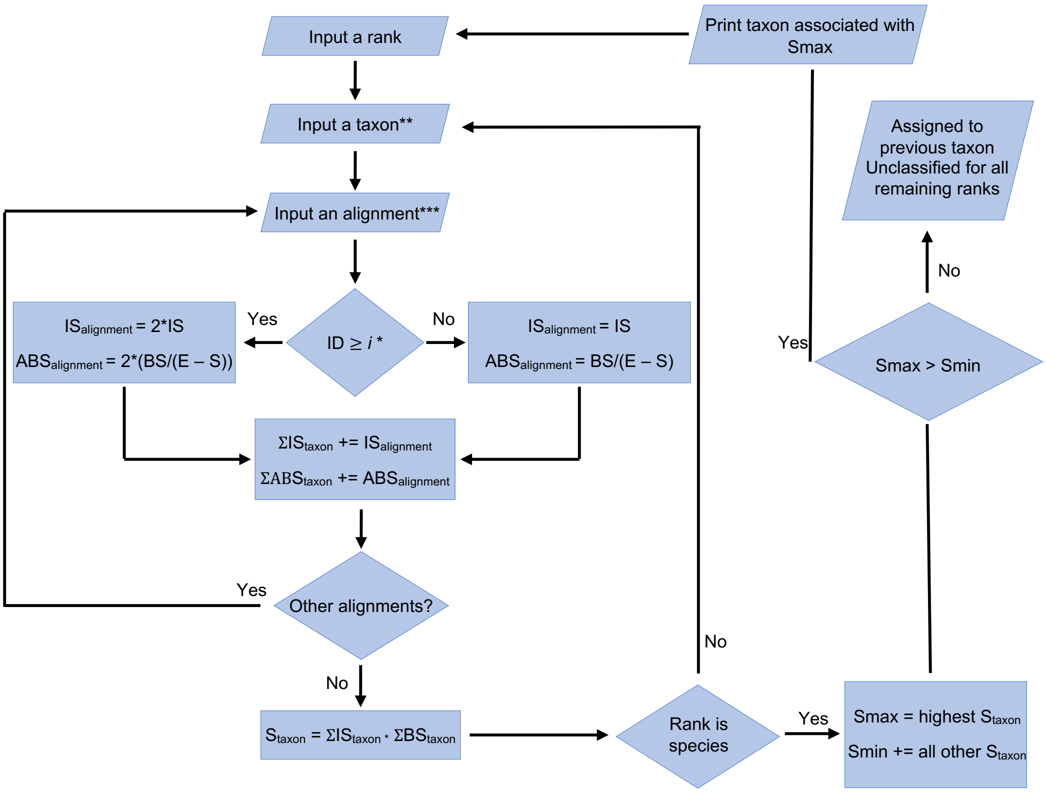

Start
Input a contig
Input the protein with highest identity score and bit-score
Does the protein overlap with already assigned proteins?
Yes
Protein is discarded
No
Protein is assigned to the contig
Print the assigned protein
Is there another protein to assign?
No
Is there another contig to assign?
Yes
Yes
Take the protein with the new highest identity score and bit-score
No
End

Figure A4.1: Flow chart of gene assignment algorithm.


Input a rank
Print taxon associated with Smax
Input a taxon**
Assigned to previous taxon Unclassified for all remaining ranks
Input an alignment***
No
$IS_{alignment} = 2*IS$
$ABS_{alignment} = 2*(BS/(E – S))$
Yes
ID ≥ *i* *
No
$IS_{alignment} = IS$
$ABS_{alignment} = BS/(E – S)$
Yes
Smax > Smin
ΣIS_{taxon} += $IS_{alignment}$
ΣABS_{taxon} += $ABS_{alignment}$
Yes
Other alignments?
No
No
$S_{taxon} = \Sigma IS_{taxon} \cdot \Sigma BS_{taxon}$
Rank is species
Yes
Smax = highest S_{taxon}
Smin += all other S_{taxon}

**i* is the identity score value (default: 80).
** the ranks are domain, phylum, class, order, family, genus, species.
*** the algorithm looks only into the alignments whose taxonomy is ‘inside’ the previously assigned rank.

Figure A4.2: Flow chart of Lowest Common Ancestor (LCA) algorithm. IS: identity score; BS: bit-score; ABS: averaged bit-score; E: end of the protein alignment; S: start of the protein alignment.

4.B Sample characteristics

Table A4.1: Read content per sample. All the counts are reported as the sum of both forward and reverse reads. Reads from Greenland (G) and Swedish (SW) datasets were 150bp long, whereas those from Svalbard (SV) dataset were 100 bp long.

Sample	Forefield	Distance	Raw reads	Trimmed reads	Trimmed reads (%)	Total bases	Mapped reads to assembly (%)
G-1	G	cryoconite	74,992,750	62,281,172	83	7,187,824,916	94
G-2	G	cryoconite	74,220,220	62,697,712	84	7,395,859,056	94
G-3	G	cryoconite	68,161,082	56,551,528	83	6,528,966,072	93
G-4	G	40	67,748,804	58,269,730	86	6,964,519,188	69
G-5	G	40	77,474,456	66,716,324	86	7,985,703,488	68
G-6	G	40	61,199,512	50,778,584	83	5,731,546,733	62
G-7	G	120	75,975,046	64,894,402	85	7,687,561,070	72
G-8	G	120	91,844,214	80,750,866	88	9,918,813,542	78
G-9	G	120	77,023,868	66,570,850	86	8,011,592,250	71
G-10	G	160	139,381,778	119,421,182	86	14,179,337,314	59
G-11	G	160	91,962,344	77,483,224	84	9,026,364,457	56
G-12	G	160	81,681,122	70,366,000	86	8,417,841,387	58
G-13	G	180	105,431,412	89,380,892	85	10,505,678,835	65
G-14	G	180	63,240,678	53,430,508	84	6,276,219,348	68
G-15	G	180	72,243,790	60,669,410	84	7,127,227,196	63
G-16	G	270	82,137,112	71,117,502	87	8,597,039,587	64
G-17	G	270	107,570,706	90,819,662	84	10,578,631,872	65
G-18	G	270	78,274,268	66,493,958	85	7,850,171,960	65
G-19	G	3800	75,392,190	63,919,782	85	7,552,035,227	57
G-20	G	3800	65,582,044	55,780,594	85	6,587,075,487	50
G-21	G	3800	91,522,052	76,517,016	84	8,858,708,214	64
G-22	G	10000	96,384,726	81,771,096	85	9,655,404,024	62
G-23	G	10000	80,981,820	68,533,676	85	7,984,763,989	59
SV-1	SV	cryoconite	22,947,810	22,269,794	97	2,215,189,703	75
SV-2	SV	basal ice	28,894,400	28,395,640	98	2,817,953,846	93
SV-3	SV	40	18,800,440	17,537,922	93	1,742,261,993	64
SV-4	SV	130	28,886,708	28,303,382	98	2,813,720,107	52
SV-5	SV	130	18,972,122	16,906,176	89	1,676,259,798	56
SV-6	SV	200	27,632,502	25,897,482	94	2,576,342,099	46
SV-7	SV	200	23,060,890	20,724,024	90	2,059,258,461	44
SV-8	SV	200	22,002,334	21,530,160	98	2,137,672,181	48
SV-9	SV	445	23,356,976	22,759,102	97	2,260,867,633	48
SV-10	SV	445	20,355,874	20,087,560	99	1,996,581,694	52
SV-11	SV	445	23,227,446	22,764,896	98	2,260,324,489	49
SV-12	SV	670	59,467,440	58,657,474	99	5,826,895,594	48
SV-13	SV	670	57,473,614	56,788,254	99	5,640,225,267	47
SV-14	SV	670	38,289,390	38,032,472	99	3,786,740,956	47
SV-15	SV	890	4,048,326	3,967,012	98	394,610,874	32
SV-16	SV	890	22,892,266	22,324,698	98	2,213,591,398	40
SV-17	SV	890	8,664,656	6,543,310	76	643,999,833	34
SV-18	SV	1150	22,630,114	22,343,416	99	2,217,910,996	45
SV-19	SV	1150	39,315,396	38,856,178	99	3,861,257,149	42
SV-20	SV	1150	69,568,490	68,761,830	99	6,822,952,060	44
SV-21	SV	1650	24,043,000	23,513,418	98	2,335,643,616	43
SV-22	SV	1650	20,762,772	20,273,822	98	2,010,967,936	51
SV-23	SV	1650	29,280,810	28,867,114	99	2,874,021,170	47
SW-1	SW	cryoconite	62,395,094	59,053,216	95	8,220,938,496	92
SW-2	SW	40	78,397,352	74,222,178	95	10,443,960,493	79
SW-3	SW	130	66,910,678	63,704,550	95	8,951,219,174	78
SW-4	SW	130	72,683,122	69,228,276	95	9,719,225,957	77
SW-5	SW	130	60,148,730	57,046,786	95	7,892,712,713	75
SW-6	SW	130	67,844,804	64,231,264	95	8,923,085,594	72
SW-7	SW	130	63,953,088	60,944,452	95	8,542,182,607	73
SW-8	SW	200	83,010,234	79,354,112	96	11,161,621,182	73
SW-9	SW	200	74,901,072	71,515,238	95	10,030,231,547	72
SW-10	SW	200	83,405,572	79,804,612	96	11,223,988,377	70
SW-11	SW	200	68,111,048	64,777,286	95	9,100,927,086	72
SW-12	SW	290	64,225,756	61,152,466	95	8,552,527,075	66
SW-13	SW	290	64,764,076	61,854,200	96	8,671,030,908	71
SW-14	SW	370	66,675,200	63,459,146	95	8,874,543,259	65
SW-16	SW	370	66,848,090	63,779,024	95	8,945,037,456	71
SW-16	SW	370	72,051,286	68,487,298	95	9,580,949,996	76
SW-17	SW	370	71,828,498	68,123,690	95	9,552,272,522	66
SW-18	SW	370	85,214,054	81,120,886	95	11,368,376,573	71
SW-19	SW	370	71,411,294	68,058,210	95	9,559,127,942	67

4.C Taxonomy and statistical analyses

Table A4.2: Statistical analyses performed on the entire dataset (soil + ice samples). Permutational multivariate analysis of variance (PERMANOVA) performed between the distance from the ice edge and the diversity index dataset (A), taxonomy dataset at the phylum-, order- and genus- level (B), gene dataset (C) and the GO (Gene Ontology) dataset (D). Mantel test performed to calculate the correlation between the distance from the ice edge and the geochemical dataset (i.e. TN + TOC) with the diversity index dataset (E), taxonomy dataset at the phylum-, order- and genus- level (F), gene dataset (G) and the GO (Gene Ontology) dataset (H). Each of these four datasets were tested with all the samples from the three different proglacial systems (G + SV + SW), only the G system, only the SV system and only the SW system. The symbol '-' is reported for Mantel test statistics were p-value ≥ 0.05.

A

Diversity indices	
R^2	**p-value**
0.12	0.01

E

Diversity indices				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	-	-	-	-
G	0.20	0.04	-	-
SV	0.45	0.00	-	-
SW	0.18	0.03	-	-

B

Taxonomy		
Rank	**R^2**	**p-value**
phylum	0.10	0.01
order	0.16	0.00
genus	0.15	0.00

F

Taxonomy				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	-	-	-	-
G	0.26	0.02	-	-
SV	0.61	0.00	-	-
SW	0.40	0.00	-	-

C

Gene	
R^2	**p-value**
0.21	0.00

G

Gene				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	0.27	0.00	-	-
G	0.28	0.02	-	-
SV	0.53	0.00	-	-
SW	0.50	0.00	0.34	0.04

D

GO categories	
R^2	**p-value**
0.08	0.00

H

GO categories				
Forefield	**Distance**		**TN + TOC**	
	r	**p-value**	**r**	**p-value**
G + SV + SW	-	-	-	-
G	0.56	0.00	-	-
SV	0.37	0.00	-	-
SW	0.25	0.03	-	-

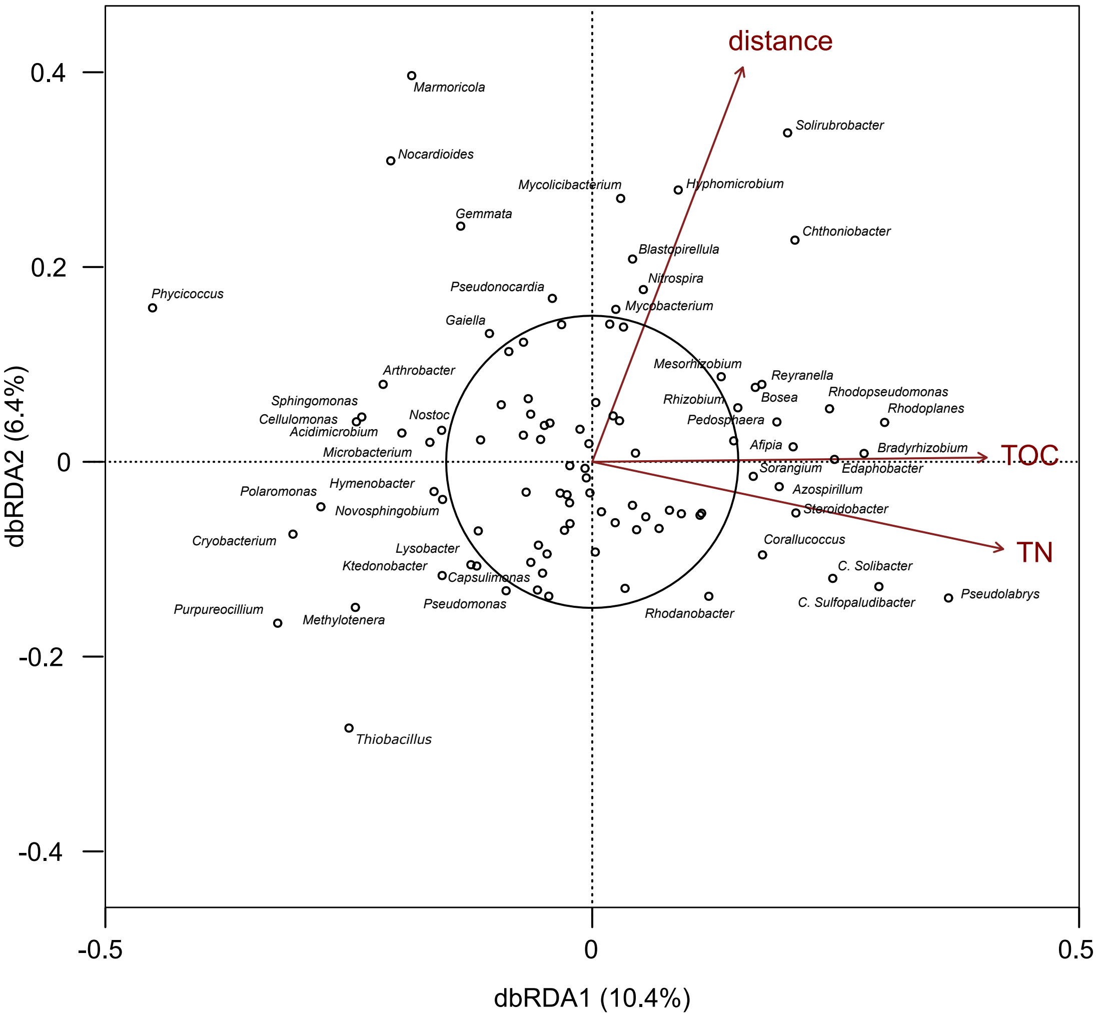


Figure A4.3: Distance-based redundancy analysis (dbRDA) bi-plot ordination performed on the Hellinger-transformed genus dataset in relation to the distance from the glacier toe, TOC and TN. Only genera that had a dbRDA1 or dbRDA2 higher than 0.15 or lower than -0.15 are displayed in the plot. Vectors indicate direction of the ice edge distance and geochemical variable effect in the bacterial community composition (Bray-Curtis similarity).

Table A4.3: Actinobacteria and Proteobacteria class abundance linear model correlation (R^2). The symbol '-' is reported when the correlation was not significant with a p-value $\geq$ 0.05. *Actinobacteria classes. **Proteobacteria classes.

	Actinobacteria*		**Acidimicrobiia***		**Rubrobacteria***		**Thermoleophilia***	
	R^2	**p-value**	**R^2**	**p-value**	**R^2**	**p-value**	**R^2**	**p-value**
Alphaproteobacteria**	0.34	0.00	0.22	0.00	0.20	0.00	-	-
Betaproteobacteria**	0.02	0.26	0.08	0.02	-	-	0.22	0.00
Gammaproteobacteria**	0.20	0.00	0.20	0.00	0.13	0.00	0.23	0.00
Deltaproteobacteria**	0.16	0.00	0.13	0.00	0.09	0.01	0.14	0.00

4.D Gene Ontology exploration

Table A4.4: Gene Ontology (GO) biological categories that significantly correlated (Mantel test statistic, p-value < 0.05) to the distances from the ice edge.

A

biological function	positive r
oxalate metabolic process	0.64
organic phosphonate catabolic process	0.64
nitrogen compound metabolic process	0.58
oligosaccharide metabolic process	0.57
chromate transport	0.57
leucine biosynthetic process	0.57
histone modification	0.56
glucose transmembrane transport	0.56
carnitine biosynthetic process	0.55
protein ubiquitination	0.54
phytochelatin biosynthetic process	0.53
copper ion homeostasis	0.51
branched-chain amino acid metabolic process	0.51
nucleoside transmembrane transport	0.51
urea metabolic process	0.50
branched-chain amino acid catabolic process	0.50
D-gluconate metabolic process	0.49
poly-hydroxybutyrate biosynthetic process	0.49
lipoprotein metabolic process	0.49
purine-containing compound salvage	0.49
xanthine metabolic process	0.49
inositol biosynthetic process	0.49
organic phosphonate metabolic process	0.48
carbohydrate transport	0.47
arabinose metabolic process	0.47
leucyl-tRNA aminoacylation	0.47
cellular aromatic compound metabolic process	0.47
arginine biosynthetic process	0.47
2-aminoethylphosphonate transport	0.47
aerobic respiration	0.47
transmembrane transport	0.46
D-xylose metabolic process	0.46
karyogamy involved in conjugation with cellular fusion	0.46
dicarboxylic acid transport	0.45
Actinobacterium-type cell wall biogenesis	0.45
CENP-A containing chromatin organization	0.45
Okazaki fragment processing involved in mitotic DNA replication	0.45
regulation of DNA double-strand break processing	0.45
regulation of histone H2B conserved C-terminal lysine ubiquitination	0.45
anaerobic ethylbenzene catabolic process	0.45
aromatic compound catabolic process	0.45
monosaccharide transmembrane transport	0.45
polyamine transport	0.45
protein-containing complex assembly	0.44
protein secretion by the type IV secretion system	0.44
positive regulation of GTPase activity	0.44
D-xylose transmembrane transport	0.44
actin filament organization	0.44
valine catabolic process	0.44
teichoic acid biosynthetic process	0.44
arginine biosynthetic process via ornithine	0.44
histone H3-K79 methylation	0.44
fungal-type cell wall beta-glucan biosynthetic process	0.44
ribosomal large subunit biogenesis	0.44
response to unfolded protein	0.44
S-adenosylmethionine biosynthetic process	0.43
intracellular signal transduction	0.43
regulation of circadian rhythm	0.43
urea catabolic process	0.43
posttranslational protein targeting to membrane, translocation	0.43
chromatin remodeling	0.43
xenobiotic transport	0.43
protein import into nucleus	0.43
glutamate biosynthetic process	0.42
glycosaminoglycan metabolic process	0.42
sodium-dependent phosphate transport	0.42
gamma-aminobutyric acid transport	0.42
establishment or maintenance of cell polarity regulating cell shape	0.42
maturation of 5.8S rRNA from tricistronic rRNA transcript	0.42
lipid transport	0.42
glyoxylate catabolic process	0.42
microtubule-based movement	0.42
molybdate ion transport	0.42
enzyme active site formation	0.42
nucleoside catabolic process	0.42
cell cycle checkpoint	0.42
response to drug	0.42
ethanolamine catabolic process	0.42
mRNA splicing, via spliceosome	0.41
polyketide metabolic process	0.41
beta-lactam antibiotic catabolic process	0.41
gluconate transmembrane transport	0.41
cell adhesion	0.41
fucose transmembrane transport	0.41
glutamine biosynthetic process	0.41
formaldehyde catabolic process	0.41
regulation of transcription by RNA polymerase II	0.41
chromatin silencing at centromere	0.41
nitrogen compound transport	0.41
histone deacetylation	0.41
ubiquitin-dependent protein catabolic process	0.40
response to methotrexate	0.40
small GTPase mediated signal transduction	0.40

B

biological function	negative r
semaphorin-plexin signaling pathway	-0.55
protein adenylylation	-0.48
DNA double-strand break processing	-0.48
intra-S DNA damage checkpoint	-0.48
N-acylethanolamine metabolic process	-0.45
reductive pentose-phosphate cycle	-0.45
U2-type prespliceosome assembly	-0.44
maintenance of rDNA	-0.44
resolution of mitotic recombination intermediates	-0.44
regulation of mitotic recombination involved in replication fork processing	-0.44
enterobacterial common antigen biosynthetic process	-0.44
photorespiration	-0.43
arginine catabolic process to glutamate	-0.43
carbon fixation	-0.42
carboxylic acid transmembrane transport	-0.42
negative regulation of TOR signaling	-0.41
cellular response to amino acid starvation	-0.41
arginine catabolic process to succinate	-0.40
rRNA (guanine-N7)-methylation	-0.40

Table A4.5: Gene Ontology (GO) biological categories that significantly correlated (Mantel test statistic, p-value < 0.05) to the total organic carbon distribution (TOC).

A

biological function	positive r
fumarate transport	0.65
succinate transmembrane transport	0.65
purine-containing compound salvage	0.58
xanthine metabolic process	0.58
pentose metabolic process	0.54
tRNA 3'-trailer cleavage, endonucleolytic	0.52
NADP metabolic process	0.51
lipoprotein metabolic process	0.51
response to drug	0.49
glycerophosphodiester transmembrane transport	0.49
sodium ion export across plasma membrane	0.48
potassium ion export across plasma membrane	0.48
response to osmotic stress	0.47
cellular monovalent inorganic cation homeostasis	0.47
histone modification	0.47
xenobiotic catabolic process	0.47
regulation of viral transcription	0.46
toxin metabolic process	0.46
peptidyl-lysine demalonylation	0.46
peptidyl-lysine desuccinylation	0.46
chromate transport	0.46
tRNA 3'-trailer cleavage	0.45
branched-chain amino acid catabolic process	0.45
valine catabolic process	0.44
D-xylose transmembrane transport	0.44
amino-acid betaine catabolic process	0.44
glycerol-3-phosphate transmembrane transport	0.43
xenobiotic transport	0.43
alpha-glucan catabolic process	0.43
cellular oligosaccharide catabolic process	0.43
alkaloid metabolic process	0.42
2,4-dichlorophenoxyacetic acid catabolic process	0.42
cholesterol biosynthetic process	0.41
galactitol transport	0.41
peptide metabolic process	0.41
L-threonine catabolic process to glycine	0.41
glucose transmembrane transport	0.41
D-xylose metabolic process	0.40
cellular response to acidic pH	0.40
XMP salvage	0.40
antibiotic catabolic process	0.40
alginic acid biosynthetic process	0.40
leucine catabolic process	0.40

B

biological function	negative r
tRNA seleno-modification	-0.57
cation transmembrane transport	-0.57
peptidoglycan biosynthetic process	-0.53
detoxification of mercury ion	-0.52
cation transport	-0.52
protein peptidyl-prolyl isomerization	-0.51
regulation of cell shape	-0.51
biotin biosynthetic process	-0.50
transcription-coupled nucleotide-excision repair, DNA damage recognition	-0.49
DNA replication	-0.49
cobalamin transport	-0.48
RNA phosphodiester bond hydrolysis	-0.48
rRNA methylation	-0.47
tRNA 3'-terminal CCA addition	-0.47
RNA repair	-0.47
protein retention in ER lumen	-0.46
manganese ion transmembrane transport	-0.46
tetrahydrobiopterin biosynthetic process	-0.46
nucleic acid phosphodiester bond hydrolysis	-0.45
DNA strand renaturation	-0.45
response to mercury ion	-0.45
cobalamin biosynthetic process	-0.45
lipopolysaccharide metabolic process	-0.44
regulation of lipid biosynthetic process	-0.44
cellular manganese ion homeostasis	-0.44
phosphorus metabolic process	-0.43
tRNA guanine ribose methylation	-0.43
photosystem II stabilization	-0.43
rhythmic process	-0.43
response to UV	-0.42
negative regulation of translation	-0.42
DNA repair	-0.42
signal transduction by protein phosphorylation	-0.42
rRNA 2'-O-methylation	-0.42
regulation of DNA repair	-0.41
tRNA pseudouridine synthesis	-0.41
photosynthesis	-0.41
cellular protein modification process	-0.41
maintenance of CRISPR repeat elements	-0.41
RNA methylation	-0.41
DNA metabolic process	-0.41
replication fork processing	-0.41
pyrimidine deoxyribonucleoside triphosphate catabolic process	-0.40
dITP catabolic process	-0.40
uracil transport	-0.40
uracil transmembrane transport	-0.40
DNA replication, synthesis of RNA primer	-0.40

Table A4.6: Gene Ontology (GO) biological categories that significantly correlated (Mantel test statistic, p-value < 0.05) to the total nitrogen distribution (TN).

A

biological function	positive r
purine-containing compound salvage	0.62
xanthine metabolic process	0.62
fumarate transport	0.61
succinate transmembrane transport	0.61
pentose metabolic process	0.60
NADP metabolic process	0.56
L-asparagine biosynthetic process	0.54
lipoprotein metabolic process	0.53
branched-chain amino acid catabolic process	0.52
2,4-dichlorophenoxyacetic acid catabolic process	0.51
valine catabolic process	0.51
alpha-glucan catabolic process	0.49
cellular oligosaccharide catabolic process	0.49
alkaloid metabolic process	0.48
response to drug	0.47
tRNA 3'-trailer cleavage, endonucleolytic	0.47
xenobiotic catabolic process	0.47
peptide metabolic process	0.47
rhamnose transmembrane transport	0.47
ferulate metabolic process	0.47
cinnamic acid catabolic process	0.47
leucine catabolic process	0.46
regulation of viral transcription	0.45
actin filament bundle assembly	0.45
cholesterol biosynthetic process	0.45
histone modification	0.44
glycerophosphodiester transmembrane transport	0.44
amino-acid betaine catabolic process	0.43
organic acid transport	0.43
xenobiotic transport	0.43
sodium ion export across plasma membrane	0.43
potassium ion export across plasma membrane	0.43
aggregation involved in sorocarp development	0.43
autophagic cell death	0.43
D-xylose transmembrane transport	0.43
locomotion	0.43
hemolysis by symbiont of host erythrocytes	0.43
cellular monovalent inorganic cation homeostasis	0.43
response to osmotic stress	0.43
cell-substrate adhesion	0.42
peptidyl-lysine demalonylation	0.42
peptidyl-lysine desuccinylation	0.42
4-hydroxyphenylacetate catabolic process	0.42
negative regulation of Wnt signaling pathway	0.42
glucose transmembrane transport	0.42
response to extracellular stimulus	0.41
gene expression	0.41
actin filament depolymerization	0.41
regulation of myosin II filament disassembly	0.41
mitotic cleavage furrow formation	0.41
intranuclear rod assembly	0.41
cellular response to acidic pH	0.41
cytokine-mediated signaling pathway	0.41
galactitol transport	0.41
aspartate biosynthetic process	0.41
antibiotic catabolic process	0.40
cellular amino acid catabolic process	0.40
tRNA 3'-trailer cleavage	0.40

B

biological function	negative r
cation transmembrane transport	-0.57
tRNA seleno-modification	-0.55
detoxification of mercury ion	-0.53
pyrimidine deoxyribonucleoside triphosphate catabolic process	-0.53
dITP catabolic process	-0.53
D-ribose catabolic process	-0.50
cation transport	-0.50
response to mercury ion	-0.50
mitochondrial ATP synthesis coupled proton transport	-0.50
RNA phosphodiester bond hydrolysis	-0.49
DNA replication	-0.49
DNA replication, synthesis of RNA primer	-0.49
regulation of cell shape	-0.48
transcription-coupled nucleotide-excision repair, DNA damage recognition	-0.48
DNA strand renaturation	-0.47
tetrahydrobiopterin biosynthetic process	-0.47
cobalamin biosynthetic process	-0.47
tRNA 3'-terminal CCA addition	-0.46
RNA repair	-0.46
peptidoglycan biosynthetic process	-0.45
cellular protein modification process	-0.45
cobalamin transport	-0.45
protein retention in ER lumen	-0.45
tRNA pseudouridine synthesis	-0.45
negative regulation of translation	-0.44
response to UV	-0.44
DNA metabolic process	-0.44
anaphase-promoting complex-dependent catabolic process	-0.44
protein polymerization	-0.44
dUTP biosynthetic process	-0.43
purine ribonucleotide biosynthetic process	-0.43
photosynthesis	-0.43
lipopolysaccharide metabolic process	-0.43
regulation of lipid biosynthetic process	-0.43
replication fork processing	-0.43
ribosome biogenesis	-0.42
plasmid maintenance	-0.42
uracil transport	-0.42
uracil transmembrane transport	-0.42
spindle assembly	-0.41
protein transport by the Tat complex	-0.41
toxin biosynthetic process	-0.41
rRNA methylation	-0.40
folic acid-containing compound biosynthetic process	-0.40
D-ribose metabolic process	-0.40
photosystem II stabilization	-0.40
RNA methylation	-0.40
regulation of DNA repair	-0.40

4.E Nitrogen fixation and rock weathering unclassified component

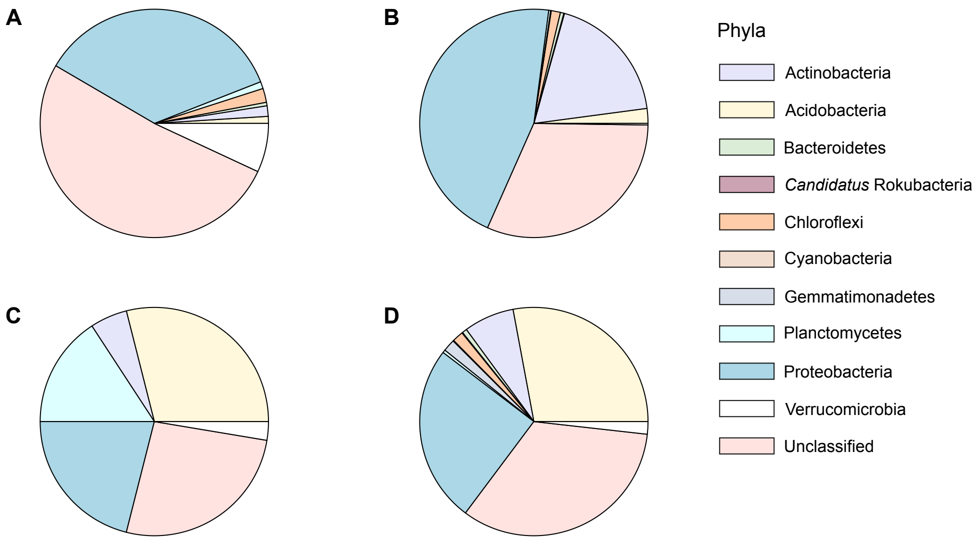


Figure A4.4: Phyla related to the unclassified coding region at the genus-level for (A) nitrogenase genes, (B) obcA genes which are involved in the oxalate biosynthesis, (C) cyanide synthase genes and (D) siderophore-related genes.

Chapter 5

Frozen soil to explore cold-adapted microbial trends and enzymes

Abstract

The polar environment represents a reservoir for the bioprospecting of new compounds and cold-adapted proteins. Cold-adapted proteins are advantageous in several industrial settings because, thanks to their low enzymatic optimum temperature, they do no require heating steps, being more energy sustainable and giving high reaction yields and fewer unwanted secondary chemical reactions. Comparison between thermophilic, mesophilic and psychrophilic proteins can give important information for the bioengineering optimization of cold-adapted proteins.

We collected frozen soil samples from 34 different locations along a transect starting at the Greenland ice sheet (GrIS) edge and going further away ranging a variety of habitats and geochemical specific site conditions with the aim to obtain a deep read coverage assembly to i) analyze the cold-adapted community in this area, studying the DNA and cDNA microbial distribution in proglacial upper permafrost layer, ii) and to reconstruct high- and medium-quality MAGs and to create a database of cold-adapted predicted proteins (CAPP database).

The cold-adapted community converged to similar taxonomic composition along the transect and showed solid permafrost-shaped community rather than microbial trends of typical of proglacial systems. However, whereas DNA profiles were conditioned by the distance from the ice edge, cDNA showed major correlation to geochemical trends showing how microbial activity is determined by environmental conditions.

The construction of the assembly of this conserved frozen soil community led to the retrieval of 69 high- and medium- quality MAGs, which will enrich the ENA MAG public repository, 213 complete biosynthetic gene clusters (BGCs) and more than three million predicted proteins. This information is part of the cold-adapted predicted protein (CAPP) database whose aim is to provide cold-adapted protein sequence information for a protein- and taxon- focused amino acid sequence modification and ultimately to facilitate protein engineering of cold-adapted enzymes.

5.1 Introduction

The use of cold-adapted proteins is advantageous in several industrial settings, such as food and pharmaceutical industries, because their enzymatic reactions do not require heating steps, being more energy sustainable and giving high reaction yields and fewer unwanted secondary chemical reactions at low temperatures (Siddiqui, 2015; Mangiagalli et al., 2020; Kaur and Gill, 2019). The use of cold-adapted enzymes is also useful for bioremediation processes where the direct use of purified enzymes in cold settings can degrade toxic compounds such as phenolic substances, hydrocarbons, plastics and pesticides both *in situ* and *ex situ* (Kumar et al., 2019; Karigar and Rao, 2011; Sharma et al., 2018). Cold-adapted enzymes are found in psychrophiles, organisms isolated from cold environments (Cavicchioli, 2016). These organisms survive to the challenging conditions thanks to cold-adapted enzymes, RNA chaperone overexpression which help with the protein folding, the translation of anti-freezing and ice-nucleation proteins and cell membrane modifications (Bakermans et al., 2009; Collins and Margesin, 2019). Further to this, they are also known to produce antimicrobial compounds because of the high microbial competition levels in cold environments (Mocali et al., 2017; Bell et al., 2013). Cold-adapted proteins are usually studied and purified from different bacterial and fungal isolates (Duarte et al., 2018; Kim et al., 2018; Shcherbakova and Troshina, 2018; Salwoom et al., 2019) and most of them were isolated from the polar regions, thus constituting a reservoir of cold-adapted proteins (Bakermans et al., 2014).

Protein engineering aims to improve the structure and activity of pre-existing proteins. This can be achieved by induced random protein mutations or targeted amino acid substitutions (Kryukova et al., 2019; Kano et al., 1997). Whereas the first approach requires a wide and expensive screening of all the randomly produced proteins, the second is cheaper as relies on the exploration of homologous protein alignments to retrieve functional amino acid substitutions. However, this approach can only be applied if sequences homologous to the studied proteins are present in online or custom protein databases. Thanks to new sequencing technologies and new bioinformatics software, it is now possible to obtain fully complete and ungapped metagenome-assembled genomes (MAGs) where gene clusters, genes and predicted proteins can be assigned with high reliability (Stewart et al., 2019; Somerville et al., 2019). This facilitates the mining of metagenomic and metatranscriptomic datasets for new gene products and protein sequences, enriching public databases with protein sequences obtained from a broad range of organisms and environments.

In this chapter I focused on two particular enzymes: β-galactosidase and polyphenol oxidases. β-galactosidase is a well-known enzyme in the food industry especially used in the diary field

for the production of lactose-free milk, by lactose hydrolysis into galactose and glucose, and the synthesis of galactooligosaccharides (GOS), by lactose transgalactosylation. In industrial settings, β-galactosidases are produced by mesophilic organisms and their enzymatic optimum temperature ranges between 30-60 °C. The use of a cold-adapted β-galactosidase would allow us to conduct the enzymatic reactions at lower temperatures, reducing production costs and food spoilage due to high treatment temperatures (Mangiagalli et al., 2020).
Polyphenol oxidases catalyze the conversion of phenols to non-toxic quinines, and therefore could be used for the treatment of phenolic polluted environments. This substance is hydrophilic and therefore highly soluble in water bodies when released into wastewaters and it has been shown to have negative effects on the micro- and macro- fauna (Babich and Davis, 1981). Phenol is used in different industrial processes as disinfectant and antiseptics because of its antimicrobial properties, and it is produced ie quantities because it is an intermediate of phenolic resins (Panadare and Rathod, 2018). Phenol is also a natural substance mainly secreted by plants. It has been proposed bacteria have developed metabolic pathways to use this molecule to defend from plant resistance mechanisms (Hernández-Romero et al., 2005).

Publicly available β-galactosidase and polyphenol oxidase sequences were analyzed and compared to their homologous sequences found in the upper layer of the Greenland permafrost (i.e. active layer). Permafrost is one of the richest cold-adapted environments where complex and active microbial communities live in constant sub-zero temperatures (Tuorto et al., 2014), overcoming also other challenges, such as high salinity, low water and low nutrient availability (Hultman et al., 2015). Cold and stress adaptation genes have been found by studying bacterial isolates and shotgun whole metagenomic DNA sequences obtained from permafrost (Ayala-Del-Río et al., 2010; Mykytczuk et al., 2013; Vishnivetskaya and Kathariou, 2005; Bakermans et al., 2009) where most of the studies have been focusing on permafrost communities changes along chronosequences (i.e. along depth gradients) and with season changes (e.g. MacKelprang et al., 2017; Schostag et al., 2015). The active layer is the upper layer of the permafrost where the soil undergoes periodic thawing/freezing. Here microbial communities show microbial seasonality and diversity and structural differences with different soil properties (Schostag et al., 2015; Ren et al., 2018; Chen et al., 2017).

We collected a total of 102 different subzero frozen soil samples from 34 different locations in the Greenland ice sheet (GrIS) proglacial system in an area extending from the ice edge to Kangerlussuaq where continuous permafrost has been previously reported (Clarhäll, 2011; Van Tatenhove and Olesen, 1994; Johansson et al., 2015; Jørgensen and Andreasen, 2007). These samples were processed using a multiple sequencing approach to construct a reliable assembly database of cold-adapted gene and protein sequences and to study the active (i.e. cDNA) and potentially active (i.e. DNA) communities across the proglacial cold frozen soil environment. The sequencing was performed with both the MinION and Illumina technologies in order to couple both the Nanopore long and error-prone sequences with the high-quality and short Illumina sequences (Bertrand et al., 2019; Giguere et al., 2020). This approach facilitates more reliable and complete assembly (De Maio et al., 2019; Bertrand et al., 2019) leading to the recovery of MAGs, the prediction of full open reading frames (ORFs) and the correct assignment of full-length genes and predicted proteins. I also coupled metagenomic data, used for the assembly reconstruction, with metatranscriptomic data used to assess the active portion of the microbial communities and the active proteins.

Thanks to this sequencing approach, I achieved three main objectives. Firstly, I characterized the frozen soil community through metagenomic and metatranscriptomic profiles, focusing on whether there were microbial trends across proglacial systems, as usually observed in surface soil (Chapter 4), and if microbial communities correlated with geochemical site-specific variables. The second aim was to construct a solid and deep-coverage assembly to recover highly complete genomes and predict the sequence of cold-adapted proteins, creating the cold-adapted predicted protein (CAPP) database. Finally, I showed how the CAPP database can be used to inform protein engineering approaches by exploring industrial and bioremediation relevant enzymes, focusing on β-galactosidase and polyphenol oxidase.

5.2 Materials and methods

5.2.1 Sampling area and strategy

In July 2018, we collected 102 soil core samples from 34 widespread locations (the two most distant samples being 43 km apart) in the Kangerlussuaq area. In particular the samples were collected following the road which connects the ice edge (point 660), Kangerlussuaq and the Sondrestrom Upper Atmospheric Research Facility (Figure 5.1), forming an artificial transect from the ice edge in the direction of the Greenland coastline.

The GrIS proglacial field is a complex system. Being an ice sheet, the ice moves in different directions (not like the linear glacier movement). Because of this, the samples did not have a linear distance from the ice edge and therefore the sample distance from the ice edge was calculated as the distance from the closest ice edge point (Table 5.1). Due to the size of the area, a variety of habitats and overlying vegetation-types were sampled: we collected frozen soil samples from the upper permafrost layer (i.e. active layer) of riverbanks, thermokarst bogs, grasslands and from heath-dominated environments (Figure A5.1). The sampling activity took

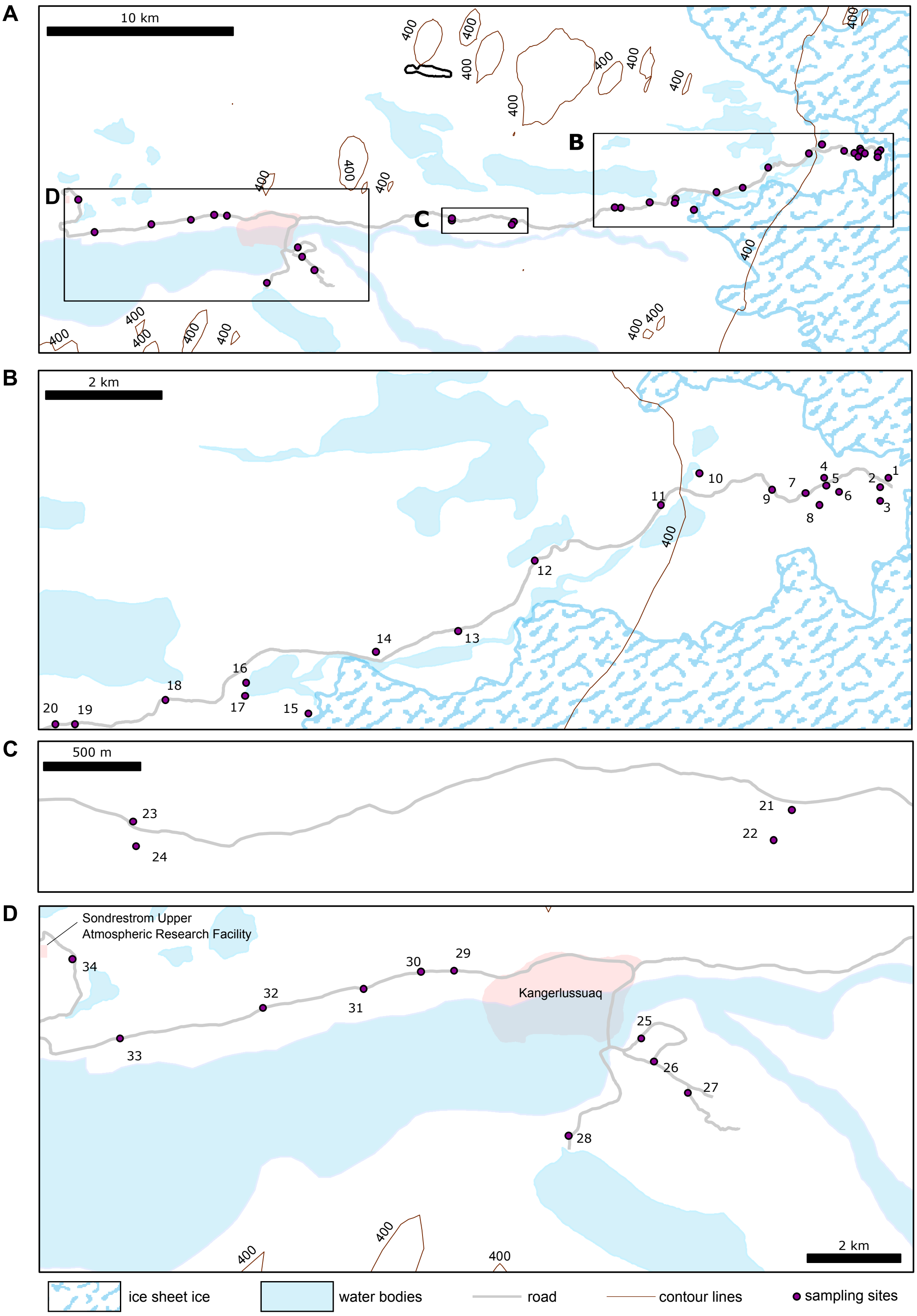


Figure 5.1: Greenland ice sheet (GrIS) proglacial system overview (A) and zoomed-in sites (B, C and D).

place during summer 2018, in July, which is the warmest month of the year in this area of Greenland. In this month, all the snow is melted and as the thawing front moves with the meltdown of the snow (Elberling et al., 2008; Hayashi et al., 2003), we can assert that the soil we sampled, being at sub-zero temperature in July, was at sub-zero temperature all year round. The soil temperature of several sampled sites was also checked on different days and the presence of frozen soil did not seem to vary even after heavy rains occurred.

Samples were collected with a 38 mm diameter soil corer (https://www.geopacks.com/products/soil-sampling-corer) which we modified with an extended shaft so that it could reach one meter depth. Each sample site comprised three technical replicates collected between 30 and 92 cm depth (Table 5.1), the sampling depth variation was due to the different soil textures (Table A5.1) in the different areas, with also a wide range of soil depths between the surface and the bedrocks. The temperature of the soil at the sample collection ranged between -0.1 °C and -2 °C (Table 5.1). Temperatures were checked with a portable digital thermometer and soil was discarded if its temperature was above 0 °C.

Soil was collected in 15 mL falcon tubes for the geochemical analyses. One gram of soil for each of the three technical replicates in each site (for a total of 3 grams per site) were merged together in a 15 mL falcon tube and preserved in 2x LifeGuard Soil Preservation solution (QIAGEN) for DNA and RNA co-extraction. Samples were kept chilled and then frozen at -20 °C in the Kangerlussuaq International Science Support (KISS) station within two days from the collection.

5.2.2 Site characterization and geochemical analyses

The site vegetation was characterized following Clarhäll (2011). Different sites were characterized into four different vegetation categories (Figure A5.1):

1. grassland. It is dominated by grass (e.g. *Calamagrostis langsdorfii*) and moss which typically create hummocks.

2. Wetland was close to lakes, rivers and bogs where the soil is wetter than in the other areas. This environment is dominated by grass, dwarf-shrubs, *Eriophorum scheuchzeri*, *Poa alpina* and *Poa pratensis.*

3. Dwarf-shrub heath. Heath with prevalence of dwarf-shrub is dominated by *Betula nana*, *Rhododendron lapponicum*, *Vaccinium uliginosum* and *Ledum palustre.*

4. Heath with prevalence of high-shrubs *Salix* species.

Geochemical analyses were performed by Dr. Fotis Sgouridis and Dr. Mahammad Rafiq in the

Glaciology department laboratories (University of Bristol). They measured soil moisture, root weight percentage, organic matter percentage, pH, DOC (Dissolved Organic Carbon), TN (Total Nitrogen), DON (Dissolved Organic Nitrogen), NO_3^{-} (nitrate), NH_4^{+} (ammonia), TP (Total Phosphorus), DOP (Dissolved Organic Phosphorus), PO_4^{3-} (phosphate) and SO_4^{2-} (sulfate). Furthermore, they also measured the ion concentrations for K (potassium), Mg (magnesium), Si (silicon), Na (sodium), Ca (calcium) and Fe (iron). Detailed methods can be in Appendix 5.A.2.

Table 5.1: Site description reporting the site distance from the ice edge, the altitude, the soil temperature, soil depth and the site vegetation characterization. The latter can be either classified as grassland, wetland, dwarf-shrub heath (short-heath) or *Salix* heath (tall-heath) (more details in Section 5.2.2 and Figure A5.1).

Site	**Distance from the ice edge** (m)	**Altitude** (m)	**Soil temperature** (°C)	**Sampling depth** (cm)	**Vegetation**
1	290	505	-0.6 ± 0.6	58.3 ± 5.9	grassland
2	400	518	-0.9 ± 0.7	42.7 ± 4.0	grassland
3	420	527	-0.8 ± 0.6	32.7 ± 2.3	grassland
4	630	453	-0.7 ± 0.6	47.3 ± 9.3	grassland
5	680	476	-1.2 ± 0.9	50.3 ± 21.0	short-heath
6	690	452	-1.5 ± 0.0	45.3 ± 11.9	short-heath
7	800	448	-1.0 ± 0.5	44.7 ± 4.6	wetland
8	870	466	-1.0 ± 0.6	46.3 ± 1.2	short-heath
9	1200	430	-1.3 ± 0.2	44.0 ± 4 .0	grassland
10	1300	336	-1.5 ± 0.3	48.7 ± 5.1	wetland
11	1550	356	-0.4 ± 0.4	44.7 ± 4.6	short-heath
12	740	336	-1.2 ± 0.3	45.3 ± 15.0	short-heath
13	950	246	-1.5 ± 0.1	48.3 ± 10.1	short-heath
14	250	248	-0.5 ± 0.3	50.7 ± 5.0	short-heath
15	80	198	-0.4 ± 0.3	55.3 ± 7.1	tall-heath
16	1150	290	-1.5 ± 0.4	36.7 ± 2.9	wetland
17	1050	217	-1.2 ± 0.3	38.3 ± 2.9	short-heath
18	2200	253	-1.5 ± 0.2	42.3 ± 7.5	short-heath
19	3600	244	-1.0 ± 0.5	56.0 ± 19.7	tall-heath
20	3950	211	-1.4 ± 0.4	43.7 ± 5.5	short-heath
21	9300	141	-0.9 ± 0.4	54.7 ± 7.0	wetland
22	9400	141	-1.6 ± 0.1	45.7 ± 7.1	tall-heath
23	12550	134	-0.8 ± 0.3	56.7 ± 15.0	tall-heath
24	12700	149.9	-1.3 ± 0.5	40.0 ± 10.0	short-heath
25	21700	66	-1.3 ± 0.3	80.7 ± 14.0	short-heath
26	21000	216	-1.2 ± 0.7	43.3 ± 12.6	short-heath
27	21500	208	-1.2 ± 0.8	43.3 ± 5.8	tall-heath
28	22700	70	-0.9 ± 0.6	54.0 ± 9.6	short-heath
29	24700	115	-0.4 ± 0.3	63.0 ± 3.0	short-heath
30	25200	122	-1.2 ± 0.7	63.7 ± 17.2	tall-heath
31	26500	86	-1.3 ± 0.4	51.7 ± 10.4	wetland
32	28600	93	-1.6 ± 0.1	45.0 ± 8.7	tall-heath
33	31700	81	-1.1 ± 0.5	46.7 ± 7.6	short-heath
34	32600	171	-1.0 ± 0.5	33.3 ± 2.9	short-heath

5.2.3 DNA preparation and sequencing

DNA and RNA were co-extracted using the RNeasy PowerSoil Total RNA Kit and RNeasy PowerSoil DNA Elution Kit (QIAGEN, Hilden, Germany) according to manufacturer's instructions. The soil, preserved in LifeGuard Soil Preservation solution (QIAGEN), was first centrifuged, the liquid was discarded, and the soil was extracted in duplicate for a total of 3 grams (1.5 g per extraction). Roots were excluded from the soil where possible.

The obtained DNA was then treated with DNase-free RNase A (Sigma-Aldrich, Darmstadt, Germany) for 30 minutes at 37 °C. DNA was then purified with a phenol:chloroform:isoamyl alcohol (25:24:1) extraction and precipitated with 100% ethanol. RNA was first purified with the TURBO DNA-free kit (Thermo Fisher Scientific, Carlsbad, CA, USA) and then concentrated with RNeasy MinElute Cleanup (QIAGEN). I used SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) to synthesize the first cDNA strand and then I followed the protocol for the synthesis and the successive clean-up of the second strand of the SuperScript Double-Stranded cDNA Synthesis Kit protocol (Thermo Fisher Scientific).

Both DNA and cDNA were sequenced with the MinION device, per flow cell (FLO-MIN106) between 4 and 6 samples were barcoded together with the Nanopore kit EXP-NBD104 and then sequenced with the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK). I then sequenced two samples with the Illumina technology: sample 34, chosen because of the high sequencing yield obtained with Nanopore (14,228,430,364 bases), and a sample composed by all the other samples pooled together. I sequenced the sample 34 separately from the others to have high coverage sequencing to obtain high quality MAGs. The Illumina libraries were prepared with the Illumina Nextera chemistry v 3 (Illumina, San Diego, CA, USA) and 150 bp PE were sequenced on the NovaSeq instrument by the commercial supplier Macrogen (Seoul, South Korea). Basecalling was performed with the Real Time Analysis (RTA) software v 3.3.3.

5.2.4 Bioinformatics analyses

Nanopore sequences were basecalled with Guppy v 3.2.2 (Oxford Nanopore Technologies) with the command `guppy_basecaller` and the parameters `--flowcell FLO-MIN106 --kit SQK-LSK109 --barcode_kits EXP-NBD104 --qscore_filtering --min_qscore 7 -q 0 -r --trim_barcodes`. All the Nanopore sequences longer than 1000 bases were assembled with metaFlye (Flye v 2.7) with the options `-g 3g -meta` (Kolmogorov et al., 2019). Nanopore reads were then mapped to the assembly with minimap2 v 2.17 (Li, 2018) and used to polish the assembly with one cycle of Racon v 1.4.15 and command line parameters `-m 8 -x -6 -g -8 -w 500` (Vaser et al., 2017). These options are the same that were used to train the neural network polishing software medaka v 0.10. The latter was run with the command `medaka_consensus` and the option `-m r941_min_high` (https://nanoporetech.github.io/medaka/index.html). Further to these two nanopore polishing steps, the assembly was polished with the Illumina reads running four cycles of Pilon v 1.23 (Walker et al., 2014). Before each Pilon step, reads were mapped back to the assembly with bwa v 0.7.17 with the mem algorithm (Li and Durbin, 2010) where the percentage of mapped reads increased from 67% to 69%, with an increase of more than 34 million reads from the beginning of the polishing steps.

Contig binning was performed with five different algorithms: MetaBAT v 2.12.1 (Kang et al., 2019), MaxBin v 2.2.7 (Wu et al., 2014), CONCOCT v 1.1.0 (Alneberg et al., 2014), BinSanity v 0.3.8 (Graham et al., 2017) and DAS Tool v 1.1.1 (Sieber et al., 2018). The latter combines the information obtained from all the other binning software and therefore was used to ultimately defined the bins. All the bins, or metagenome-assembled Genomes (MAGs) in this case, were checked with CheckM v 1.1.2 (Parks et al., 2015). The parameters I considered to quality check the MAGs were completeness, contamination and strain heterogeneity. The first is expressed as percentage of how many of the single copy genes are found in a MAG. Contamination is the percentage of how many of the single copy genes from non-related taxonomy are present. Strain heterogeneity indicates how many variants of the single copy genes are present.

We considered as high-quality MAGs the bins that have a completeness higher than 90% and a contamination lower than 5%. High-quality MAGs will be uploaded the European Nucleotide Archive (ENA). Medium-quality MAGs were defined with a threshold of 50% and 10% for completeness and contamination, respectively. These thresholds were defined in different published works (e.g. Almeida et al., 2019; Stewart et al., 2019; Bowers et al., 2017).

Coding regions were assigned to the polished assembly with Prokka v 1.14.6 (Seemann, 2014). This software uses Prodigal v 2.6.3 for Open Reading Frame (ORF) prediction (Hyatt et al., 2012), barrnap v 0.9 for ribosomal RNA prediction (https://github.com/tseemann/barrnap) and ARAGORN v 1.2.38 to predict tRNA and tmRNA coding genes (Laslett and Canback, 2004). Taxonomy was assigned to the contigs after performing Diamond v 0.9.22 with the parameters `-e 0.000001 -F 15 --range-culling --range-cover 10 --id 50 --top 5 -f 6 -p 55 -c1 -b4.0 --compress 1` (Buchfink et al., 2015) and using the NCBI non-redundant (nr) database v 5 (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz) (Sayers et al., 2020). The Diamond output was then used to run the `longMeta-assignment` command to obtain taxonomy annotation for each assembled contig (Chapter 4).

Nanopore reads were then mapped back to the assembly with minimap2 (`-ax map-ont` for DNA and `-ax splice` for cDNA sequences) and taxonomy trends were analyzed for both metagenomic and metatranscriptomic data. The first were analyzed with the LongMeta pipeline where all the commands were run with default parameters except for `longMeta-coverage` where `--perc-limit` was set to 0.001. Whereas, DNA trends were estimated as coverages, the cDNA trends were estimated as read count associated to each different taxonomy using a custom script.

5.2.5 Cold-adapted predicted protein (CAPP) database

The coding regions and proteins predicted from the high-quality assembly were used to construct the cold-adapted predicted protein database (CAPP database) which can be found at https://www.cerealsdb.uk.net/CAPP/CAPP.txt.gz together with the protein sequence file https://www.cerealsdb.uk.net/CAPP/CAPP.faa.gz and the gene sequence file https://www.cerealsdb.uk.net/CAPP/CAPP.fasta.gz. The CAPP.txt.gz file contains, associated to each entry, the following information:

- Entry name.
- How the coding region was assigned, if by similarity to a protein motif or by similarity to a protein sequence.
- Gene name.
- Protein name.
- Enzyme Commission (EC) number.
- Taxonomy assignment.
- MAG name, if the protein was from one of the high-quality or medium-quality MAGs.
- Field reporting whether the CDS corresponded to one or more cDNA reads. cDNA sequences were mapped to the DNA sequences with minimap2 (`-ax splice`).
- nr database accession numbers, if there were matches between CAPP entries and nr proteins. nr proteins were mapped to the CAPP proteins with Diamond, command `-e 0.000001 --id 70 --query-cover 90`.

The cold-adapted protein database consists in 3,076,838 coding regions/proteins. Genes coding for rRNA, tRNA and tmRNA were excluded from the database.

CAPP amino acid frequencies were compared to protein sequences associated to 10 known thermophilic and 10 known psychrophilic genomes (Table A5.2). The data were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/genome/) and the amino acid frequencies were calculated with a custom script.

I then explored two different protein products extracted from the CAPP database: β-galactosidase (EC 3.2.1.23) and polyphenolic oxidase (EC 1.10.3.-). To work with expressed and functional proteins at sub-zero temperatures, I used only proteins that were mapped back to the Nanopore cDNA sequences. These proteins, and the nr proteins that aligned to them, were then aligned with mafft v 7.271 with parameters `--localpair --maxiterate 1000` (Katoh and Standley, 2013) and then used to construct a phylogenetic tree with iqtree v 1.6.12 using default parameter and performing 1000 bootstraps (Nguyen et al., 2015). The newick trees were then drawn as cladograms with the software FigTree v 1.4.4 (http://tree.bio.ed.ac.uk./software/figtree/). These trees were built in order to show the sequence relation and clustering without any purpose to show evolutionary relationships between the different protein clusters, as no accurate root was chosen for the tree reconstruction. Sequences used for the tree construction can be found in Appendix 5.D.

Whereas the polyphenolic oxidase protein sequences included in the tree were encoded by only one gene, the β-galactosidase sequences were encoded by three homologous genes (i.e. BoGH2A, lacZ and bga). In order to further investigate the specific amino acid usage in the protein sequences, I selected one sequence cluster from each tree. I selected protein clusters where both the nr and CAPP proteins were present in a similar number in order to perform a balanced permutational multivariate analysis of variance (PERMANOVA) to detect protein positions with significant differences in amino acid composition between the proteins belonging to the two different databases (CAPP and nr database; more details in the next section).

Finally, secondary metabolite gene clusters were predicted using antiSMASH v 5.1.2 (Blin et al., 2019) on high- and medium- quality MAGs. antiSMASH was run with command parameters `--genefinding-tool prodigal-m --smcog-trees`.

5.2.6 Statistical analyses

All the statistical analyses and the result plots were performed in the R environment v 6.6.1 (R Core Team 2019, 2019) thanks to the R packages vegan v 2.5-6 (Oksanen, 2017), gplots v 3.0.3 (Warnes, 2012), ggplot2 v 3.3.0 (Wickham, 2016), ggfortify v 0.4.10 (Tang et al., 2016), tidyr v 1.0.3 (Wickham and Henry, 2019), plyr v 1.8.6 (Wickham, 2011) and gridExtra v 2.3 (Auguie, 2017).

Mantel tests were performed to analyze the relation between site distance from the ice edge, geochemical variables and taxonomic trends. Mantel test statistics (r) were performed with 9999 permutations and considered significant only for p-value < 0.05. Mantel tests were calculated using Spearman's rank correlation coefficients on Euclidean matrices when considering the distance from the ice edge and the geochemical dataset and Bray-Curtis matrices for the taxonomic datasets.

Spearman's rank-order correlation coefficient (r_s) was calculated to i) detect correlations between the distance from the ice edge and the other geochemical variables and ii) detect correlations between the DNA and cDNA datasets and the geochemical variables. The r_s was considered significant only if p-value < 0.05.

Permutational multivariate analysis of variance (PERMANOVA; 9999 permutations) was performed at each position of the protein sequences of the β-galactosidase and polyphenol oxidase sequences selected from the respective trees. The tested variables were the different amino acid frequencies for the analyzed protein sequences, whereas the tested factor was the database from which the proteins were retrieved: nr database versus the CAPP database (i.e. mesophilic vs psychrophilic proteins). The R^2 values were considered significant only for p-value < 0.05.

All taxonomic profiles are reported as relative abundances calculated on coverage values for DNA dataset and as read counts for the cDNA dataset.

5.3 Results

5.3.1 Site characterization

The Mantel test statistic (r) performed between the geochemical dataset and site distance from the ice edge (p-value < 0.05) was 0.21. The Mantel test statistic (r) performed between grain size distribution dataset (Table A5.1) and site distance was also significant with r equals to 0.32.

Spearman statistical analyses showed a significant positive correlation (p-value < 0.05) between the site distance from the ice edge and the vegetation complexity, pH, root weight percentage, ${PO_4}^{3-}$ (phosphate), Na (sodium) and Fe (iron) concentration. Whereas, it showed a significant negative correlation with altitude, soil temperature, organic matter percentage, total nitrogen (TN), ${NO_3}^{-}$ (nitrate), ${SO_4}^{2-}$ (sulfate) and Ca (calcium) concentration (Table 5.2 and Figure A5.2). The vegetation complexity increased with the distance having grassland environment closer to ice edge and then development of dwarf shrubs and shrubs going further away. The other geochemical factors did not show any significance with the distance (Figure A5.3).

5.3.2 Assembly result and MAG reconstruction

For each sample, I obtained between 144,844,933 (site 32) and 14,228,430,364 bases (site 34) from the Nanopore DNA sequencing. I also obtained a total of 800 Gb of data from the Illumina

Table 5.2: Spearman's rank correlation coefficients (r_s) calculated between the site distance from the ice edge and site variables. Variables that resulted to be significantly correlated (p-value < 0.05) with the distance from the ice edge are colored in red (positive correlation) or blue (negative correlation). The symbol '-' is reported for r_s were p-value ≥ 0.05. *in the vegetation complexity definition the grassland and wetland environments were assumed to be the least complex, followed by dwarf-shrub heath and *Salix* heath. **grain size was defined as the peak size value in the grain size distribution.

Site variables	r_s	p-value
altitude	-0.82	0.00
vegetation complexity*	0.34	0.00
soil temperature	-0.20	0.00
pH	0.46	0.00
soil depth	-	-
soil moisture	-	-
root	0.20	0.04
organic matter	-0.27	0.00
grain size**	-	-
TN	-0.23	0.02
DON	-	-
TP	-	-
DOP	-	-
DOC	-	-
NO_3^-	-0.38	0.00
NH_4^+	-	-
PO_4^{3-}	0.32	0.00
SO_4^{2-}	-0.24	0.01
Na	0.27	0.01
K	-	-
Mg	-	-
Ca	-0.39	0.00
Fe	0.26	0.01
Si	-	-

sequencing. The final assembly constructed with Nanopore sequences and polished with Illumina reads consisted in 2,834,571,546 bases. The assembly N50 was 40,615 bases. The minimum contig length was 1000 bases, length under which all the contigs were removed, and the maximum was 3,693,048 bases long. The latter corresponded to a complete and circularized bacterial genome (MAG-01). Illumina and Nanopore reads coverage over the assembly was 164x and 18x, respectively. The different used binning algorithms gave different numbers of recovered bins: Metabat2 gave 1005 bins, MaxBin2 found 517, Concoct 236, BinSanity found 618 and DAS Tool found 267. I considered as definitive the DAS Tool results as it combines all the information obtained from the other tools to create consistent bins. The high-quality MAGs (completeness $\leq$ 90 and contamination $\leq$ 5) were 9. The medium-quality MAGs (completeness $\geq$ 50 and contamination $\leq$ 10) were 60.

All the high-quality MAGs had consistent taxonomy across all the contigs, and they all belonged to Bacteria. The high-quality MAGs belonged to the phyla Chloroflexi (4 MAGs), Acidobacteria (2 MAGs), Proteobacteria (2 MAGs), Bacteroidetes (1 MAG). MAG-01 was a complete MAG

Table 5.3: High-quality MAGs. The table reports MAG completeness, contamination and strain heterogeneity as calculated from CheckM. It also reports the number of contigs in each MAG and the taxonomic classification as phylum, class, order, family and genus. The taxonomic classification was reported only when more than the 70% of the contigs where assigned to the same taxon.

MAGs	**Completeness** (%)	**Contamination** (%)	**Strain heterogeneity**	**Contig number**	**Taxonomy**
MAG-01	99.6	0.0	0 .0	1	Proteobacteria, Deltaproteobacteria, Desulfuromonadales, Geobacteraceae
MAG-02	99.1	0.9	0.0	5	Chloroflexi
MAG-03	98.2	0.9	100.0	5	Chloroflexi
MAG-04	98.2	4.6	16.7	19	Chloroflexi
MAG-05	96.8	3.8	50.0	129	Acidobacteria
MAG-06	96.0	3.2	0.0	30	Proteobacteria
MAG-07	96.0	0.9	100.0	11	Chloroflexi
MAG-08	91 .0	1.2	25.0	25	Bacteroidetes, Chitinophagia, Chitinophagales, Chitinophagaceae
MAG-09	90.2	4.9	0.0	138	Acidobacteria

(cMAG), circularized and ungapped, belonging to the bacterial family Geobacteraceae (Table 5.3).

Medium-quality MAGs were also assigned to the phyla Chloroflexi (6 MAGs), Acidobacteria (8 MAGs), Proteobacteria (11 MAGs) and Bacteroidetes (4 MAGs) but also to Actinobacteria (14 MAGs), Verrucomicrobia (2 MAGs) and other phyla such as Gemmatimonadetes and Nitrospirae. MAG-20 was the only one assigned to Archaea (phylum Thaumarchaeota) with a completeness of 85% and a contamination of 5% (Table A5.3). Only nine high- and medium-quality MAGs were assigned at the family or genus level.

The assembly was annotated with Prodigal which found 3,109,960 open reading frames (ORFs). Almost the 99% of these were CDSs (coding DNA sequences), 1% were rRNA genes and less than the 0.1% were represented by tRNA and tmRNA genes. About one sixth of the ORFs were mapped back to the cDNA transcripts. Between the 50-84% and the 50-81% of the DNA reads (longer than 1000 bp) and cDNA reads (longer than 100 bp) mapped back to the assembly.

5.3.3 Metagenomic trends

Mantel test statistics (r) showed a significant (p-value < 0.05) correlation between the phylum-, class- and genus-level taxonomic datasets and the site distance from the ice edge ($r = 0.25$, $r = 0.27$, $r = 0.27$) and the taxonomic and geochemical datasets ($r = 0.20$, $r = 0.22$, $r = 0.22$).

Bacteria represented between the 93 and the 99% of the organisms in all the samples, Archaea between the 0-7%, viruses the 0-1% and Eukaryota, all belonging to fungal organisms, had a maximum of 1% in the samples (Figure A5.4A).

Microbial distribution did not show particular trends along the succession (Figure A5.4). The most abundant phyla were the Proteobacteria (33-52%), followed by Actinobacteria (12-22%), Bacteroidetes (4-24%), Firmicutes (2-12%), Cyanobacteria (1-7%), Acidobacteria (1-5%), Chlo-

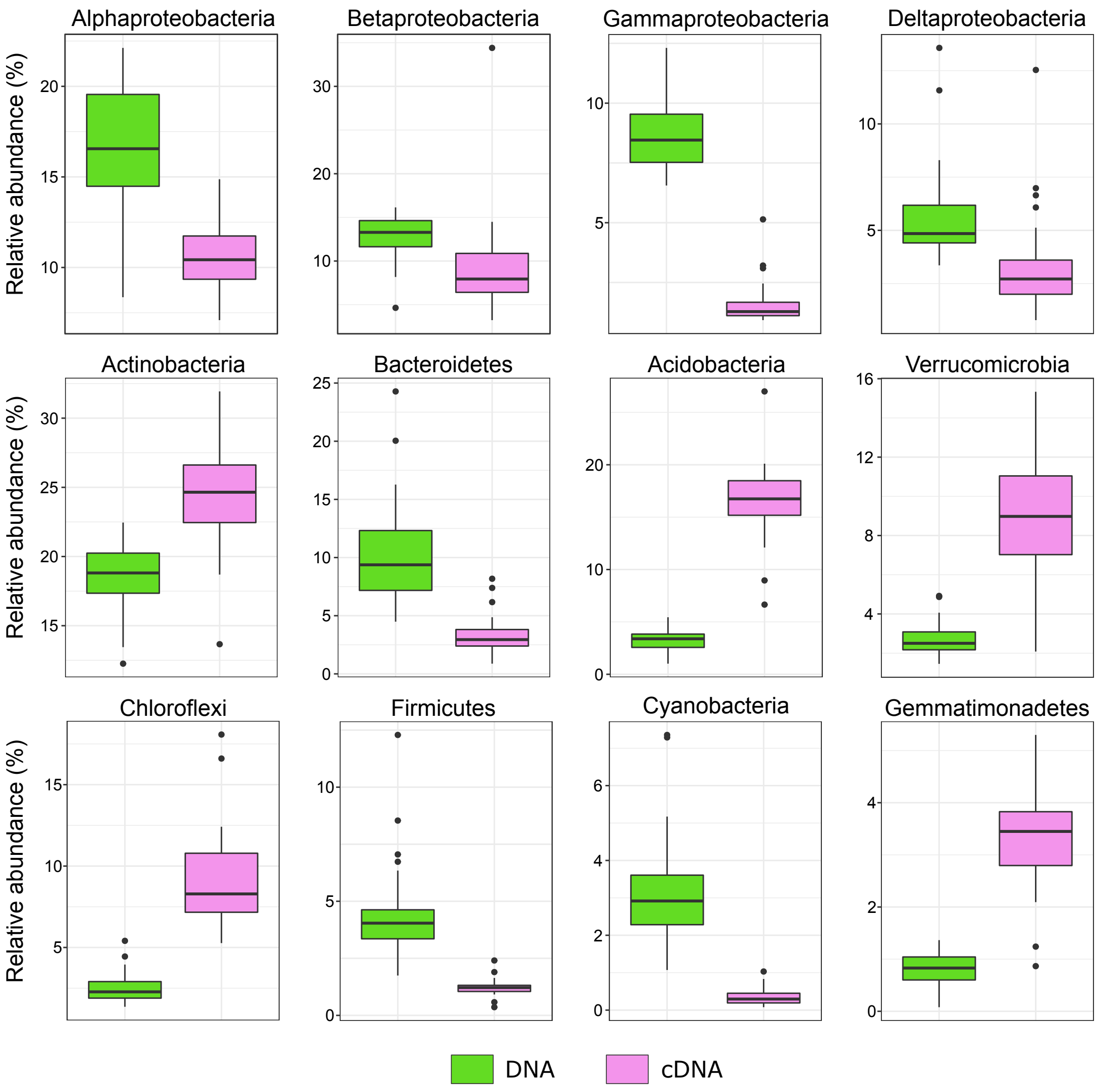


Figure 5.2: Most abundant phyla and Proteobacteria classes in the DNA and cDNA profiles. Only taxa with a minimum relative abundance of 5% in at least one sample of the DNA or cDNA datasets are reported.

roflexi (1-5%) and Verrucomicrobia (3-5%). The Proteobacteria classes had an abundance of 3-14% for Deltaproteobacteria, 5-16% for Betaproteobacteria, 7-12% for Gammaproteobacteria and 8-22% for Alphaproteobacteria (Figure 5.2). Other bacterial classes present with a high relative abundance were Actinobacteria (11-20%), Chitinophagia (1-10%), Clostridia (1-9%) and Bacteroidia (0-13%). The most abundant archaeal phyla were Euryarchaeota representing up to the 6% in some sites and Thaumarchaeota representing up to the 2%.

5.3.4 Metatranscriptomic trends

The Nanopore cDNA sequencing output ranged between 4,802,857 (site 33) and 1,025,359,890 bases (site 2) per sample. In each sample, between the 87 and 98% of the Nanopore cDNA reads, that mapped back to the assembly, fell inside predicted genes. Of these, more than the 98%

of the mapped cDNA reads fell inside one ORF and the 2% spanned more than one ORF. In almost all the samples, more than 98% of the mapped reads were represented by genes coding for ribosomal RNA. The exceptions were site 5, 18 and 25 were the ribosomal reads only represented the 94, 91 and 57% of the reads, respectively.

Mantel test statistics (r) showed no correlation between cDNA taxonomy dataset and site distance. However the correlation was significant between the geochemical dataset and the cDNA dataset with an r equals to 0.23, 0.36, 0.40 at phylum-, class- and genus- level respectively (p-value < 0.05).

In this dataset, the most abundant phyla were the same as in the metagenomics dataset. However, the relative abundances varied between DNA and cDNA datasets. Proteobacteria varied between 14-53%, Actinobacteria between 14-32%, Bacteroidetes 1-8%, Firmicutes 0-2%, Cyanobacteria 1-7%, Acidobacteria 7-27%, Chloroflexi 5-18%, Verrucomicrobia 2-16% and Gemmatimonadetes 1-5%. The most abundant classes were Alphaproteobacteria (7-15%), Gammaproteobacteria (1-5%), Deltaproteobacteria (1-13%), Betaproteobacteria (3-34%) and Actinobacteria (13-30%) (Figure 5.2 and A5.4D).

5.3.5 Microbial dataset and site characteristic comparison

Comparing the phyla relative abundances between the DNA and cDNA datasets, the phyla Actinobacteria, Acidobacteria, Verrucomicrobia, Chloroflexi and Gemmatimonadetes were the only ones present with a higher abundance in most of the sites for the cDNA dataset (Figure 5.2).

Looking at how the DNA and cDNA datasets varied in relation to different site variables (e.g. geochemical data), we can observe that, compared to the DNA dataset, the cDNA dataset had more bacterial classes that showed significant (p-value < 0.05) Spearman's correlation coefficients (r_s) (Figure 5.3). In the cDNA dataset, most of the classes showed a positive correlation with the ion and nutrient distributions. In particular, the classes Ignavibacteria, Thermomicrobia, Spirochaetia, Bacteroida, Cytophagia, Betaproteobacteria, Gammaproteobacteria and Chitinophagia showed a positive correlation with the Ca, Mg, $SO_4{}^{2-}$, DOC, DON and TN concentrations and the soil moisture and the organic matter percentage. No correlation between cDNA relative abundances and the distance from the ice edge was observed (Figure 5.3B).

When the Spearman's correlation coefficient (r_s) was calculated at the genus level, the DNA and the cDNA datasets showed significant r_s for 225 and 116 genera (out of 941 and 860 in total), respectively. The DNA dataset had more genera (145) correlating with the distance from the ice edge and other site-based characteristics, such as altitude, soil temperature, soil depth and

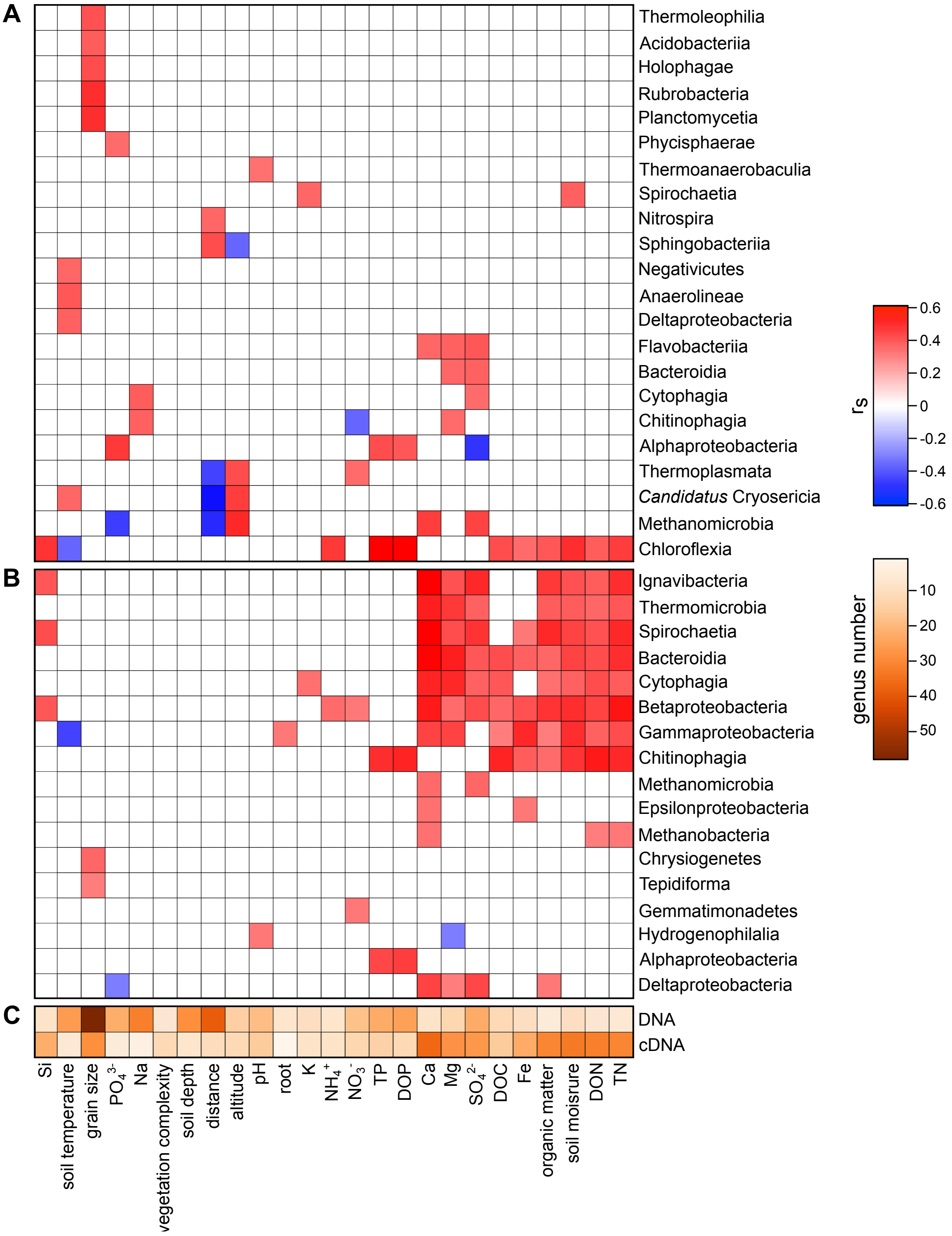


Figure 5.3: Taxonomic correlation with site characteristics and geochemical variables. Spearman's rank correlation coefficients (r_s) between taxonomic classes that significantly correlated (p-value $<$ 0.05) with at least one of the site variables for DNA (A) and cDNA (B). Number of genera that showed a significant correlation (p-value $<$ 0.05) in the DNA or cDNA dataset (C).

grain size, and also TP, DOP, NO_3^- PO_4^{3-}, Na and K, compared to the cDNA dataset (70). In the cDNA dataset, more genera (280) correlated with vegetation, soil moisture, organic content, TN, DON, DOC and ion concentrations, such as NH_4^+, SO_4^{2-}, Mg, Ca, Fe and Si (200) (Figure 5.3C).

5.3.6 Cold-adapted predicted protein (CAPP) database exploration

The cold-adapted predicted protein database (CAPP database) consisted of more than 2000 enzyme classes (according to the Enzyme Classification system) where the most common classes were histidine kinase (EC 2.7.13.3) with 14429 entries, non-specific serine/threonine protein kinase (EC 2.7.11.1) with 10138, ABC-type vitamin B12 transporter (EC 7.6.2.8) with 8029, D-inositol-3-phosphate glycosyltransferase (EC 2.4.1.250) with 6959 and DNA helicase (EC 3.6.4.12) with 6575. Table 5.4 reports a non-exhaustive list of enzymes with industrial relevance that were found in the CAPP database. These enzymes include, for example, those involved in food processing (e.g. β-galactosidases and α-amylases), in molecular biology protocols (e.g. DNA ligases and alkaline phosphatases) and also enzymes used in the bioremediation field, such as cellulases and polyphenol oxidases.

The comparison between the amino acid usage of heat and cold adapted proteins obtained from known psychrophilic and thermophilic organisms (Table A5.2) showed a separation between cold and heat-adapted genomes along the second component analysis (PC2) which explained the 17.5% of the observed variance (Figure 5.4). This clustering was explained by an increase of the amino acids glutamic acid (E) and leucine (L) in the heat adapted proteins and an increase of methionine (M), serine (S), glutamine (Q), cysteine (C), threonine (T), aspartic acid (D), histidine (H) in the cold adapted proteins (Figure 5.4). The CAPP proteins clustered closer to the known cold adapted proteins (e.g. *Rhodoferax antarcticus* and *Cryobacterium psychrophilum*).

In order to further explore the CAPP database and to identify specific amino acid substitutions in specific industrial relevant proteins, I further analyzed the enzymes β-galactosidase (EC 3.2.1.23) and polyphenol oxidase (EC 1.10.3.-).

The β-galactosidase sequences separated into three main different clusters, one for each of the three genes that code for this multi-component enzyme (i.e. BoGH2A, lacZ and bga). The tree was constructed with 100 protein sequences from the CAPP database and 207 sequences from the nr database. Sixty-eight of the database sequences were assigned by similarity with protein sequences and 42 were assigned by similarity with β-galactosidase protein motives. The nr sequences were derived from freshwater and soil metagenomes and were assigned mainly to the phyla Acidobacteria, Actinobacteria, Bacteroidetes, Chloroflexi and the Proteobacteria classes Alphaproteobacteria and Betaproteobacteria (Figure 5.5).

Detailed amino acid composition was studied for the Chloroflexi cluster marked in Figure 5.5. PERMANOVA performed at each alignment position showed significant (p-value < 0.05) differences in the amino acid composition between the CAPP and the nr sequences at 24 positions. Valine (V) and glutamic acid (E) were the most observed amino acids in the nr proteins for the highlighted protein positions. In the CAPP proteins they were not always substituted by the same amino acids. In particular V was substituted in six positions mainly with alanine (A), leucine (L), isoleucine (I) and serine (S) in the CAPP sequences. The amino acid glutamic acid corresponded in 4 positions with alanine (A), aspartic acid (D) and glutamine (Q). Vice-versa, in the CAPP database, the most substitutions were with the amino acids alanine (A) and leucine (L) but did not correspond all to the same amino acid in the nr database (Table 5.5A).

The phylogenetic tree built with polyphenol oxidase proteins reports 33 and 54 entries from the CAPP and the nr database, respectively. All the nr entries were from temperate soil and freshwater metagenomes, except from one entry that can be found in the Verrucomicrobia cluster and that was isolate from the Antarctic region. All the CAPP proteins were assigned to these genes by similarity with UniProt sequences. The tree sequences belonged mainly to the phyla: Acidobacteria, Proteobacteria (i.e. Alphaproteobacteria) and Chloroflexi (Figure 5.6).

Table 5.4: Non-exhaustive list of CAPP enzymes, found in the cold-adapted predicted protein database, relevant for food processing, molecular biology and bioremediation applications. The table reports the Enzyme Commission (EC) number associated to the protein, the total number of protein sequences present in the CAPP dataset and the number of CAPP proteins that matched to cDNA sequences.

Enzyme application	EC number	Enzyme name	CAPP entries	cDNA mapped CAPP entries
Food processing	3.2.1.23	β-galactosidase	686	100
	3.2.1.1	Alpha-amylase	366	89
Molecular biology	6.5.1.1	DNA ligase (ATP)	1104	247
	3.1.3.1	Alkaline phosphatase	368	82
	3.2.2.27	Uracil-DNA glycosylase	332	35
Bioremediation	3.2.1.21	β-glucosidase	653	126
	1.10.3.-	Polyphenol oxidase	599	118
	3.2.1.4	Cellulase	214	43
	1.13.11.2	Catechol 2,3-dioxygenase	312	46

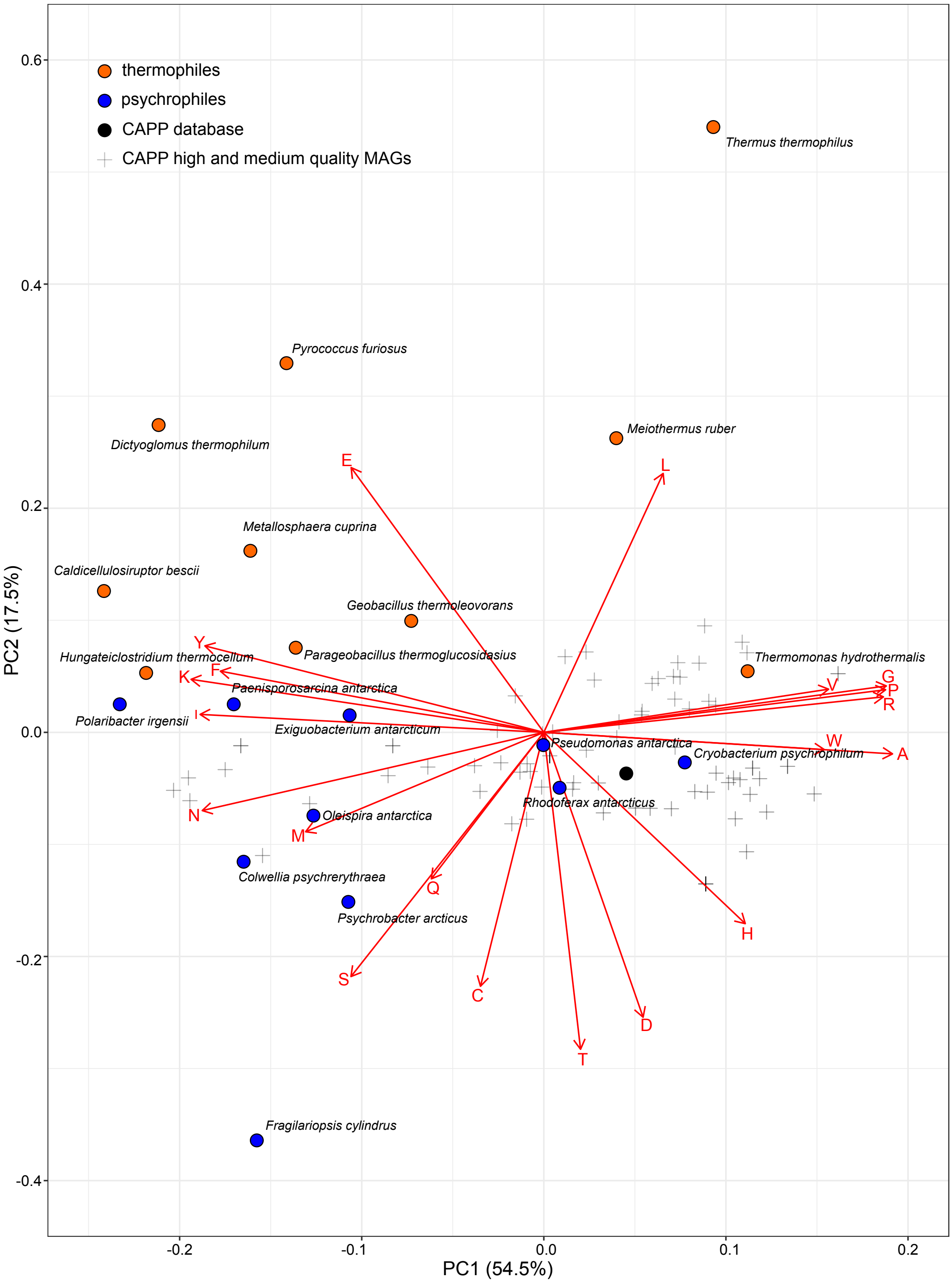


Figure 5.4: Principal component analysis (PCA) of the amino acid composition of proteins from known thermophilic and psychrophilic organisms, from the cold-adapted predicted protein (CAPP) database and from the high and medium quality MAGs retrieved from the CAPP database. Vectors indicate direction of the amino acid effect in the genomes' sequence composition (Bray-Curtis similarity).

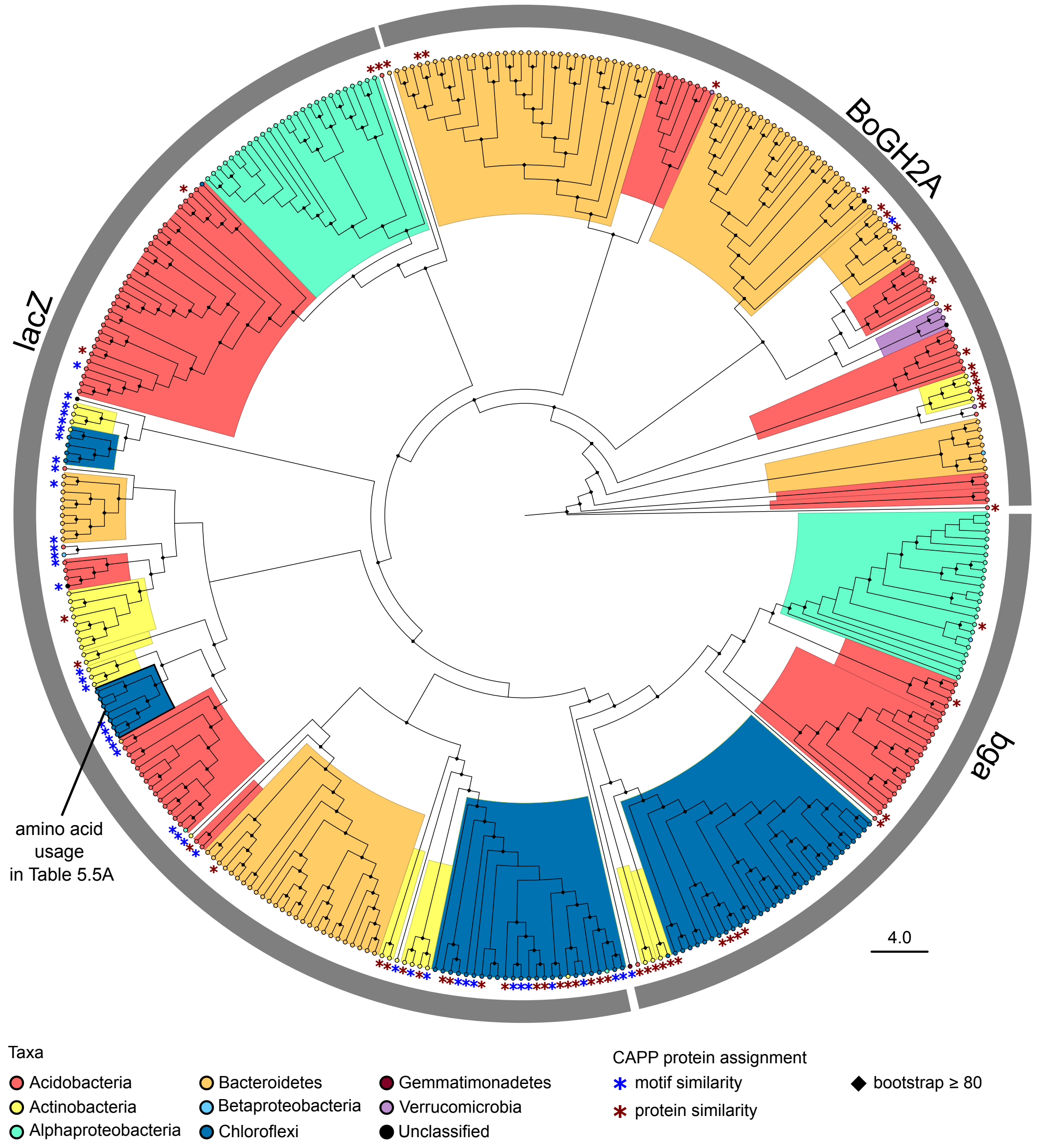


Figure 5.5: Phylogenetic tree constructed for the protein β-galactosidase with both the CAPP database proteins, marked by '*', and nr proteins. Sequence accession numbers are reported in Appendix 5.D.

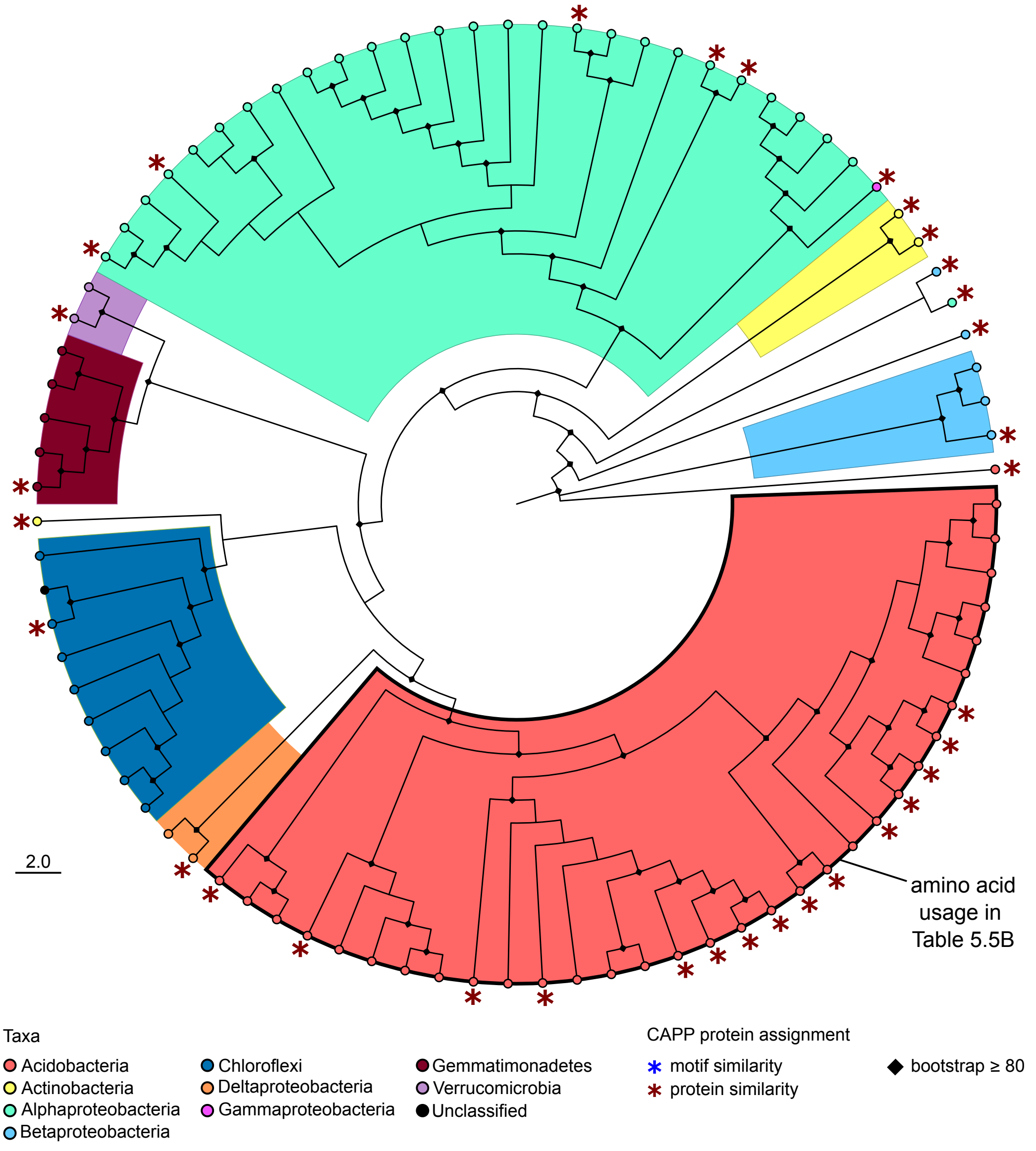


Figure 5.6: Phylogenetic tree constructed for the protein polyphenol oxidase with both the CAPP database proteins, marked by '*', and nr proteins. Sequence accession numbers are reported in Appendix 5.D.

A

Protein position	Amino acids in nr database												Amino acids in CAPP database															
	D	E	F	G	H	I	K	M	P	Q	R	V	A	D	E	G	I	K	L	M	N	Q	R	S	T	V	W	Y
323	-	-	-	-	-	-	-	-	1.0	-	-	-	0.7	-	-	-	-	-	-	-	-	-	-	0.3	-	-	-	-
357	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	0.3	-	-	-	-	-	0.7	-	-	-	-	-	-
426	-	-	-	-	-	-	-	1.0	-	-	-		-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-
469	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-
513	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.7	-	-	-	-	-	-	-	0.3	-
571	-	1.0	-	-	-	-	-	-	-	-	-	-	0.3	0.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
574	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-
590	-	1.0	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-
594	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-
622	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.3	-	-	-	-	-	-	0.3
651	-	-	-	-	-	-	-	-	-	-	-	1.0	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
652	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-
662	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	0.3	-	-	-	0.3	-	-	-	-	-	-
696	-	-	-	1.0	-	-	-	-	-	-	-	-	0.7	-	-	-	-	-	-	-	-	-	0.3	-	-	-	-	-
746	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-
765	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-
770	-	1.0	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
775	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-
786	-	-	-	-	-	-	-	-	-	-	-	1.0	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
793	-	-	-	-	-	-	1.0	-	-	-	-	-	-	0.3	-	-	-	-	-	-	-	-	-	-	0.3	-	-	-
842	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	1.0	-	-	-	-	-	-	-	-	-
872	-	-	-	-	-	-	-	-	-	-	-	1.0	-	-	-	-	-	-	0.3	-	-	-	-	-	0.3	-	-	-
890	-	-	-	-	-	-	-	-		1.0			-	-	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-
902	1.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.3	-	-	-	-	-	-	-	0.3	-	-	-	-
	2.0	4.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	2.0	6.0	4.7	1.7	3.0	0.7	1.0	0.3	4.0	1.0	0.3	1.0	0.3	0.7	0.7	2.0	0.3	0.3

B

Protein position	Amino acids in nr database															Amino acids in CAPP database											
	A	D	E	F	G	H	I	K	M	P	Q	R	S	T	V	A	D	E	I	L	M	P	Q	R	S	T	V
75	-	-	-	0.2	-	-	0.1	-	0.5	-	-	-	-	-	0.1	-	-	-	0.1	-	0.9	-	-	-	-	-	-
89	0.1	0.2	0.2	-	-	-	-	0.1	-	-	-	-	0.1	-	-	0.1	0.3	-	-	-	-	-	-	-	0.1	0.5	-
91	0.1	0.1	0.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.9	-	-	-	-	-	-	-	-	-
127	-	-	-	-	-	-	-	-	-	-	-	-	-	0.1	0.8	-	-	-	0.3	0.4	-	-	0.1	-	-	-	0.2
128	-	-	-	0.1	-	-	-	-	-	-	-	-	-	-	0.8	-	-	-	0.4	-	-	-	-	-	-	-	0.6
177	-	-	-	-	0.5	0.1	-	0.1	-	-	-	0.4	-	-	-	-	-	-	-	-	-	-	-	0.1	-	-	-
245	0.2	-	-	-	-	-	0.2	-	-	0.4	-	-	-	0.1	-	-	-	-	0.1	-	-	0.9	-	-	-	-	-
247	0.5	-	-	-	0.1	-	-	-	-	-	-	0.2	0.1	-	-	0.1	-	-	-	-	-	-	0.3	0.1	0.5	-	-
252	0.1	-	-	-	-	-	-	0.2	-	0.1	0.1	0.1	-	0.2	0.1	-	-	-	-	-	-	0.1	-	-	-	0.9	-
	0.9	0.4	0.7	0.2	0.5	0.1	0.3	0.3	0.5	0.5	0.1	0.8	0.2	0.4	1.8	0.1	0.3	0.9	0.8	0.4	0.9	1.0	0.4	0.2	0.6	1.4	0.8

Table 5.5: Amino acid substitutions for the β-galactosidase (A) and the polyphenol oxidase protein (B). Only the protein positions where the PERMANOVA performed on the amino acid distribution between the cold-adapted predicted protein (CAPP) and the nr database was significant (p-value < 0.05) and with an $R^2 > 0.1$ are shown. The values associated with the amino acids correspond to the proportion of proteins that for a specific position is associated to a specific amino acid. The β-galactosidase table was constructed looking at 4 proteins from the nr database and 3 proteins for the CAPP database whereas the polyphenol oxidase table was constructed with 17 and 15 proteins from the nr and CAPP database, respectively.

I then explored in detail the amino acid frequency in the Acidobacteria cluster highlighted in Figure 5.6 composed by 15 and 17 CAPP and nr proteins, respectively. The amino acid that was mostly present in differing positions (highlighted by the PERMANOVA) in the nr database was valine (V). This amino acid corresponded to different amino acids in the CAPP proteins, mainly to isoleucine (I), methionine (M) and threonine (T) (Table 5.5B). As in Table 5.5A, position substitution in Table 5.5B did not show univocal correspondences between one amino acid in the nr database to another in the CAPP database.

Contigs belonging to the high- and medium-quality MAGs had 231 complete secondary metabolite biosynthetic gene clusters (BGCs) identified by antiSMASH (Blin et al., 2019). The most abundant BGCs codified for terpene, Type III Polyketide Synthase (T3PKS), Non-Ribosomal Peptide Synthetase (NRPS) and NRPS-like, bacteriocin, β-lactone and arylpolyene. The main producers were the phyla Acidobacteria, Proteobacteria, Actinobacteria and Chloroflexi (Table 5.6).

Table 5.6: Complete biosynthetic gene clusters (BGCs) predicted from high- and medium-quality MAGs. *Type III polyketide synthases. **Nonribosomal peptide synthetases. ***Type I polyketide synthases. ****Linear azol(in)e-containing peptides. *****Heterocyst glycolipid synthase-like PKS

			Phyla												
		Cluster number	Acidobacteria	Proteobacteria	Actinobacteria	Chloroflexi	Bacteroidetes	*Ca.* Cryosericota	Gemmatimonadetes	Nitrospirae	Verrucomicrobia	Thaumarchaeota	*Ca.* Rokubacteria	Unclassified
BGCs	terpene	52	12	11	11	3	1	-	1	3	4	-	1	5
	T3PKS*	24	6	6	1	3	5	-	-	-	-	-	-	3
	NRPS-like	23	11	2	2	1	-	-	-	-	-	-	-	7
	bacteriocin	21	7	3	3	-	-	1	2	1	1	-	1	2
	NRPS**	20	7	4	2	1	1	-	-	-	-	-	-	5
	β-lactone	18	-	4	6	3	-	5	-	-	-	-	-	-
	arylpolyene	12	-	5	1	-	3	-	-	1	-	-	-	2
	lassopeptide	8	4	-	1	-	-	-	2	-	-	-	-	1
	T1PKS***	7	3	2	1	-	-	-	-	-	-	-	-	1
	indole	5	1	-	3	1	-	-	-	-	-	-	-	-
	LAP****	4	-	-	1	1	-	1	-	-	-	-	-	1
	resorcinol	4	-	-	1	-	3	-	-	-	-	-	-	-
	N-acyl amino acids	4	-	1	1	-	-	-	-	-	-	-	-	2
	lanthipeptide	3	-	-	-	-	-	2	1	-	-	-	-	-
	TfuA-related	3	-	1	-	1	-	-	-	-	-	1	-	-
	hglE-KS*****	2	1	-	1	-	-	-	-	-	-	-	-	-
	phosphonate	2	1	1	-	-	-	-	-	-	-	-	-	-
	thiopeptide	2	1	-	1	-	-	-	-	-	-	-	-	-
	hserlactone	1	-	1	-	-	-	-	-	-	-	-	-	-
	phenazine	1	-	-	-	-	-	-	-	1	-	-	-	-
	sactipeptide	1	-	-	-	-	-	1	-	-	-	-	-	-
	ectoine	1	-	-	1	-	-	-	-	-	-	-	-	-
	siderophore	1	-	1	-	-	-	-	-	-	-	-	-	-
	ladderane	1	-	-	-	-	-	-	-	-	1	-	-	-
	oligosaccharide	1	1	-	-	-	-	-	-	-	-	-	-	-

5.4 Discussion

I studied frozen soil samples collected along a transect starting at the Greenland ice sheet (GrIS) edge and going further away ranging a variety of habitats and geochemical specific site conditions with the aim to analyze the cold-adapted community in this area, studying the DNA and cDNA microbial distribution in proglacial upper permafrost layer, to reconstruct high- and medium-quality MAGs and to create a database of cold-adapted predicted proteins. This was achieved thanks to the use of multiple sequencing approaches (i.e. metagenomics and metatranscriptomics) and technologies (i.e. Nanopore and Illumina) that led to the construction of a deep read coverage assembly.

The GrIS proglacial system showed an active (i.e. cDNA) and potentially active (i.e. DNA) consistent microbial community across sites (Figure A5.4) where only the DNA dataset significantly correlated with the site distance from the ice edge ($r = 0.27$). However, no trends between autotrophs and heterotrophs, usually observed in surface soil in proglacial successions, were observed (Fernández-Martínez et al., 2017; Liu et al., 2012; Schmidt et al., 2008). The encountered permafrost taxa were previously found in this environment (Jansson and Taş, 2014; Steven et al., 2006) and the most active organisms in our dataset, belonging to Acidobacteria, Actinobacteria, Verrucomicrobia, Chloroflexi (Figure 5.2), were previously identified as active at a wide range of sub-zero temperatures (Tuorto et al., 2014). Organisms that were predominant in the DNA but not in the cDNA datasets (e.g. Gammaproteobacteria or Firmicutes) could be in a wide range of physiological states. Organisms enter a state of dormancy when the environmental conditions are not favorable to their growth but can revert this process and increase their activity when the conditions become more favorable (Lebre et al., 2017; Jansson and Hofmockel, 2018). In the permafrost, organisms are present at different metabolic states (Steven et al., 2006; Jansson and Taş, 2014). Incubation experiments performed on permafrost soil have shown quick shifts in the microbial communities with the changing conditions showing how the microbial organisms in a dormant or low activity state are ready to activate when more favorable environmental conditions are present (Luláková et al., 2019). This suggests that microbial community shifts due to the permafrost thawing could be quick and that microbial positive feedback mechanism to the global warming, where microbial communities degrade the once trapped organic carbon, could significantly increase the emission of greenhouse gasses (e.g. CO_2 and CH_4) (Mondav et al., 2017; Schuur et al., 2015; Ernakovich and Wallenstein, 2015; MacKelprang et al., 2011). In this dataset, organisms belonging to the classes Bacteroidia, Cythophagia, Betaproteobacteria, Gammaproteobacteria and Chitinophagia also showed a significant positive correlation with the DOC concentration (Figure 5.3) showing how these taxa, for example, could potentially increase and benefit by the permafrost thawing.

The DNA dataset showed both a significant correlation with the site distance from the ice edge and the geochemical variables because it may reflect both the non-active community trapped in the dark and cold permafrost by its formation, being the permafrost an isolated and impermeable layer from the rest of the soil (Woo et al., 2008), and also the more active cold-adapted community that was shaped by environment selection factors set by this challenging environment (Malard et al., 2019; Ren et al., 2018). Where the challenges set by the frozen soil environment could have created a converged and ubiquitous community able to survive these environmental conditions in the broad variety of sampled sites, from thermokarst bogs to *Salix* heath.

The cDNA trends showed a stronger statistical correlation with the geochemical dataset compared to the DNA dataset (r equals to 0.40 and 0.22 for the cDNA and DNA dataset, respectively) and also more taxa in the cDNA dataset were found to be correlated to geochemical variables (Figure 5.3). The cDNA higher correlation showed how, locally, the organism activity is strongly conditioned by factors such as nutrients (e.g. sulfates and DOC) and ion concentration (e.g. magnesium, calcium and iron) where, for example, the latter can also directly influence and limit the microbial activity serving as enzyme cofactors (Miethke, 2013; Salama et al., 2020; Pasternak et al., 2010).

The construction of the assembly of this conserved frozen soil community retrieved 69 high- and medium- quality MAGs and 213 complete biosynthetic gene clusters (BGCs), highlighting the good assembly quality. The transcription of these gene clusters leads to secondary metabolite biosynthesis which constitutes a huge resource for antimicrobial and insecticide compounds (Weber et al., 2015; Castro et al., 2014). The BGCs found in this dataset comprised gene clusters for the biosynthesis of ribosomally synthesized peptides, such as the bacteriocins, nonribosomal peptides (NRP), synthetized by non-ribosomal peptide synthetase (NRPS), and also polyketides (PK), synthesized by Type I and III polyketide synthases (T1PKS and T3PKS). These compounds (i.e. bacteriocins, NRPs and PKs) are widely used in the pharmaceutical industries as antimicrobial compounds and drugs (Egan et al., 2017; Shen, 2003; Du and Lou, 2010).

The genomes of this frozen soil community were also used to create the CAPP database, a database of cold-adapted predicted proteins. The aim of this database is to provide cold-adapted functional protein sequences for a better-informed protein design and modeling and ultimately to facilitate protein engineering of cold-adapted enzymes. This database has also few added values. Firstly, it is possible to retrieve only protein sequences that were active at subzero temperatures thanks to their correspondences to the metatranscriptomic dataset. Secondary, because we also geochemically characterized each site (e.g. ion and nutrient concentrations), it is possible to trace back at which conditions these proteins were transcribed, helping to understand in which conditions a specific protein would potentially work.

As said above, the CAPP database, and similar databases, could be used in the protein engineering field to explore amino acid differences between psychrophilic and mesophilic proteins, and to take an informed approach towards single amino acid protein modifications to create proteins active at lower temperatures. It has been observed that cold-adapted proteins have overall a more flexible structure compared to proteins with higher temperatures in order to permits a higher structural flexibility at close-to-freezing point temperatures (Åqvist et al., 2017; Margesin and Collins, 2019). Some amino acids can have different stabilizing properties on protein structures where, for examples, some (e.g serine, threonine, asparagine, glutamine, histidine, tyrosine, tryptophan, aspartate, glutamate, arginine, proline and lysine) have been shown to facilitate protein stabilizing structures, such as hydrogen bonds and salt bridges, therefore reducing protein flexibility (Chao et al., 2020). However, amino acid sequences and substitutions between homologous proteins (i.e. cold-adapted and heat-adapted proteins) are highly protein and position specific where even one amino acid substitution can lead to protein malfunction and to unbalanced trade-off between protein activity and stability (Siddiqui, 2015; Siddiqui and Cavicchioli, 2006). This highlights how it is important to assume a protein-focused approach for protein modification.

When looking at the amino acid composition of the CAPP proteins compared to known psychrophilic and thermophilic proteins, it was evident how the clustering between cold and heat adapted organisms was only a secondary factor (Figure 5.4), and that the differential amino acid frequencies of these genomes were also shaped by other unidentified factors (e.g. other geochemical variables or taxonomic preferential amino acid usage).

For this reason, in this chapter I have looked at two different enzymes (i.e. β-galactosidase and polyphenol oxidase) and restricted comparisons within phyla to show how the enzymes of the CAPP database can be explored by comparing them to organisms isolated from temperate environments. This method allowed the exploration of a well-known enzyme such as β-galactosidase, which is well characterized for its use in molecular procedures (Serebriiskii and Golemis, 2000; Juers et al., 2012), and of polyphenol oxidase enzymes.

The tree constructed for both these enzymes formed consistent taxonomic clusters where occasionally sequences belonging to other taxa were found (Figure 5.5 and Figure 5.6). The latter may be a result of microbial lateral gene transfer or may represent an error in the contig annotation.

No common amino acid substitutions were observed in any of the enzyme protein positions (Table 5.5). However, in both enzymes the most common replaced amino acid from the nr database (i.e. mesophilic proteins) to the CAPP database (i.e. psychrophilic proteins) was valine. A higher presence of the amino acid valine in the protein structure was shown to increase protein stability (i.e. decrease structural flexibility) in several proteins (e.g. Okoniewska et al., 2000; Gryk et al., 1996; Tanaka et al., 1993). To clarify, this result does not mean that, overall, the content of valine was higher in nr compared to the CAPP proteins, but it simply indicates that the amino acid valine, which was shown to have stabilizing structural effects, was generally substituted from heat-adapted to cold-adapted protein homologous at certain positions.

In β-galactosidase protein positions 571, 651, 770 and 786, amino acids in the nr proteins were substituted mainly to alanine which has been showed to increase protein flexibility (Hespen et al., 2018; Kokubo et al., 2013).

The tendency of the amino acid valine to be substituted with other amino acids in the CAPP database and the tendency of the amino acid alanine to substitute others is an expected result where, sure enough, heat-adapted proteins are expected to have a more stable protein structure compared to the cold-adapted ones.

However, both valine and alanine did not correspond to univocal substitutions between psychrophilic and mesophilic proteins. Once again, this highlights how amino acid substitutions, aiming to a change of protein flexibility and activity, must be protein- and position-driven and how the construction of a cold-adapted protein database could help to conduct informed protein modifications and to create proteins with low enzymatic temperature optimum.

5.5 Conclusion

To conclude, thanks to the use of a hybrid approach where both Nanopore and Illumina technologies were combined, I was able to reconstruct a high reliable deep coverage assembly from permafrost active layer microbial communities of the GrIS proglacial system.

The cold-adapted community converged to similar taxonomic composition along a transect spanning 43 km and ranging different environments and did not show typical microbial trends of the proglacial systems but a more permafrost adapted community. However, whereas DNA profiles were conditioned by the distance from the ice edge, cDNA had higher correlation to geochemical trends showing how microbial communities are shaped by environmental conditions.

The soil cold-adapted assembly was also used to reconstruct high-quality MAGs, which will enrich the ENA MAG public repository, and to create a database of gene and protein sequences suitable to explore amino acid substitutions between known thermophilic, mesophilic and psychrophilic predicted proteins, called the CAPP database. This database is only a first step towards protein studies and further protein investigations, gene cloning and protein synthesis and enzymatic tests are needed. Whereas this approach may not be the best for the biomining of new enzymes and metabolic pathways, it is suitable to look at single amino acid substitutions and therefore to inform protein models that aim to lower the optimal enzymatic reaction temperature of mesophilic proteins, bringing many advantages in the industrial and bioremediation sectors.

Acknowledgements

I thank Dr. Gary Barker for all the help during the sampling activity and Annabel Summerfield for the help with the DNA and RNA extraction. Thanks to Dr. Fotis Sgouridis and Dr. Mahammad Rafiq for the geochemical analyses.

A modified version of this chapter is in preparation to be submitted to a peer-reviewed journal: 'Varliero, G., Rafiq, M., Sgouridis, F., Summerfield, A., Barker, G. A. L., Frozen soil to explore cold-adapted microbial trends and enzymes'.

Appendix

5.A Site characterization

5.A.1 Vegetation

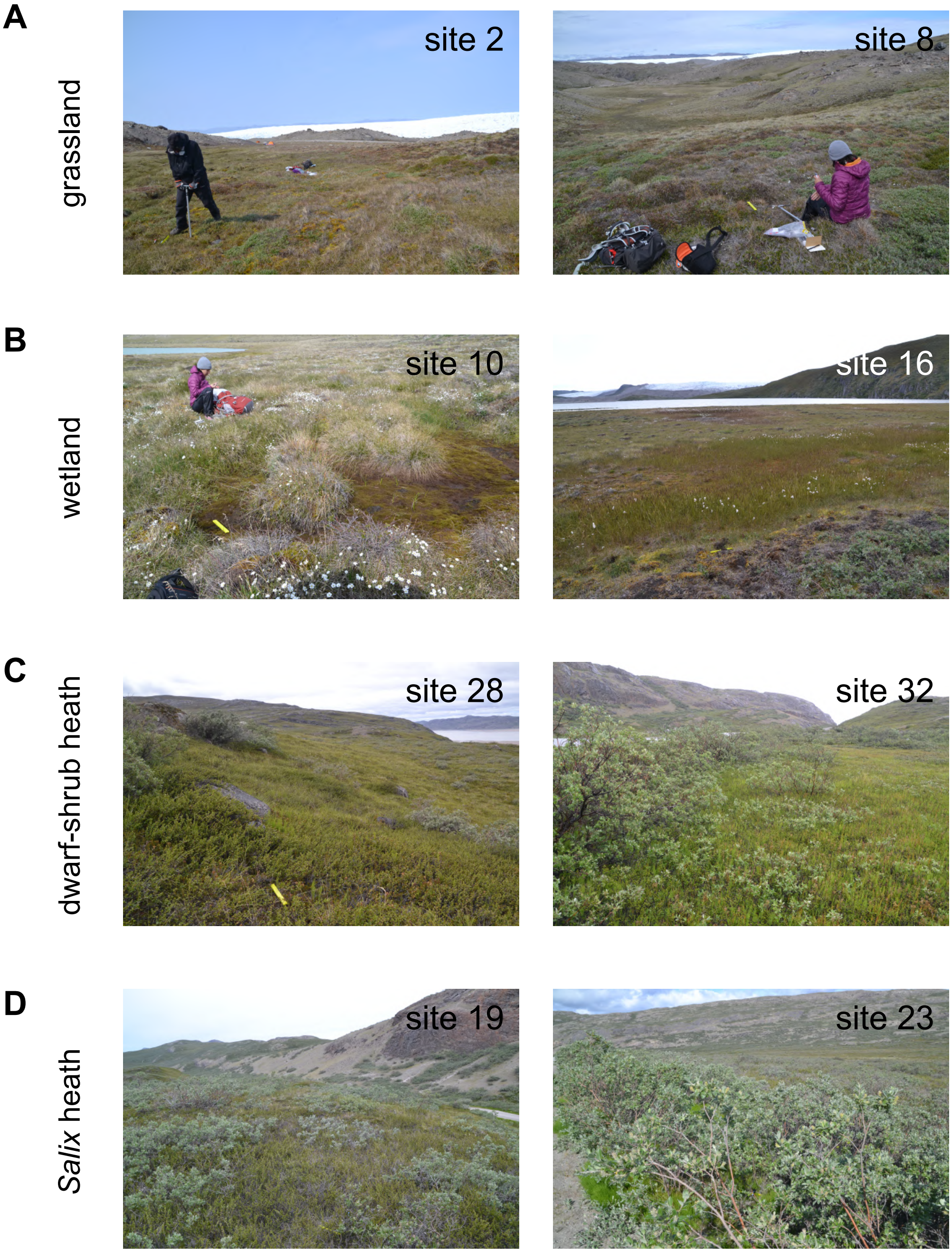


Figure A5.1: Site vegetation characterization. Sites defined as grassland were dominated by grass (e.g. *Calamagrostis langsdorfii*) and moss (A). Wetland was close to lakes, rivers and bogs where there is a dominance of grass, dwarf-shrubs, *Eriophorum scheuchzeri* and *Poa alpina* and *Poa pratensis* (B). The vegetation defined as dwarf-shrub heath is dominated by *Betula nana*, *Rhododendron lapponicum*, *Vaccinium uliginosum* and *Ledum palustre* (C); *Salix* heath is dominated by high-shrubs of the genus *Salix* (D).

5.A.2 Geochemical analyses

Soil samples were manually homogenised, and all visible roots were separated by sieving before the subsequent analyses were performed to the sieved fraction (< 2 mm). Root biomass and soil moisture content were determined gravimetrically by drying at 105 °C for 24 hours. Organic matter content was determined by loss on ignition after furnacing 1 g of dried soil at 550 °C for 4 hours. Following treatment of soils with loss on ignition (to remove organic matter), the absolute particle size distribution of the mineral soil fraction was determined with optical laser diffraction using a MS3000 Mastersizer (Malvern Instruments Ltd., UK).

Field moist soils (1 g) were extracted at a ratio of 5:1 with 5 mL 2M KCl for the determination of exchangeable ammonium (NH_4^+), while 5 g were extracted with 25 mL of deionised water for the determination of dissolved nitrate (NO_3^-), phosphate (PO_4^{3-}), silicon (Si), total iron (Fe), major anions and cations and dissolved organic carbon (DOC). The soil slurries were continuously shaken on a reciprocating shaker at 200 rpm for 1 hour before being centrifuged at 5000 rpm for 10 minutes followed by filtration with 0.22 µm 25 mm PES syringe filters. Ammonium was analysed spectrophotometrically on a Gallery Plus Automated Photometric Analyser (Thermo Fisher Scientific, UK) using a salicylate-hypochlorite alkaline reaction method measured at 660 nm, nitrate using a hydrazine-sulfanilamide reaction method measured at 540 nm, phosphate using the molybdenum blue method measured at 880 nm, silicon using an ammonium molybdate – ascorbic acid reaction method measured at 700 nm and total iron using a hydroxylamine-ferrozine reaction method measured at 562 nm. The limits of detection were 0.01 mg N, P, Si or Fe L^{-1}, the samples were blank corrected, while the precision as a relative standard deviation (RSD) was < 2%. Soil pH was also measured in the deionised water extracts using the Gallery Plus built in probe (calibrated 4–7 pH).

DOC concentrations were quantified using a Shimadzu TOC-L Organic Carbon Analyzer, with a high salinity module. Non-purgeable organic carbon (NPOC) was measured after acidification of samples with 9M H_2SO_4 and catalytic combustion (680 °C) of dissolved organic carbon to carbon dioxide, which was then measured by infrared absorption. The limit of detection was 0.01 mg C L^{-1}, the samples were blank corrected, while the precision as a relative standard deviation (RSD) was < 5%. For the simultaneous determination of total dissolved nitrogen (TDN) and phosphorus (TDP), an aliquot (5 mL) of the deionised water soil extracts was digested using the potassium persulfate oxidation method of Johnes and Heathwaite (1992) modified for the CEM MarsXpress microwave digestion unit. TDN and TDP were measured colorimetrically as nitrate and phosphate, respectively as above. Dissolved organic nitrogen (DON) and phosphorus (DOP) were calculated by difference (DON = TDN - NO_3^- - NH_4^+ and

DOP = TDP - PO_4^{3-}).

Major anions and cations were measured simultaneously in the deionised water soil extracts using an ICS5000 ion chromatograph (Thermo Fisher Scientific, UK). Anions were separated isocratically on an AS11-HC 2-mm column at 0.25 ml min^{-1} flow rate using 24 mM KOH eluent. Cations were separated isocratically on an CS12 2 mm column at 0.25 mL min^{-1} flow rate using 20 mM MSA eluent. The limit of detection was 0.001 mg L^{-1} for all measured anions and cations, the samples were blank corrected, while the precision as a relative standard deviation (RSD) was < 2%.

5.A.3 Geochemical data

Table A5.1: Grain size distribution across the different sites.

Sites	**Grain size**				
	< 2 µm	**2–8 µm**	**8-15 µm**	**15–50 µm**	**> 50 µm**
1	0.19 ± 0.08	5.60 ± 1.51	9.02 ± 2.59	45.45 ± 4.01	39.72 ± 7.96
2	0.08 ± 0.07	6.38 ± 0.49	11.29 ± 1.05	46.63 ± 3.11	35.62 ± 3.52
3	0.09 ± 0.02	5.79 ± 0.60	10.01 ± 1.57	48.96 ± 2.42	35.13 ± 4.08
4	0.02 ± 0.01	4.57 ± 0.80	8.36 ± 0.92	46.82 ± 3.67	40.23 ± 5.16
5	0.09 ± 0.04	5.31 ± 0.19	10.8 ± 1.35	50.83 ± 4.81	32.97 ± 6.05
6	0.05 ± 0.05	4.79 ± 1.00	8.27 ± 1.18	44.76 ± 0.90	42.14 ± 2.73
7	0.03 ± 0.04	6.02 ± 1.33	9.63 ± 1.03	43.75 ± 3.20	40.58 ± 5.21
8	0.06 ± 0.02	4.90 ± 0.41	8.85 ± 0.75	44.18 ± 1.93	42.01 ± 2.86
9	0.04 ± 0.05	5.39 ± 1.16	9.28 ± 2.22	46.50 ± 2.28	38.80 ± 5.49
10	0.02 ± 0.01	4.68 ± 0.14	7.19 ± 0.16	40.91 ± 1.31	47.20 ± 1.57
11	0.05 ± 0.07	5.58 ± 0.51	9.71 ± 1.36	48.31 ± 2.23	36.34 ± 3.93
12	0.04 ± 0.05	4.62 ± 0.99	7.19 ± 1.09	42.11 ± 2.82	46.03 ± 4.93
13	0.03 ± 0.05	5.35 ± 1.09	8.73 ± 0.81	43.64 ± 4.26	42.25 ± 3.20
14	0.04 ± 0.05	4.71 ± 1.86	8.15 ± 3.28	46.98 ± 3.01	40.12 ± 8.19
15	0.03 ± 0.03	3.95 ± 0.17	5.91 ± 0.29	44.33 ± 1.19	45.77 ± 0.75
16	0.00 ± 0.00	3.56 ± 0.61	6.08 ± 1.18	43.46 ± 1.28	46.91 ± 2.01
17	0.00 ± 0.01	2.88 ± 0.69	4.58 ± 1.24	40.15 ± 1.75	52.38 ± 3.15
18	0.03 ± 0.04	3.26 ± 1.44	5.25 ± 3.16	37.23 ± 4.69	54.23 ± 9.34
19	0.15 ± 0.11	4.71 ± 1.19	6.33 ± 1.78	43.70 ± 6.73	45.12 ± 9.76
20	0.01 ± 0.02	3.60 ± 0.16	5.21 ± 0.36	40.92 ± 1.80	50.25 ± 1.33
21	0.00 ± 0.00	0.76 ± 0.36	0.59 ± 0.64	6.29 ± 5.20	92.35 ± 6.20
22	0.02 ± 0.02	3.48 ± 1.07	5.17 ± 1.98	41.04 ± 4.85	50.29 ± 7.23
23	0.08 ± 0.05	4.91 ± 1.14	7.27 ± 2.85	46.94 ± 4.33	40.80 ± 8.15
24	0.15 ± 0.17	5.76 ± 1.29	7.65 ± 1.71	41.35 ± 4.34	45.09 ± 7.14
25	0.03 ± 0.01	2.68 ± 0.49	3.31 ± 0.59	19.95 ± 4.92	74.03 ± 5.88
26	0.12 ± 0.09	5.16 ± 0.93	8.87 ± 2.35	46.27 ± 9.49	39.59 ± 12.81
27	0.18 ± 0.08	5.99 ± 0.86	9.10 ± 0.97	45.58 ± 2.73	39.14 ± 2.96
28	0.33 ± 0.16	6.10 ± 2.18	7.2 ± 3.42	32.02 ± 14.29	54.34 ± 19.53
29	0.49 ± 0.21	7.59 ± 0.27	9.81 ± 2.45	40.77 ± 15.75	41.34 ± 17.74
30	0.56 ± 0.29	9.29 ± 2.46	11.76 ± 0.65	45.72 ± 3.47	32.67 ± 2.40
31	0.27 ± 0.08	7.79 ± 1.38	9.77 ± 1.07	44.39 ± 4.10	37.78 ± 6.54
32	0.10 ± 0.03	9.08 ± 0.88	13.18 ± 0.47	50.07 ± 3.64	27.57 ± 2.44
33	0.35 ± 0.05	8.20 ± 0.98	11.90 ± 2.12	48.46 ± 2.26	31.09 ± 4.96
34	0.70 ± 0.42	10.47 ± 3.52	11.21 ± 3.48	38.45 ± 7.85	39.18 ± 12.29

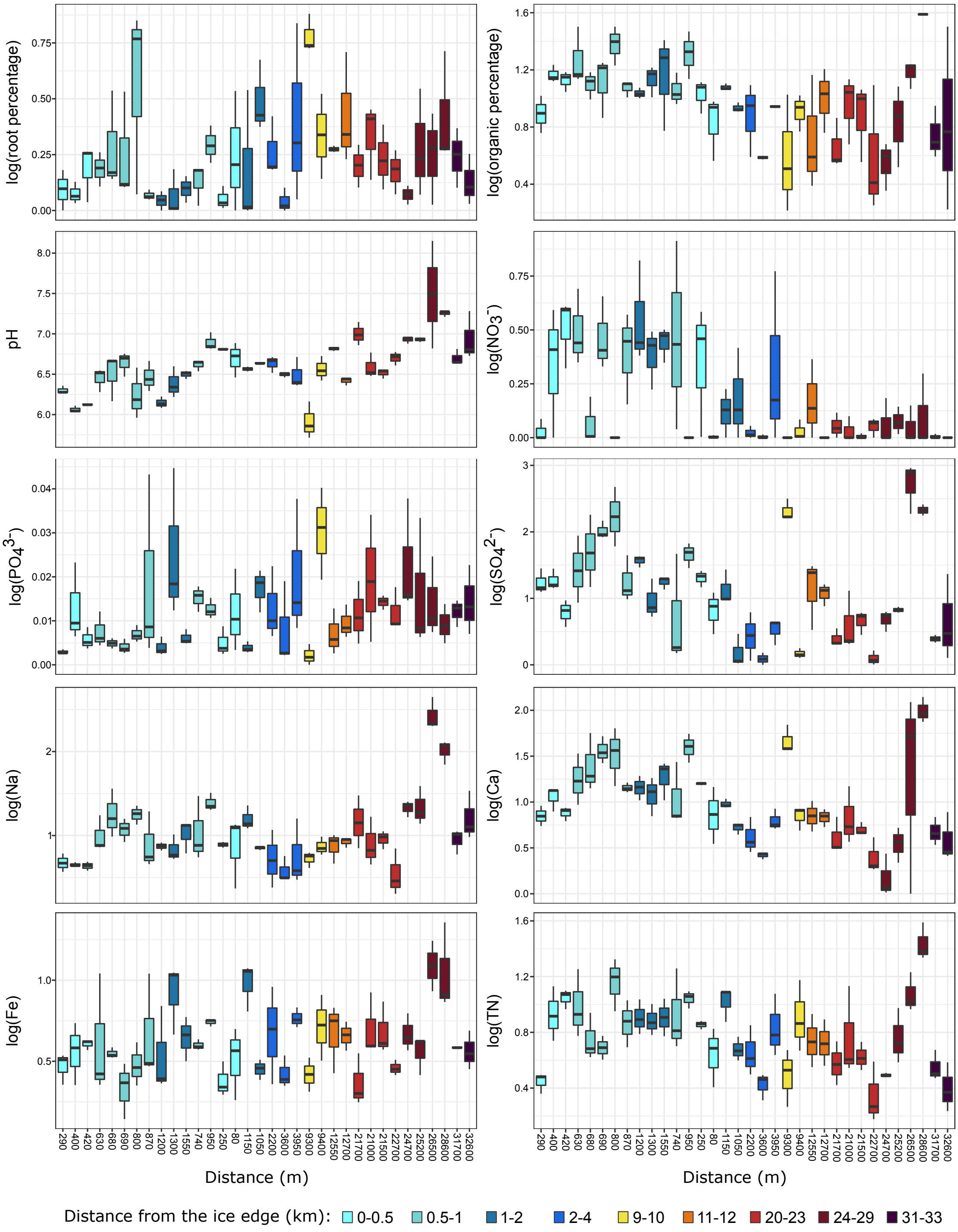


Figure A5.2: Geochemical trends across the proglacial sampling sites. The figure reports only the variables that were significantly (p-value < 0.05) correlated to site distance from the ice edge in the Spearman's rank correlation test (results are shown in Table 5.2). All the values, except from pH, are expressed as $\log_{10}(x + 1)$.

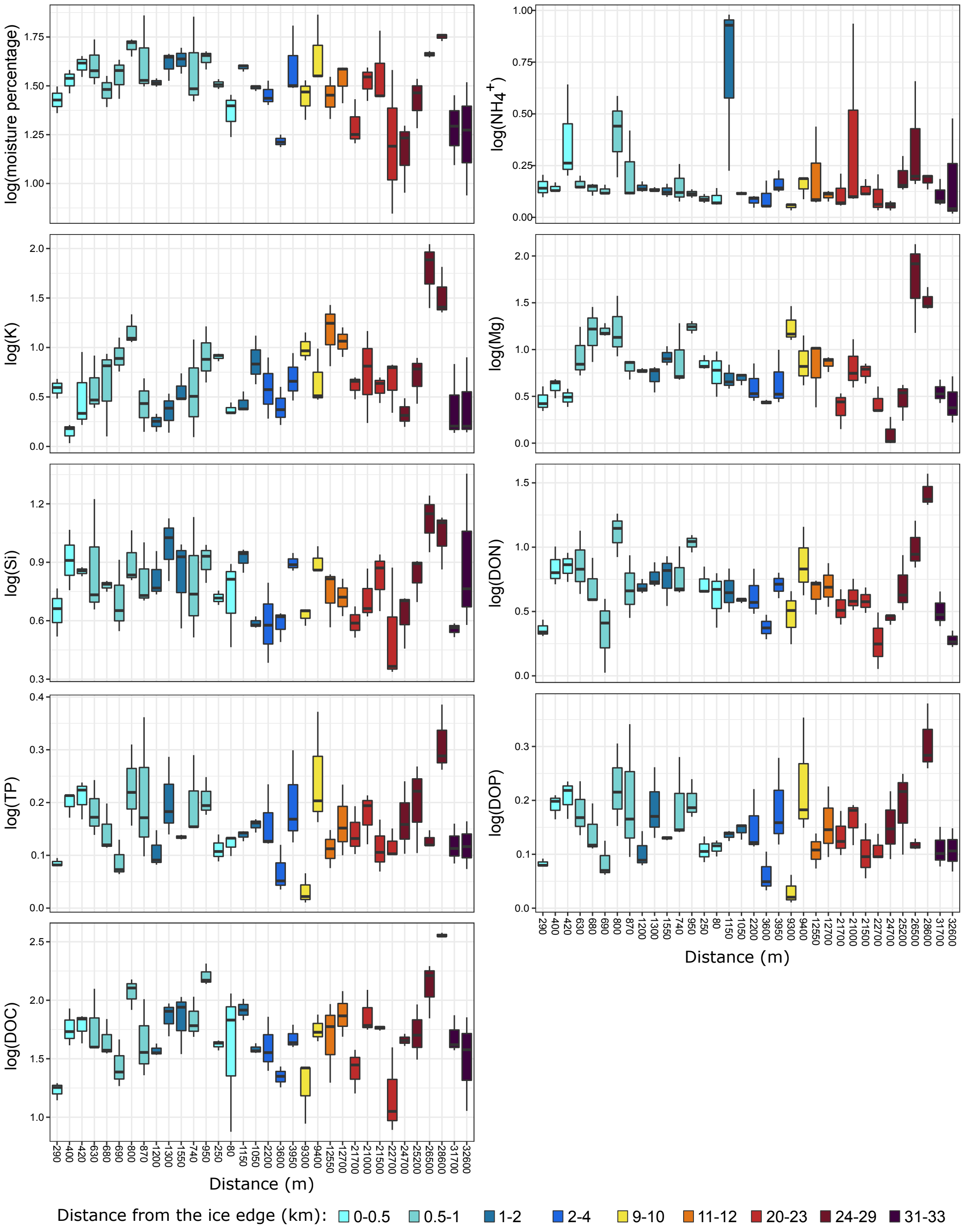


Figure A5.3: Geochemical trends across the proglacial sampling sites. The figure reports only the variables that were not significantly (p-value ≥ 0.05) correlated to site distance from the ice edge in the Spearman's rank correlation test (results are showed in Table 5.2). All the values are expressed as $\log_{10}(x + 1)$.

5.A.4 Taxonomic trends

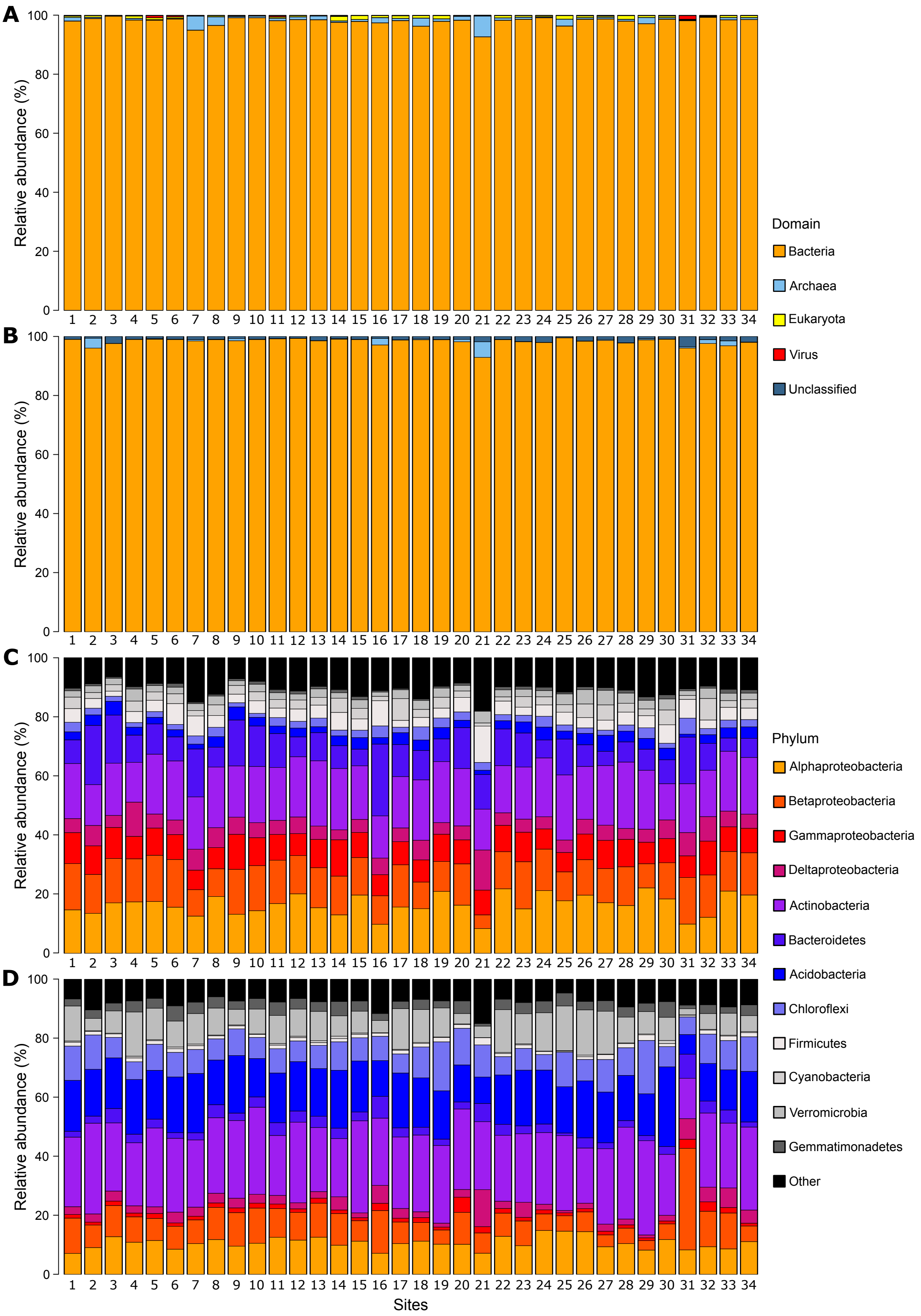


Figure A5.4: Taxonomic DNA and cDNA composition at each site at the domain-level for DNA (A) and cDNA (B), and at the phylum-level (and Proteobacteria classes) for DNA (C) and cDNA (D).

5.B Psychrophilic and thermophilic genomes

Table A5.2: NCBI genome accession number used to explore the amino acid usage from psychrophilic and thermophilic organisms in Figure 5.4.

NCBI accession	Species	Organism
GCF_000007305.1	*Pyrococcus furiosus*	thermophiles
GCF_001295365.1	*Parageobacillus thermoglucosidasius*	
GCF_000091545.1	*Thermus thermophilus*	
GCF_000204925.1	*Metallosphaera cuprina*	
GCF_000022325.1	*Caldicellulosiruptor bescii*	
GCF_000015865.1	*Hungateiclostridium thermocellum*	
GCF_001610955.1	*Geobacillus thermoleovorans*	
GCF_000020965.1	*Dictyoglomus thermophilum*	
GCF_000024425.1	*Meiothermus ruber*	
GCF_900129205.1	*Thermomonas hydrothermalis*	
GCF_000299435.1	*Exiguobacterium antarcticum*	psychrophiles
GCF_000764185.1	*Colwellia psychrerythraea*	
GCA_001750085.1	*Fragilariopsis cylindrus*	
GCF_004367585.1	*Paenisporosarcina antarctica*	
GCF_001955735.1	*Rhodoferax antarcticus*	
GCF_004365915.1	*Cryobacterium psychrophilum*	
GCF_000967895.1	*Oleispira antarctica*	
GCF_000153225.1	*Polaribacter irgensii*	
GCF_000012305.1	*Psychrobacter arcticus*	
GCF_001647715.1	*Pseudomonas antarctica*	

5.C MAGs

Table A5.3: Medium-quality MAGs. The table reports MAG completeness, contamination and strain heterogeneity as calculated from CheckM. It also reports the number of contigs in each MAG and the taxonomic classification as phylum, class, order, family and genus. The taxonomic classification was reported only when more than the 70% of the contigs where assigned to the same taxon.

MAGs	**Completeness** (%)	**Contamination** (%)	**Strain heterogeneity**	**Contig number**	**Taxonomy**
MAG-10	98.0	8.9	45.8	36	Proteobacteria, Betaproteobacteria
MAG-11	99.0	8.1	7.1	8	Actinobacteria, Actinobacteria
MAG-12	96.6	8.0	47.1	59	-
MAG-13	95.7	6.5	62.5	19	Chloroflexi
MAG-14	93.1	7.2	16.7	86	Actinobacteria, Actinobacteria
MAG-15	90.0	7.1	26.7	73	Verrucomicrobia
MAG-16	89.8	7.5	6.2	51	Actinobacteria, Actinobacteria
MAG-17	89.1	5.3	22.2	67	Chloroflexi
MAG-18	86.5	5.9	57.1	56	Actinobacteria, Actinobacteria
MAG-19	85.9	3.4	25.0	106	Acidobacteria
MAG-20	85.4	4.8	40.0	96	Thaumarchaeota
MAG-21	84.8	3.3	0.0	98	Actinobacteria, Actinobacteria
MAG-22	84.0	8.4	36.4	60	Gemmatimonadetes
MAG-23	82.6	9.6	37.9	82	Actinobacteria, Actinobacteria
MAG-24	81.9	6.9	45.8	70	Actinobacteria, Actinobacteria
MAG-25	81.1	9.5	53.8	31	Actinobacteria, Actinobacteria
MAG-26	79.2	3.7	0.0	32	-
MAG-27	79.2	9.3	70.0	25	Chloroflexi
MAG-28	78.6	3.9	0.0	48	Actinobacteria, Actinobacteria
MAG-29	78.4	1.8	0.0	57	*Candidatus* Cryosericota, *Ca.* Cryosericia, *Ca.* Cryosericales, *Ca.* Cryosericaceae, *Ca.* Cryosericum
MAG-30	77.7	7.4	0.0	70	-
MAG-31	77.7	5.4	0.0	18	Nitrospirae, Nitrospira, Nitrospirales, Nitrospiraceae, *Nitrospira*
MAG-32	77.7	1.1	0.0	24	-
MAG-33	76.8	8.9	68.7	42	Proteobacteria, Alphaproteobacteria, Sphingomonadales, Sphingomonadaceae, *Sphingomonas*
MAG-34	74.1	8.8	72.7	26	Chloroflexi
MAG-35	74.0	4.4	50.0	97	Gemmatimonadetes
MAG-36	73.8	7.5	52.6	63	Actinobacteria, Actinobacteria
MAG-37	73.5	3.4	0.0	47	Actinobacteria
MAG-38	72.5	3.3	26.7	61	Proteobacteria, Alphaproteobacteria, Rhizobiales, Hyphomicrobiaceae, *Hyphomicrobium*
MAG-39	71.5	6.4	42.9	46	Proteobacteria, Betaproteobacteria
MAG-40	71.2	0.5	100.0	31	Chloroflexi
MAG-41	70.9	6.1	53.8	167	Acidobacteria
MAG-42	70.3	1.8	80.0	45	Bacteroidetes, Chitinophagia, Chitinophagales, Chitinophagaceae
MAG-43	69.3	3.8	50.0	51	Acidobacteria
MAG-44	68.7	2.2	16.7	24	Proteobacteria, Alphaproteobacteria, Rhizobiales, Bradyrhizobiaceae, *Bradyrhizobium*
MAG-45	67.7	2.5	50.0	52	Bacteroidetes
MAG-46	67.3	9.5	0.0	77	Actinobacteria, Actinobacteria
MAG-47	66.5	6.1	28.6	188	Acidobacteria, Acidobacteriia, Bryobacterales, Solibacteraceae
MAG-48	65.9	7.3	28.1	46	Proteobacteria, Alphaproteobacteria, Rhizobiales
MAG-49	65.2	1.4	60.0	77	Proteobacteria, Deltaproteobacteria
MAG-50	64.6	3.4	40.0	60	Acidobacteria
MAG-51	64.4	0.5	0.0	22	Proteobacteria, Betaproteobacteria
MAG-52	64.0	5.8	33.3	79	-
MAG-53	64.0	0.0	0.0	1	*Candidatus* Saccharibacteria
MAG-54	62.8	3.0	50.0	39	Verrucomicrobia
MAG-55	61.7	0.0	0.0	77	-
MAG-56	60.3	9.7	23.1	199	-
MAG-57	59.4	5.2	0.0	63	Actinobacteria, Actinobacteria
MAG-58	56.7	0.8	0.0	45	*Candidatus* Rokubacteria
MAG-59	55.5	0.2	0.0	72	Acidobacteria
MAG-60	54.3	7.5	57.7	93	Proteobacteria, Alphaproteobacteria
MAG-61	54.2	7.1	68.7	53	Proteobacteria, Alphaproteobacteria, Rhizobiales
MAG-62	53.7	3.6	8.3	75	Bacteroidetes
MAG-63	53.2	5.5	59.1	174	Bacteroidetes, Chitinophagia, Chitinophagales
MAG-64	53.0	5.7	16.7	106	Acidobacteria
MAG-65	52.9	0.4	0.0	33	Proteobacteria, Epsilonproteobacteria, -, -, *Sulfurovum*
MAG-66	52.4	9.4	76.9	87	Actinobacteria, Actinobacteria
MAG-67	51.8	4.0	41.2	46	Actinobacteria, Actinobacteria
MAG-68	51.7	7.9	0.0	67	Chloroflexi
MAG-69	50.5	2.6	33.3	204	Acidobacteria

5.D Tree sequences

Sequence accession numbers used to construct phylogenetic tree for β-galactosidase (Figure 5.5):

CAPP database: EHDLDABF_45316, EHDLDABF_64722, EHDLDABF_91228, EHDLDABF_91281, EHDLDABF_95383, EHDLDABF_95401, EHDLDABF_166629, EHDLDABF_240844, EHDLDABF_267849, EHDLDABF_292543, EHDLDABF_318823, EHDLDABF_331561, EHDLDABF_347189, EHDLDABF_360348, EHDLDABF_371318, EHDLDABF_407185, EHDLDABF_430994, EHDLDABF_468492, EHDLDABF_501598, EHDLDABF_566100, EHDLDABF_577598, EHDLDABF_659512, EHDLDABF_837302, EHDLDABF_870477, EHDLDABF_877154, EHDLDABF_877694, EHDLDABF_913836, EHDLDABF_931980, EHDLDABF_950071, EHDLDABF_970631, EHDLDABF_977904, EHDLDABF_1009044, EHDLDABF_1021885, EHDLDABF_1086177, EHDLDABF_1086178, EHDLDABF_1097039, EHDLDABF_1221760, EHDLDABF_1262834, EHDLDABF_1267799, EHDLDABF_1267902, EHDLDABF_1323377, EHDLDABF_1346125, EHDLDABF_1357884, EHDLDABF_1361164, EHDLDABF_1366487, EHDLDABF_1406383, EHDLDABF_1470756, EHDLDABF_1493494, EHDLDABF_1503156, EHDLDABF_1532113, EHDLDABF_1542948, EHDLDABF_1545466, EHDLDABF_1582646, EHDLDABF_1599313, EHDLDABF_1599357, EHDLDABF_1613950, EHDLDABF_1673633, EHDLDABF_1707264, EHDLDABF_1729044, EHDLDABF_1732353, EHDLDABF_1745514, EHDLDABF_1757123, EHDLDABF_1791209, EHDLDABF_1822912, EHDLDABF_1822913, EHDLDABF_1822984, EHDLDABF_1823430, EHDLDABF_1843861, EHDLDABF_1975752, EHDLDABF_2083605, EHDLDABF_2099415, EHDLDABF_2119990, EHDLDABF_2120441, EHDLDABF_2121330, EHDLDABF_2145751, EHDLDABF_2217339, EHDLDABF_2261931, EHDLDABF_2266289, EHDLDABF_2299159, EHDLDABF_2466288, EHDLDABF_2477296, EHDLDABF_2527918, EHDLDABF_2594762, EHDLDABF_2606570, EHDLDABF_2611901, EHDLDABF_2679847, EHDLDABF_2716359, EHDLDABF_2716360, EHDLDABF_2766648, EHDLDABF_2792713, EHDLDABF_2807891, EHDLDABF_2893601, EHDLDABF_2925164, EHDLDABF_2929874, EHDLDABF_2929916, EHDLDABF_2957101, EHDLDABF_2997334, EHDLDABF_3058317, EHDLDABF_3058332, EHDLDABF_3107198

nr database: WP_023697273.1, PYS72254.1, WP_127508724.1, SIS00308.1, HAH24410.1, OQY95594.1, OLD25772.1, PYV70045.1, RWB75876.1, WP_050061565.1, PYS64427.1, TME01883.1, WP_114790757.1, MRS01668.1, TMD09167.1, PYV47517.1, WP_017263926.1, PYT84220.1, OJV99997.1, TMI71839.1, SEP43528.1, BBE16927.1, WP_127121645.1, PYQ39194.1, OJU48565.1, TIT19796.1, MZP66540.1, PWT79888.1, WP_026256956.1, TMC54283.1, SFF79344.1, WP_052273167.1, TMB97468.1, OFY69781.1, WP_113952081.1, RPJ81704.1, WP_104841375.1, PYV70391.1, REF35853.1, WP_109486978.1, PYX38256.1, WP_089915368.1, GEC36186.1, WP_127613867.1, WP_064244200.1, WP_158912181.1, WP_090652988.1, WP_133331184.1, PYV69121.1, WP_158910847.1, WP_054281808.1, WP_114778600.1, RYY85798.1, WP_089919784.1, OJU53914.1, TME10281.1, HAH22220.1, TMI96660.1, TMD55350.1, MTK54465.1, MZP66850.1, WP_144888490.1, PYV70416.1, WP_012706331.1, WP_158749341.1, TME18627.1, PYV46009.1, WP_127664169.1, WP_134150408.1, RYY66422.1, RYX82899.1, WP_127515784.1, OLC89633.1, WP_121812171.1, WP_128770900.1, WP_123489692.1, PYV80412.1, RYY39133.1, WP_119408651.1, TMG58523.1, TIN91115.1, TDQ11067.1, WP_014266677.1, TMI97840.1, PYS36517.1, WP_013850251.1, TMC79922.1, WP_131852161.1, WP_103787511.1, PYX66750.1, WP_096351857.1, SCW80931.1, WP_115851867.1, TMD30386.1, WP_129132668.1, WP_092656998.1, WP_057935181.1, TAL59189.1, WP_121228762.1, PYX82779.1, WP_023712995.1, TMD33564.1, WP_027042739.1, WP_128531994.1, WP_048907999.1, PYR55679.1, PYS97483.1, OFY64085.1, HAH24005.1, WP_144845657.1, WP_106808516.1, OJV42166.1, WP_158818890.1, WP_026442485.1, PYT04137.1, TMF96728.1, GET31153.1, HAL83229.1, OLC22105.1, OFY63939.1, PYT04509.1, WP_133575235.1, WP_015242746.1, TMI93442.1, TDW48136.1, PYS79983.1, HAX94450.1, WP_037451808.1, NBB31009.1, WP_131557281.1, PYT59732.1, WP_114206465.1, TMC81361.1, WP_136826256.1, WP_097527037.1, WP_090533197.1, RZJ51477.1, TMF00975.1, OFX36309.1, MZQ50475.1, OLB89323.1, PYX30768.1, HAF24895.1, TMG14717.1, PYS81685.1, WP_055878653.1, HCX31060.1, WP_023767869.1, WP_150169169.1, WP_146790111.1, WP_127773716.1, WP_027031004.1, PYY11344.1, TME49682.1, OLE85709.1, PYS60469.1, WP_007803138.1, HBZ22309.1, PYS85599.1, TMF69965.1, TMF40691.1, WP_072836659.1, WP_073488532.1, PYU04614.1, PYX56774.1, OLB86806.1, TME14957.1, WP_131556827.1, WP_158827323.1, OGO55307.1, TMG23361.1, PYY17466.1, TME80271.1, WP_020708282.1, HBZ20649.1, WP_158820806.1, WP_153494913.1, RYY23638.1, WP_020479467.1, HBC79891.1, WP_121352668.1, WP_127597254.1, KIC94338.1, PYS51651.1, WP_131029608.1, WP_140572075.1, PLS77592.1, TMD01836.1, WP_123170581.1, PYV80970.1, PYY11011.1, WP_127676860.1, OLB18064.1, WP_141509533.1, HBB97186.1, WP_114940644.1, OLD28577.1, HAO77450.1, PYS74568.1, TME55635.1, WP_127668496.1, TJV40165.1, PYV80319.1, HAF18197.1, WP_090465134.1, WP_078063451.1, WP_133270437.1, TMC68379.1, WP_017276349.1, TMC42996.1, PYS83182.1, WP_020715497.1, WP_003527974.1, WP_141815845.1, HAF14031.1, PYV65995.1, PYT84404.1, WP_116252546.1, WP_014530059.1, PYY06821.1, HBC79922.1, WP_083437941.1, TMG23181.1, WP_109931595.1, TMF21141.1, HBC78998.1, PYY07516.1, WP_131998744.1, TME20193.1, TMI66028.1, PYS64816.1, TJV06621.1, SPF42223.1, PYX53713.1, PYU03259.1, WP_129003026.1, WP_084192346.1, NBP68606.1, WP_091142190.1, WP_158275322.1, PYQ45410.1, TMI69791.1, WP_134336876.1, WP_040120938.1, PWT89885.1, WP_027049018.1, TME15395.1, PYV54013.1, RYZ19055.1, WP_036984935.1, WP_147051973.1, WP_023806641.1, PYT21661.1, WP_023735738.1, WP_023689291.1, OLC49625.1, PYS92344.1, WP_109607543.1, OLC27975.1, WP_127640216.1, SDR72641.1, RYZ22059.1, HAQ17943.1, WP_092881809.1, WP_066406228.1, WP_068705763.1, WP_027167073.1, PYX80291.1, WP_097540432.1, TMI80668.1

Sequence accession numbers used to construct phylogenetic tree for polyphenol oxydase (Figure 5.6):

CAPP database: EHDLDABF_70248, EHDLDABF_70605, EHDLDABF_95264, EHDLDABF_224291, EHDLDABF_262193, EHDLDABF_265054, EHDLDABF_297465, EHDLDABF_430395, EHDLDABF_436066, EHDLDABF_436113, EHDLDABF_590943, EHDLDABF_645602, EHDLDABF_829503, EHDLDABF_888330, EHDLDABF_1354491, EHDLDABF_1632464, EHDLDABF_1632481, EHDLDABF_1639228, EHDLDABF_1779284, EHDLDABF_1848849, EHDLDABF_1927488, EHDLDABF_1944653, EHDLDABF_2146878, EHDLDABF_2247588, EHDLDABF_2280476, EHDLDABF_2306856, EHDLDABF_2327830, EHDLDABF_2405424, EHDLDABF_2458255, EHDLDABF_2469082, EHDLDABF_2534830, EHDLDABF_2830123, EHDLDABF_3096181

nr database: MSO29417.1, PYS72837.1, PYS43783.1, HAF18775.1, OGR24435.1, OLE18990.1, PYR62836.1, HAF12435.1, MSO68536.1, WP_068015720.1, TMD33946.1, WP_115516831.1, TME56079.1, PYS66576.1, TMJ61823.1, TMF44238.1, TMJ72845.1, WP_115694578.1, TMJ06204.1, MSP45182.1, WP_137043544.1, PYO50958.1, TMJ76533.1, WP_011685509.1, OJY44468.1, TMB89188.1, PWT82990.1, TMJ36258.1, PYS74442.1, HBB94777.1, TMC98478.1, TMK11227.1, TMK11950.1, TAN32529.1, MSO83078.1, PYO08029.1, HCU12400.1, MPZ38244.1, WP_106859038.1, TMJ84751.1, TMD89804.1, MSO45167.1, PZR72320.1, HBB87388.1, PYS34685.1, WP_056910038.1, TMH04654.1, PYS56382.1, TMJ94997.1, PYS21966.1, OLB73649.1, PYO77657.1, TMJ90682.1, TMK33988.1

Chapter 6

General discussion

6.1 General conclusions

Understanding how environmental microbial communities are shaped and structured is pivotal to comprehend how the environment works and is influenced by biological processes. Global dispersion, diversified metabolisms and high mutation rates put microorganisms in a central position in the biogeochemical cycles and at the base of the food chains, where phototrophic and lithotrophic organisms enrich the environment with easily bioavailable nutrients (Barton et al., 2010; Singh et al., 2010; Kivisaar, 2003; Ram and Hadany, 2014). The scientific world became aware of this microbial biogeographic patterns, their ubiquity and diversity thanks to the advent of sequencing technologies, which led to an overall substitution of isolation and culturing techniques and to an increasing use of DNA and RNA environmental sequencing for the study of microbial diversity (Gutleben et al., 2018; Shokralla et al., 2012). Arctic exploration especially benefited from these technological innovations as the isolation and culturing of polar organisms pose extra challenges due to the difficulty of reproducing environmental conditions in laboratory settings.

Arctic communities must be studied because their habitat have been undergoing an unprecedented change due to global warming (Meier et al., 2007; King et al., 2019; Braun et al., 2019). These communities are going to shift in relation to the changes in the environmental conditions, possibly initiating positive feedback to the global warming itself, and a better understanding of the polar microorganisms now will certainty allow a better understanding of future environmental shifts (Williamson et al., 2019; Makhalanyane et al., 2015; Wang et al., 2020). Furthermore, microbial communities living in this habitat are of great interest because, thanks to their adaptation to the environment, they represent a wide reservoir for the bioprospecting of new compounds and cold-adapted enzymes (Segawa et al., 2013; Bell et al., 2013; Mangiagalli et al., 2020; Collins and Margesin, 2019).

In my thesis I have explored different Arctic environment microbial datasets, from glacial environments to proglacial systems and permafrost. The overall aim is to extend our knowledge on this environment, showing also how different environmental microbial communities can be bioinformatically explored, and implementing new pipelines, software and databases where needed.

I implemented PhyloPrimer, an online software to design primers amplifying taxon-specific genes in environmental samples which was successfully validated with the design of primers targeting different *Streptococcus* organisms (Chapter 2). Using a new sampling method and an amplicon metagenomic approach, I then compared, for the first time, microbial communities sampled from clear (i.e. frozen englacial water channel) and cloudy ice (i.e. meteoric ice) showing that microbial communities clustered in relation to the microbial distribution on the ice surface but also to the ice types, suggesting an important role of the englacial channels in the development of a glacial microbial community and in the microbial dispersion within the glacier (Chapter 3). In Chapter 4 I presented LongMeta, a pipeline to analyze both gene and taxon content in whole shotgun metagenomic data which I used to explore rock weathering and nitrogen fixation processes in proglacial systems. These processes were shown to be driven by a plurality of organisms where, however, different diazotrophic and rock weathering organisms were present at different stages of the soil succession. Finally, I created a reliable assembly using whole shotgun data obtained with Illumina and Nanopore sequencing technologies which allowed me to characterize the cold-adapted microbial communities of the upper permafrost layer and to create a solid database of cold-adapted enzymes (CAPP database) to inform, ultimately, bioengineering studies for the design of cold-adapted industrially relevant enzymes, such as β-galactosidase and polyphenol oxidase (Chapter 5).

6.2 Bioinformatics approaches

Each chapter presented and implemented a different approach for the analyses of microbial communities. Chapter 2 and 3 focused on an amplicon metagenomic approach whereas Chapter 4 and 5 on a whole shotgun metagenomic approach (Figure 1.2) (Liu et al., 2020).

PhyloPrimer was developed with biomonitoring and detection studies in mind and therefore was designed in order to develop highly taxon-specific primers to detect and verify the presence of specific organisms within an environmental community (Chapter 2). For this reason, even if the target of these studies usually represents only a minor part of the community, they can still be considered metagenomic studies as the amplification is performed on the entire DNA content of the environmental community. Going more towards the classic definition of amplicon metagenomics where a wider portion of the community is studied, PhyloPrimer can also be used to develop primers for universal gene amplification (e.g. 16S rRNA and 18S rRNA genes). However, the design of universal primers is usually not required as there are many that are publicly available and broadly used. In Chapter 3, for example, I have used primers designed by Takahashi et al. (2014) and already used by hundreds of studies (e.g. Kasai et al., 2015; Zamanzadeh et al., 2016; Naylor et al., 2017; McAnulty et al., 2017). The choice of predesigned primers, or at least the same gene regions, is often advantageous as it also allows less biased comparisons among different environmental communities (Thompson et al., 2017; Gilbert et al., 2018).

Data preparation and analysis in these two chapters required only minimal DNA amount, and it was cheap, fast and, above all, suitable to answer the study hypotheses. However, when a wider taxonomic and functional characterization of the microbial communities is required, a whole shotgun metagenomic approach must be applied.

Chapter 4 and 5 both focus on whole shotgun metagenomic data, however, using different approaches for the data analysis. In Chapter 4, where I analyzed three different proglacial datasets, I used a supervised approach for both taxonomic and functional annotation where the proglacial dataset was first aligned to known protein sequences and then homologous regions were identified using LongMeta. In Chapter 5, where I analyzed frozen soil microbial communities, I used an unsupervised approach in the functional (ORF prediction) and taxonomic bin definition. ORFs and bins were then assigned to their function and taxonomy by homology assignments with known sequences. The used approach was chosen in relation to the assembly quality, upon which both approaches rely, and the study aims.

The assembly coverage in Chapter 4 was lower than the one in Chapter 5: 10x and 164x, respectively. The assembly in Chapter 4 was 31 Gb whereas the one in Chapter 5 was 3 Gb. However 56.3% and 0.2% of the assembly belonged to sequences shorter than 1000 bp, and 89.7% and 3.1% of the assembly belonged to sequences shorter than 5000 bp for Chapter 4 and 5, respectively. N50 was therefore higher in Chapter 5 with 40,615 bp compared to 2,240 bp in proglacial system dataset of Chapter 4.

The higher assembly quality in Chapter 5 was due to different factors. First of all, in Chapter 4 the assembly was created from a highly fragmented dataset (i.e. 100/150 bp Illumina reads) whereas, it was created from the longer Nanopore reads in Chapter 5. The use of ONT technologies in genome reconstruction has indeed been shown as a game-changer when working with microbial sequences leading to more complete assemblies (Ayling et al., 2020; Nicholls et al., 2019; Giguere et al., 2020; De Maio et al., 2019; Somerville et al., 2019). Further to the read length, also the community characteristics and complexity may also have influenced the assembly quality. In Chapter 5, the cold-adapted permafrost microbial communities were quite conserved across samples (Figure A5.4), whereas proglacial surface soil communities from Chapter 4 were more diversified as they showed clearer microbial trends and also biographic differences across the different proglacial systems (Figure 4.8).

We saw a similar assembly quality trend in Chapter 4 where the datasets with longer read lengths (i.e. Greenland and Sweden with PE 2x150 bp) assembled better than the one with shorter read length (i.e. Svalbard with PE 2x100 bp). Further, highest coverages and therefore better assembly yields were observed in the ice rather than in the soil samples, the latter having a higher microbial complexity (Figure 4.6).

The different approaches (i.e. supervised and unsupervised annotation) were also applied because of the different intents of the data analysis. One of the objectives of Chapter 5 was to create a database of cold-adapted proteins (CAPP database). To do so, I had to use ab initio gene prediction to detect also unknown coding regions in this way not relying on homology to known proteins. In Chapter 4, instead, I explored known processes and therefore the similarity-based supervised approach was suitable for the data analysis.

Part of my thesis aims was to develop new tools and pipelines. Whereas I simply used preexisting pipelines for the analysis of the data in Chapter 3, I presented new tools, pipelines or databases in the other chapters.

I developed PhyloPrimer to facilitate the design of taxonomic-specific primers (Chapter 2) and LongMeta to create an easy reproducible pipeline for the analyses of metagenomic data (Chapter 4). I also applied the pipeline in Chapter 5, showing its reproducibility across different datasets and also its assignment reliability where contig taxonomy showed to be consistent within the same predicted taxonomic bins (Table 5.3 and A5.3).

In Chapter 5, I have developed a cold-adapted database of predicted proteins that, once publicly available, will be used to inform cold-adapted protein design and modeling. The work done in this chapter provides an example of the importance of data and discipline integration to obtain complete results. Sure enough, in this chapter DNA (i.e. metagenomics), RNA (i.e. metatrascriptomics) data and environmental variables were integrated together to give a complete picture of the microbial communities. And the further integration of bioengineering studies, cloning technologies and other omic disciplines (e.g. proteomics) will allow to take advantage of the CAPP database for the design of cold-adapted enzymes (Gutleben et al., 2018; Aguiar-pulido et al., 2016).

6.3 Glacial and proglacial communities

I have investigated microbial communities found in the glacial environment (Chapter 3), along proglacial soil successions (Chapter 4) and in the upper layer of permafrost soil (Chapter 5). These communities can not be directly compared between each other as they were analyzed with different approaches. However, we saw how they overall reflected what previously found in the literature. The glacial communities were dominated by typical glacial autotrophic (e.g. Cyanobacteria), spore-forming (e.g. Clostridium) and recalcitrant compound degrading (e.g. Bacteroidetes) organisms. However, for the first time, I observed how microbial communities

presented differences between different glacial areas and englacial and meteoric ice, showing how microbial communities in the englacial channel may differ and shape accordingly to water channel fluxes at the time of the ice formation, potentially indicating a pivotal role of englacial channels in the shaping of glacial microbial communities (Chapter 3).
Proglacial systems showed glacial influxes in the first soil stages and consequent autotrophic/heterotrophic trends along the soil successions, whereas the permafrost active layer environment showed a typical and consistent cold-adapted microbial community across sites.

In all the studied systems there was a correlation of the microbial community trends and clustering with system's geochemical variables and/or site distance separation, also confirming previous observations of microbial core community but also site endemisms and microdiversities (Delgado-Baquerizo et al., 2018; Malard et al., 2019). For instance, on the glacier surface the main taxonomic clustering was found between two main glacial surface areas characterized by different geochemical settings (Table 3.2, Figure 3.3 and 3.5) (Chapter 3). Microbial trends in the three soil successions (Chapter 4) followed different trends in different systems. Microbial communities in the Sweden and Svalbard systems followed distance relation from the ice edge, whereas, microbial communities in the Greenland system were mainly correlated with the geochemical variables (Table 4.1). In Chapter 5, the frozen soil community, similarly to the Greenland dataset in Chapter 4, showed similar correlation patterns both for distance and biochemical variables. However, in the frozen soil data we saw higher geochemical correlation values with the cDNA dataset, showing how differential presence of nutrients and enzyme cofactors can influence microbial activity.

Finally, even if with different sampling scales, the Greenland area studied in Chapter 4 and Chapter 5 was the same (extending from the ice edge in proximity of point 601 to Kangerlussuaq). However, microbial communities in the surface (Chapter 4) and the deepest frozen soil (Chapter 5) showed different microbial diversity, especially in the sampling sites closer to the ice edge. Whereas these samples were enriched with autotrophs in the surface soil, the frozen soil showed the presence of the same organisms along all the succession and also a less oscillating microbial community, which was observed, for example, in the surface soil along all the surface soil successions (Actinobacteria vs Proteobacteria; Figure 4.8A). This difference could be explained by the intrinsic differences between surface and the deepest frozen soil. Surface soil is a less challenging environment, with its higher temperatures and water availability, and it is more conditioned by the surrounding environmental conditions (e.g. proglacial gradients), not being isolated as the permafrost. Therefore, whereas surface soil microbial communities follow gradients set by the environment itself, and also symbiotic relationships with plants, the frozen soil communities are more isolated and mainly shaped by the challenging conditions, showing a stabler microbial composition.

6.4 Future work

Bioinformatics is a field in continuous evolution. The application of the next generation sequencing technologies to environmental and biomedical sciences is broad and brings the scientific and bioinformatics communities to constantly work on the improvement of sequencing quality and data analyses, where the output of more reads and longer sequences will allow a better resolution of microbial communities. This will lead to more MAG-based studies where complete and almost complete metagenome-assembled genomes (MAGs) are retrieved from metagenomic data, leading to a database enrichment and, in turn, to an increasingly better annotation due to the increasing number of known sequences. Continuing the exploration of environments such as the Arctic which is mostly unexplored will both benefit and help to the construction of the publicly sequence databases. The increase in generated data and databases' size will also pose few challenges where more computational power and bigger storage spaces will be needed for data analyses and to store and place sequence data and metadata. An important challenge will be also the functional characterization of all novel genes found in MAGs, which may be more difficult than their bioinformatics discovery.

All the results and pipelines presented in this thesis will be submitted to peer-reviewed journals. However, I would like to do some follow-up work and analyses connected to the data presented in Chapter 3 and 5.

The study presented in Chapter 3 showed promising results where I concluded that the englacial channel systems may have a role in the dispersion of glacial microbial communities. However, because of our indirect sampling method, the results are all based on the assumption that when we sample the clear ice bands we are sampling frozen englacial water. In the future I would like to directly sample flowing water from the englacial channels and see if same trends are seen. Furthermore, I think that working with metatrascriptomic data would give a further proof that englacial communities are active and can drive the creation of a typical glacial community.

The study I have developed in Chapter 5 is a start for the creation of a common and solid database constituted by cold-adapted proteins. The analyses I performed on the β-galactosidase and polyphenol oxidase sequences could be applied similarly to other enzymes or to the mined biosynthetic gene clusters (BGCs). Furthermore, it would be interesting to check (e.g. with protein modeling software or gene cloning and analysis) whether the detected modifications in the studied proteins (i.e. β-galactosidase and polyphenol oxidase) would lead to a higher protein flexibility and to a decrease in the optimal enzymatic temperature compared to the mesophilic homologous.

Finally, the comparison of surface and frozen soil microbial communities from Chapter 4 and 5 was only partially reliable because the samples were collected across different sampling scales and years and were also analyzed with different methods. However, because of the interesting results, where a clear difference was observed between surface and subsurface soil communities, it would be interesting to conduct a study where both the surface soil layer and the surface permafrost layer were sampled at the same sites in order to compare more consistently the microbial communities.

Bibliography

Acevedo-Rocha, C. G., Fang, G., Schmidt, M., Ussery, D. W., and Danchin, A. (2013). From essential to persistent genes: A functional approach to constructing synthetic life. *Trends in Genetics* 29.5, pp. 273–279.

ACIA (2005). Impacts of a warming Arctic. Tech. rep., pp. 1–139.

Addo, M. A. and Dos Santos, P. C. (2020). Distribution of Nitrogen-Fixation Genes in Prokaryotes Containing Alternative Nitrogenases. *ChemBioChem*, pp. 1–12.

Adékambi, T., Drancourt, M., and Raoult, D. (2009). The rpoB gene as a tool for clinical microbiologists. *Trends in Microbiology* 17.1, pp. 37–45.

Aguiar-pulido, V., Huang, W., Suarez-ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. (2016). Approaches for Microbiome Analysis. 12, pp. 5–16.

Ahmad, F., Ahmad, I., and Khan, M. S. (2008). Screening of free-living rhizospheric bacteria for their multiple plant growth promoting activities. *Microbiological Research* 163.2, pp. 173–181.

Ai, L., Liu, J., Jiang, Y., Guo, W., Wei, P., and Bai, L. (2019). Specific PCR method for detection of species origin in biochemical drugs via primers for the ATPase 8 gene by electrophoresis. *Microchimica Acta* 186.9, pp. 14–17.

Ali, P., Shah, A. A., Hasan, F., Hertkorn, N., Gonsior, M., Sajjad, W., and Chen, F. (2020). A Glacier Bacterium Produces High Yield of Cryoprotective Exopolysaccharide. *Frontiers in Microbiology* 10.February, pp. 1–16.

Allawi, H. T. and SantaLucia, J. (1998a). Nearest neighbor thermodynamic parameters for internal G·A mismatches in DNA. *Biochemistry* 37.8, pp. 2170–2179.

– (1998b). Nearest-neighbor thermodynamics of internal A·C mismatches in DNA: Sequence dependence and pH effects. *Biochemistry* 37.26, pp. 9435–9444.

– (1998c). Thermodynamics of internal C·T mismatches in DNA. *Nucleic Acids Research* 26.11, pp. 2694–2701.

Allawi, H. T. and Santalucia, J. (1997). Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochemistry* 36.34, pp. 10581–10594.

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., and Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature* 568.7753, pp. 499–504.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods* 11.11, pp. 1144–1146.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215.3, pp. 403–10. arXiv: arXiv:1611.08307v1.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21.1, pp. 1–16.

Amir, A., Daniel, M., Navas-Molina, J., Kopylova, E., Morton, J., Xu, Z. Z., Eric, K., Thompson, L., Hyde, E., Gonzalez, A., and Knight, R. (2017). Deblur Rapidly Resolves Single-. *American Society for Microbiology* 2.2, pp. 1–7.

An, L. Z., Chen, Y., Xiang, S. R., Shang, T. C., and Tian, L. D. (2010). Differences in community composition of bacteria in four glaciers in western China. *Biogeosciences* 7.6, pp. 1937–1952.

Anantharaman, K., Hausmann, B., Jungbluth, S. P., Kantor, R. S., Lavy, A., Warren, L. A., Rappé, M. S., Pester, M., Loy, A., Thomas, B. C., and Banfield, J. F. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME Journal* 12.7, pp. 1715–1728.

Anesio, A. M., Hodson, A. J., Fritz, A., Psenner, R., and Sattler, B. (2009). High microbial activity on glaciers: Importance to the global carbon cycle. *Global Change Biology* 15.4, pp. 955–960.

Anesio, A. M. and Laybourn-Parry, J. (2012). Glaciers and ice sheets as a biome. *Trends in Ecology & Evolution* 27.4, pp. 219–225.

Anesio, A. M., Lutz, S., Chrismas, N. A., and Benning, L. G. (2017). The microbiome of glaciers and ice sheets. *npj Biofilms and Microbiomes* 3.1, pp. 0–1.

Åqvist, J., Isaksen, G. V., and Brandsdal, B. O. (2017). Computation of enzyme cold adaptation. *Nature Reviews Chemistry* 1.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics* 25, pp. 25–29.

Auer, L., Mariadassou, M., O'Donohue, M., Klopp, C., and Hernandez-Raquet, G. (2017). Analysis of large 16S rRNA Illumina data sets: Impact of singleton read filtering on microbial community description. *Molecular Ecology Resources* 17.6, e122–e132.

Auguie, B. (2017). gridExtra: functions in Grid graphics. R Package Version 2.3. *CRAN PROJECT.*

Ayala-Del-Río, H. L., Chain, P. S., Grzymski, J. J., Ponder, M. A., Ivanova, N., Bergholz, P. W., Bartolo, G. D., Hauser, L., Land, M., Bakermans, C., Rodrigues, D., Klappenbach, J., Zarka, D., Larimer, F., Richardson, P., Murray, A., Thomashow, M., and Tiedje, J. M. (2010). The genome sequence of psychrobacter arcticus 273-4, a psychroactive siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. *Applied and Environmental Microbiology* 76.7, pp. 2304–2312.

Ayling, M., Clark, M. D., and Leggett, R. M. (2020). New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics* 21.2, pp. 584–594.

Baas-Becking, L. G. M. (1934). Geobiologie, of Inleiding Tot de Milieukunde: Met Literatuurlijst en Ind.

Babich, H. and Davis, D. L. (1981). Phenol: A review of environmental and health risks. *Regulatory Toxicology and Pharmacology* 1.1, pp. 90–109.

Bae, E. and Phillips, G. N. (2004). Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. *Journal of Biological Chemistry* 279.27, pp. 28202–28208.

Bagshaw, E. A., Tranter, M., Fountain, A. G., Welch, K. A., Basagic, H., and Lyons, W. B. (2007). Biogeochemical evolution of cryoconite holes on Canada Glacier, Taylor Valley, Antarctica. *Journal of Geophysical Research: Biogeosciences.*

Baines, S. L., Carter, G., Jennison, A. V., Graham, R. M., Sintchenko, V., Wang, Q., Rockett, R. J., Timms, V. J., Martinez, E., Ballard, S., Tomita, T., Isles, N., Kristy, A., Pitchers, W., Stinear, T. P., Williamson, D. A., Benjamin, P., Seemann, T., Diagnostic, M., and Public, U. (2019). Complete microbial genomes for public health in Australia and Southwest Pacific. *bioRxiv*, p. 829663.

Baird, D. J. and Hajibabaei, M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology* 21.8, pp. 2039–2044.

Bajerski, F. and Wagner, D. (2013). Bacterial succession in Antarctic soils of two glacier forefields on Larsemann Hills, East Antarctica. *FEMS Microbiology Ecology* 85.1, pp. 128–142.

Bakermans, C., Bergholz, P. W., Ayala-del-Río, H., and Tiedje, J. (2009). Genomic Insights into Cold Adaptation of Permafrost Bacteria. *Permafrost Soils, Soil Biology 16*, pp. 159–168.

Bakermans, C., Bergholz, P. W., Rodrigues, D. F., and Vishnivetskaya, T. A. (2014). Genomic and Expression Analyses of Cold-Adapted Microorganisms. *Polar Microbiology: Life in a Deep Freeze*, pp. 126–155.

Balaji, S., Gopi, K., and Muthuvelan, B. (2013). A review on production of poly β hydroxybutyrates from cyanobacteria for the production of bio plastics. *Algal Research* 2.3, pp. 278–285.

Bamber, J. L. (1988). Enhanced radar scattering from water inclusions in ice. *Journal of Glaciology* 34.118, pp. 293–296.

Banerji, A., Bagley, M., Elk, M., Pilgrim, E., Martinson, J., and Santo Domingo, J. (2018). Spatial and temporal dynamics of a freshwater eukaryotic plankton community revealed via 18S rRNA gene metabarcoding. *Hydrobiologia* 818.1, pp. 71–86.

Barka, E. A., Vatsa, P., Sanchez, L., Nathalie Gaveau-Vaillant, C. J., Klenk, H.-P., Clément, C., Ouhdouch, Y., and P. van Wezeld, G. (2016). Taxonomy, Physiology, and Natural Products of Actinobacteria. *American Society for Microbiology* 80.1, pp. 1–43.

Barton, L. L., Mandl, M., and Loy, A. (2010). Geomicrobiology: Molecular and environmental perspective. Ed. by L. A. Barton LL, Mandl M. Springer.

Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research* 47.D1, pp. D506–D515.

Baxevanis, A. D. (2020). Assessing pairwise sequence similarity: BLAST and FASTA. *Bioinformatics*. Ed. by A. D. Baxevanis, G. D. Bader, and D. S. Wishart. forth edit. John Wiley & Sons. Chap. Assessing, pp. 45–78.

Bell, T. H., Callender, K. L., Whyte, L. G., and Greer, C. W. (2013). Microbial competition in polar soils: A review of an understudied but potentially important control on productivity. *Biology* 2.2, pp. 533–554.

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., Dvornicic, M., Soldo, J. P., Koh, J. Y., Tong, C., Ng, O. T., Barkham, T., Young, B., Marimuthu, K., Chng, K. R., Sikic, M., and Nagarajan, N. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology* 37.8, pp. 937–944.

Blazewicz, S. J., Barnard, R. L., Daly, R. A., and Firestone, M. K. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: Limitations and uses. *ISME Journal* 7.11, pp. 2061–2068.

Bleidorn, C. (2017). Phylogenomics: An introduction. Springer, Cham.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. (2019). AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* 47.W1, W81–W87.

Blumer, C. and Haas, D. (2000). Mechanism, regulation, and ecological role of bacterial cyanide biosynthesis. *Archives of Microbiology* 173.3, pp. 170–177.

Boetius, A., Anesio, A. M., Deming, J. W., Mikucki, J. A., and Rapp, J. Z. (2015). Microbial ecology of the cryosphere: Sea ice and glacial habitats. *Nature Reviews Microbiology* 13.11, pp. 677–690.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30.15, pp. 2114–2120.

Bommarito, S., Peyret, N., and SantaLucia, J. (2000). Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Research* 28.9, pp. 1929–1934.

Borin, S., Ventura, S., Tambone, F., Mapelli, F., Schubotz, F., Brusetti, L., Scaglia, B., D'Acqui, L. P., Solheim, B., Turicchia, S., Marasco, R., Hinrichs, K. U., Baldi, F., Adani, F., and Daffonchio, D. (2010). Rock weathering creates oases of life in a High Arctic desert. *Environmental Microbiology* 12.2, pp. 293–303.

Bowers, R. M. et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35.8, pp. 725–731.

Boyce, R., Chilana, P., and Rose, T. M. (2009). iCODEHOP: A new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. *Nucleic Acids Research* 37.SUPPL. 2, pp. 222–228.

Boyd, E. S., Hamilton, T. L., Havig, J. R., Skidmore, M. L., and Shock, E. L. (2014). Chemolithotrophic primary production in a subglacial ecosystem. *Applied and Environmental Microbiology* 80.19, pp. 6146–6153.

Bradley, J. A., Arndt, S., Šabacká, M., Benning, L. G., Barker, G. L., Blacker, J. J., Yallop, M. L., Wright, K. E., Bellas, C. M., Telling, J., Tranter, M., and Anesio, A. M. (2016). Microbial dynamics in a High Arctic glacier forefield: A combined field, laboratory, and modelling approach. *Biogeosciences* 13.19, pp. 5677–5696.

Bradley, J. A., Singarayer, J. S., and Anesio, A. M. (2014). Microbial community dynamics in the forefield of glaciers. *Proceedings of the Royal Society B: Biological Sciences* 281.1795.

Brankatschk, R., Töwe, S., Kleineidam, K., Schloter, M., and Zeyer, J. (2011). Abundances and potential activities of nitrogen cycling microbial communities along a chronosequence of a glacier forefield. *ISME Journal* 5.6, pp. 1025–1037.

Braun, M. H., Malz, P., Sommer, C., Farías-Barahona, D., Sauter, T., Casassa, G., Soruco, A., Skvarca, P., and Seehaus, T. C. (2019). Constraining glacier elevation and mass changes in South America. *Nature Climate Change* 9.2, pp. 130–136.

Braun, V. and Hantke, K. (2011). Recent insights into iron import by bacteria. *Current Opinion in Chemical Biology* 15.2, pp. 328–334.

Bridgham, S. D. and Ye, R. (2015). Organic Matter Mineralization and Decomposition.

Brown, B. L., Watson, M., Minot, S. S., Rivera, M. C., and Franklin, R. B. (2017). MinIONTM nanopore sequencing of environmental metagenomes: A synthetic approach. *GigaScience* 6.3, pp. 1–10.

Brown, G. H. (2002). Glacier meltwater hydrochemistry. *Applied Geochemistry* 17.7, pp. 855–883.

Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L., and Leddy, M. B. (2020). Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. *PLoS ONE* 15.2, pp. 1–21.

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, pp. 266–279.

Bryce, C., Blackwell, N., Schmidt, C., Otte, J., Huang, Y. M., Kleindienst, S., Tomaszewski, E., Schad, M., Warter, V., Peng, C., Byrne, J. M., and Kappler, A. (2018). Microbial anaerobic Fe(II) oxidation – Ecology, mechanisms and environmental implications. *Environmental Microbiology* 20.10, pp. 3462–3483.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12.1, pp. 59–60.

Calderoli, P. A., Collavino, M. M., Behrends Kraemer, F., Morrás, H. J., and Aguilar, O. M. (2017). Analysis of nifH-RNA reveals phylotypes related to Geobacter and Cyanobacteria as important functional components of the N2-fixing community depending on depth and agricultural use of soil. *MicrobiologyOpen* 6.5, pp. 1–15.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal* 11.12, pp. 2639–2643.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13, pp. 581–583.

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., and Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic acids research* 47.18, e103.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., Mcdonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). Correspondence QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group* 7.5, pp. 335–336.

Carbon, S. et al. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*.

Carrivick, J. L. and Heckmann, T. (2017). Short-term geomorphological evolution of proglacial systems. *Geomorphology* 287, pp. 3–28.

Carte, A. E. (1961). Air bubbles in ice. *Proceedings of the Physical Society* 77.3, pp. 1958–1967.

Carvalhais, L. C., Dennis, P. G., Tyson, G. W., and Schenk, P. M. (2012). Application of metatranscriptomics to soil environments. *Journal of Microbiological Methods* 91.2, pp. 246–251.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology* 73.1, pp. 278–288.

Casillo, A., Parrilli, E., Sannino, F., Mitchell, D. E., Gibson, M. I., Marino, G., Lanzetta, R., Parrilli, M., Cosconati, S., Novellino, E., Randazzo, A., Tutino, M. L., and Corsaro, M. M. (2017). Structure-activity relationship of the exopolysaccharide from a psychrophilic bacterium: A strategy for cryoprotection. *Carbohydrate Polymers* 156, pp. 364–371.

Castello, J. and Rogers, S. (2005). Life in Ancient Life. Ed. by Princeton University Press.

Castro, A. P. de, Fernandes, G. d. R., and Franco, O. L. (2014). Insights into novel antimicrobial compounds and antibiotic resistance genes from soil metagenomes. *Frontiers in Microbiology* 5.SEP, pp. 1–9.

Catania, G. A., Neumann, T. A., and Price, S. F. (2008). Characterizing englacial drainage in the ablation zone of the Greenland ice sheet. *Journal of Glaciology* 54.187, pp. 567–578.

Cavaliere, M., Feng, S., Soyer, O. S., and Jiménez, J. I. (2017). Cooperation in microbial communities and their biotechnological applications. *Environmental Microbiology* 19.8, pp. 2949–2963.

Cavicchioli, R. (2016). On the concept of a psychrophile. *ISME Journal* 10.4, pp. 793–795.

Chao, Y. C., Merritt, M., Schaefferkoetter, D., and Evans, T. G. (2020). High-throughput quantification of protein structural change reveals potential mechanisms of temperature adaptation in Mytilus mussels. *BMC Evolutionary Biology* 20.1, pp. 1–18.

Chen, I. M. A., Markowitz, V. M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Andersen, E., Huntemann, M., Varghese, N., Hadjithomas, M., Tennessen, K., Nielsen, T., Ivanova, N. N., and Kyrpides, N. C. (2017). IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research* 45.D1, pp. D507–D516.

Chen, K. and Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology* 1.2, pp. 0106–0112.

Chen, L. X., Anantharaman, K., Shaiber, A., Murat Eren, A., and Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Research* 30.3, pp. 315–333.

Chen, Y., Li, X. K., Si, J., Wu, G. J., Tian, L. D., and Xiang, S. R. (2016). Changes of the bacterial abundance and communities in shallow ice cores from Dunde and Muztagata glaciers, Western China. *Frontiers in Microbiology* 7.NOV, pp. 1–16.

Choi, G. M. and Im, W. T. (2018). Paraburkholderia azotifigens sp. nov., a nitrogen-fixing bacterium isolated from paddy soil. *International Journal of Systematic and Evolutionary Microbiology* 68.1, pp. 310–316.

Chrismas, N. A., Anesio, A. M., and Śanchez-Baracaldo, P. (2018). The future of genomics in polar and alpine cyanobacteria. *FEMS Microbiology Ecology* 94.4, pp. 1–10.

Chrismas, N. A., Barker, G., Anesio, A. M., and Sánchez-Baracaldo, P. (2016). Genomic mechanisms for cold tolerance and production of exopolysaccharides in the Arctic cyanobacterium Phormidesmis priestleyi BC1401. *BMC Genomics* 17.1, pp. 1–14.

Christner, B. C., Mikucki, J. A., Foreman, C. M., Denson, J., and Priscu, J. C. (2005). Glacial ice cores: A model system for developing extraterrestrial decontamination protocols. *Icarus* 174.2 SPEC. ISS. Pp. 572–584.

Chuang, L. Y., Cheng, Y. H., and Yang, C. H. (2013). Specific primer design for the polymerase chain reaction. *Biotechnology Letters* 35.10, pp. 1541–1549.

Cid, F. P., Rilling, J. I., Graether, S. P., Bravo, L. A., De La Luz Mora, M., and Jorquera, M. A. (2016). Properties and biotechnological applications of ice-binding proteins in bacteria. *FEMS Microbiology Letters* 363.11, pp. 1–12.

Clarhäll, A. (2011). SKB studies of the periglacial environment – report from field studies in Kangerlussuaq, Greenland 2008 and 2010. *Svensk Kärnbränslehantering AB* March.

Cock, P. J. and Whitworth, D. E. (2007). Evolution of gene overlaps: Relative reading frame bias in prokaryotic two-component system genes. *Journal of Molecular Evolution* 64.4, pp. 457–462.

Cogley, J. G. (2011). Mass-balance terms revisited. *Journal of Glaciology* 56.200, pp. 997–1001.

Collins, T. and Margesin, R. (2019). Psychrophilic lifestyles: mechanisms of adaptation and biotechnological tools. *Applied Microbiology and Biotechnology* 103.7, pp. 2857–2871.

Cook, J., Edwards, A., Takeuchi, N., and Irvine-Fynn, T. (2016). Cryoconite: The dark biological secret of the cryosphere. *Progress in Physical Geography* 40.1, pp. 66–111.

Cristescu, M. E. (2019). Can Environmental RNA Revolutionize Biodiversity Science? *Trends in Ecology and Evolution* 34.8, pp. 694–697.

Cuffey, K. and Paterson, W. (2010). The physics of glaciers. 4th Editio. Elsevier.

Čuklina, J., Hahn, J., Imakaev, M., Omasits, U., Förstner, K. U., Ljubimov, N., Goebel, M., Pessi, G., Fischer, H. M., Ahrens, C. H., Gelfand, M. S., and Evguenieva-Hackenberg, E. (2016). Genome-wide transcription start site mapping of Bradyrhizobium japonicum grown free-living or in symbiosis - a rich resource to identify new transcripts, proteins and to study gene regulation. *BMC Genomics* 17.1, pp. 1–19.

Dahal, B., NandaKafle, G., Perkins, L., and Brözel, V. S. (2017). Diversity of free-Living nitrogen fixing Streptomyces in soils of the badlands of South Dakota. *Microbiological Research* 195, pp. 31–39.

Daims, H., Lücker, S., and Wagner, M. (2016). A New Perspective on Microbes Formerly Known as Nitrite-Oxidizing Bacteria. *Trends in Microbiology* 24.9, pp. 699–712.

D'Amico, S., Collins, T., Marx, J. C., Feller, G., and Gerday, C. (2006). Psychrophilic microorganisms: Challenges for life. *EMBO Reports* 7.4, pp. 385–389.

Dangendorf, S., Marcos, M., Wöppelmann, G., Conrad, C. P., Frederikse, T., and Riva, R. (2017). Reassessment of 20th century global mean sea level rise. *Proceedings of the National Academy of Sciences of the United States of America* 114.23, pp. 5946–5951.

Dani, K. G., Mader, H. M., Wolff, E. W., and Wadham, J. L. (2012). Modelling the liquid-water vein system within polar ice sheets as a potential microbial habitat. *Earth and Planetary Science Letters* 333-334, pp. 238–249.

Davidson, E. A. and Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440.7081, pp. 165–173.

Davies, M. R., McIntyre, L., Mutreja, A., Lacey, J. A., Lees, J. A., Towers, R. J., Duchêne, S., Smeesters, P. R., Frost, H. R., Price, D. J., Holden, M. T., David, S., Giffard, P. M., Worthing, K. A., Seale, A. C., Berkley, J. A., Harris, S. R., Rivera-Hernandez, T., Berking, O., Cork, A. J., Torres, R. S., Lithgow, T., Strugnell, R. A., Bergmann, R., Nitsche-Schmitz, P., Chhatwal, G. S., Bentley, S. D., Fraser, J. D., Moreland, N. J., Carapetis, J. R., Steer, A. C., Parkhill, J., Saul, A., Williamson, D. A., Currie, B. J., Tong, S. Y., Dougan, G., and Walker, M. J. (2019). Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nature Genetics* 51.6, pp. 1035–1043.

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6.226, pp. 1–14.

De Maayer, P., Anderson, D., Cary, C., and Cowan, D. A. (2014). Some like it cold: Understanding the survival strategies of psychrophiles. *EMBO Reports* 15.5, pp. 508–517.

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., Wick, R., Oun, M. A., Stubberfield, E., Hoosdally, S. J., Crook, D. W., Peto, T. E., Sheppard, A. E., Bailey, M. J., Read, D. S., Anjum, M. F., Sarah Walker, A., and Stoesser, N. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics* 5.9.

De Wit, R. and Bouvier, T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology* 8.4, pp. 755–758.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology* 34.5, pp. 518–524.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., Vere, N. de, Pfrender, M. E., and Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology* 26.21, pp. 5872–5895.

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-gonzález, A., Eldridge, D. J., Bardgett, R. D., Maestre, F. T., Singh, B. K., and Fierer, N. (2018). Bacteria Found in Soil. *Science* 325.February, pp. 320–325.

Deming, J. W. and Young, J. N. (2017). The role of exopolysaccharides in microbial adaptation to cold habitats. *Psychrophiles: From Biodiversity to Biotechnology: Second Edition.*

Deng, J., Gu, Y., Zhang, J., Xue, K., Qin, Y., Yuan, M., Yin, H., He, Z., Wu, L., Schuur, E. A., Tiedje, J. M., and Zhou, J. (2015). Shifts of tundra bacterial and archaeal communities along a permafrost thaw gradient in Alaska. *Molecular Ecology* 24.1, pp. 222–234.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* 72.7, pp. 5069–5072.

Deslippe, J. R., Hartmann, M., Simard, S. W., and Mohn, W. W. (2012). Long-term warming alters the composition of Arctic soil microbial communities. *FEMS Microbiology Ecology* 82.2, pp. 303–315.

Dieser, M., Broemsen, E. L., Cameron, K. A., King, G. M., Achberger, A., Choquette, K., Hagedorn, B., Sletten, R., Junge, K., and Christner, B. C. (2014). Molecular and biogeochemical evidence for methane cycling beneath the western margin of the Greenland Ice Sheet. *ISME Journal* 8.11, pp. 2305–2316.

Dieser, M., Nocker, A., Priscu, J. C., and Foreman, C. M. (2010). Viable microbes in ice: Application of molecular assays to McMurdo Dry Valley lake ice communities. *Antarctic Science.*

Dobinski, W. (2011). Permafrost. *Earth-Science Reviews* 108.3-4, pp. 158–169.

D'Souza, G., Shitut, S., Preussger, D., Yousif, G., Waschina, S., and Kost, C. (2018). Ecology and evolution of metabolic cross-feeding interactions in bacteria. *Natural Product Reports* 35.5, pp. 455–488.

Du, L. and Lou, L. (2010). PKS and NRPS release mechanisms. *Natural Product Reports* 27.2, pp. 255–278.

Du, X., Sang, P., Xia, Y. L., Li, Y., Liang, J., Ai, S. M., Ji, X. L., Fu, Y. X., and Liu, S. Q. (2017). Comparative thermal unfolding study of psychrophilic and mesophilic subtilisin-like serine proteases by molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics* 35.7, pp. 1500–1517.

Duarte, A. W. F., Barato, M. B., Nobre, F. S., Polezel, D. A., Oliveira, T. B. de, Santos, J. A. dos, Rodrigues, A., and Sette, L. D. (2018). Production of cold-adapted enzymes by filamentous fungi from King George Island, Antarctica. *Polar Biology* 41.12, pp. 2511–2521.

Dubnick, A., Wadham, J., Tranter, M., Sharp, M., Orwin, J., Barker, J., Bagshaw, E., and Fitzsimons, S. (2017). Trickle or treat: The dynamics of nutrient export from polar glaciers. *Hydrological Processes* 31.9, pp. 1776–1789.

Dumbrell, A. J., Nelson, M., Helgason, T., Dytham, C., and Fitter, A. H. (2010). Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME Journal* 4.3, pp. 337–345.

Edwards, A., Anesio, A. M., Rassner, S. M., Sattler, B., Hubbard, B., Perkins, W. T., Young, M., and Griffith, G. W. (2011). Possible interactions between bacterial diversity, microbial activity and supraglacial hydrology of cryoconite holes in Svalbard. *ISME Journal* 5.1, pp. 150–160.

Edwards, A. and Cook, S. (2015). Microbial dynamics in glacier forefield soils show succession is not just skin deep. *Molecular Ecology* 24.5, pp. 963–966.

Egan, K., Ross, R. P., and Hill, C. (2017). Bacteriocins: antibiotics in the age of the microbiome. *Emerging Topics in Life Sciences* 1.1, pp. 55–63.

Eiler, A. (2006). Evidence for the ubiquity of mixotrophic bacteria in the upper ocean: Implications and consequences. *Applied and Environmental Microbiology* 72.12, pp. 7431–7437.

Elberling, B., Nordstrøm, C., Grøndahl, L., Søgaard, H., Friborg, T., Christensen, T. R., Ström, L., Marchand, F., and Nijs, I. (2008). High-Arctic Soil CO2 and CH4 Production Controlled by Temperature, Water, Freezing and Snow. *Advances in Ecological Research* 40.07, pp. 441–472.

Elbrecht, V., Hebert, P. D., and Steinke, D. (2018). Slippage of degenerate primers can cause variation in amplicon length. *Scientific Reports* 8.1, pp. 1–5.

Els, N., Baumann-Stanzer, K., Larose, C., Vogel, T. M., and Sattler, B. (2019a). Beyond the planetary boundary layer: Bacterial and fungal vertical biogeography at Mount Sonnblick, Austria. *Geo: Geography and Environment* 6.1.

Els, N., Larose, C., Baumann-Stanzer, K., Tignat-Perrier, R., Keuschnig, C., Vogel, T. M., and Sattler, B. (2019b). Microbial composition in seasonal time series of free tropospheric air and precipitation reveals community separation. Vol. 35. 4. Springer Netherlands, pp. 671–701.

Engel, P. and Moran, N. A. (2013). The gut microbiota of insects - diversity in structure and function. *FEMS Microbiology Reviews* 37.5, pp. 699–735.

Englander, S. W. and Mayne, L. (2014). The nature of protein folding pathways. *Proceedings of the National Academy of Sciences of the United States of America* 111.45, pp. 15873–15880.

Ernakovich, J. G. and Wallenstein, M. D. (2015). Permafrost microbial community traits and functional diversity indicate low activity at in situ thaw temperatures. *Soil Biology and Biochemistry* 87, pp. 78–89.

Escobar-Zepeda, A., De León, A. V. P., and Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics* 6.DEC, pp. 1–15.

Essinger, S. D., Reichenberger, E., Morrison, C., Blackwood, C. B., and Rosen, G. L. (2015). A toolkit for ARB to integrate custom databases and externally built phylogenies. *PLoS ONE* 10.1, pp. 1–7.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science* 320.5879, pp. 1034–1039.

Federico, A., Dallio, M., Di Sarno, R., Giorgio, V., and Miele, L. (2017). Gut microbiota, obesity and metabolic disorders. *Minerva Gastroenterologica e Dietologica* 63.4, pp. 337–344.

Feng, Q., Chen, W. D., and Wang, Y. D. (2018). Gut microbiota: An integral moderator in health and disease. *Frontiers in Microbiology* 9.FEB, pp. 1–8.

Fernandez, L., Bertilsson, S., and Peura, S. (2020). Non-cyanobacterial diazotrophs dominate nitrogen-fixing communities in permafrost thaw ponds. *Limnology and Oceanography* 65.S1, S180–S193.

Fernández-Martínez, M. A., Pointing, S. B., Pérez-Ortega, S., Arróniz-Crespo, M., Allan Green, T. G., Rozzi, R., Sancho, L. G., and Ríos, A. de los (2016). Functional ecology of soil microbial communities along a glacier forefield in Tierra del Fuego (Chile). *International Microbiology* 19.3, pp. 161–173.

Fernández-Martínez, M. A., Pérez-Ortega, S., Pointing, S. B., Allan Green, T. G., Pintado, A., Rozzi, R., Sancho, L. G., and Ríos, A. de los (2017). Microbial succession dynamics along glacier forefield chronosequences in Tierra del Fuego (Chile). *Polar Biology* 40.10, pp. 1939–1957.

Ferreira, C. M., Vilas-Boas, Â., Sousa, C. A., Soares, H. M., and Soares, E. V. (2019). Comparison of five bacterial strains producing siderophores with ability to chelate iron under alkaline conditions. *AMB Express* 9.1.

Fierer, N., Jackson, J. A., Vilgalys, R., and Jackson, R. B. (2005). Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. *Applied and Environmental Microbiology* 71.7, pp. 4117–4120.

Fierer, N. and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America* 103.3, pp. 626–631.

Fierer, N., Nemergut, D., Knight, R., and Craine, J. M. (2010). Changes through time: Integrating microorganisms into the study of succession. *Research in Microbiology* 161.8, pp. 635–642.

Fontaneto, D. and Hortal, J. (2012). Microbial biogeography: is everything small everywhere? Ed. by P. R. Ogilvie, L A and Hirschl. Horizon Scientific Press.

Fountain, A. G., Campbell, J. L., Schuur, E. A. G., Stammerjohn, S. E., Williams, M. W., and Ducklow, H. W. (2012). The Disappearing Cryosphere: Impacts and Ecosystem Responses to Rapid Cryosphere Loss. *BioScience* 62.4, pp. 405–415.

Fountain, A. G., Jacobel, R. W., Schlichting, R., and Jansson, P. (2005). Fractures as the main pathways of water flow in temperate glaciers. *Nature* 433.7026, pp. 618–621.

Fountain, A. G., Tranter, M., Nylen, T. H., Lewis, K. J., and Mueller, D. R. (2004). Evolution of cryoconite holes and their contribution to meltwater runoff from glaciers in the McMurdo Dry Valleys, Antarctica. *Journal of Glaciology* 50.168, pp. 35–45.

Fountain, A. G. and Walder, J. S. (1998). Water flow through temperate glaciers. *Reviews of Geophysics* 36.3, pp. 299–328.

Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., and Huttenhower, C. (2015). Sequencing and beyond: Integrating molecular 'omics' for microbial community profiling. arXiv: 0402594v3 `[arXiv:cond-mat]`.

Frey, B., Rieder, S. R., Brunner, I., Plötze, M., Koetzsch, S., Lapanje, A., Brandi, H., and Furrer, G. (2010). Weathering-Associated bacteria from the damma glacier forefield: Physiological capabilities and impact on granite dissolution. *Applied and Environmental Microbiology* 76.14, pp. 4788–4796.

Gaby, J. C. and Buckley, D. H. (2012). A comprehensive evaluation of PCR primers to amplify the nifH gene of nitrogenase. *PLoS ONE* 7.7.

Gadberry, M. D., Malcomber, S. T., Doust, A. N., and Kellogg, E. A. (2005). Primaclade - A flexible tool to find conserved PCR primers across multiple species. *Bioinformatics* 21.7, pp. 1263–1264.

Galperin, M. Y. (2016). Genome Diversity of Spore-Forming Firmicutes. *The Bacterial Spore.*

Ganzert, L., Jurgens, G., Munster, U., and Wagner, D. (2007). Methanogenic communities in permafrost-affected soils of the Laptev Sea coast, SiberianArctic, characterized by16S rRNA gene fingerprints. *FEMS Microbiol Ecol* 59.2, p. 476.

Garcia-Lopez, E., Maria Moreno, A., and Cid, C. (2019). Microbial Community Structure and Metabolic Networks in Polar Glaciers. *Metagenomics - Basics, Methods and Applications.* IntechOpen.

Garcias-Bonet, N., Arrieta, J. M., Duarte, C. M., and Marbà, N. (2016). Nitrogen-fixing bacteria in Mediterranean seagrass (Posidonia oceanica) roots. *Aquatic Botany* 131, pp. 57–60.

Garibyan, L. and Avashia, N. (2013). Polymerase chain reaction. *Journal of Investigative Dermatology* 133.3, pp. 1–4.

Garnier, S. (2018). viridis: Default Color Maps from 'matplotlib'. *R package version 0.5.1.*

Giguere, D. J., Bahcheli, A. T., Joris, B. R., Paulssen, J. M., Gieg, L. M., and Flatley, M. W. (2020). Complete and validated genomes from a metagenome. *bioRxiv.*

Gil, R., Silva, F. J., Pereto, J., and Moya, andres (2004). Histone antibody validation table. *Microbiology and molecular biology reviews : MMBR* 68.3, pp. 518–537.

Gilbert, J. A., Jansson, J. K., and Knight, R. (2018). Earth Microbiome Project and Global Systems Biology. *mSystems* 3.3, pp. 1–4.

Gillings, M. R. (2017). Lateral gene transfer, bacterial genome evolution, and the Anthropocene. *Annals of the New York Academy of Sciences* 1389.1, pp. 20–36.

Glass, E. M. and Meyer, F. (2011). The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, pp. 325–331. arXiv: NIHMS150003.

Goel, N., Singh, S., and Aseri, T. C. (2013). A comparative analysis of soft computing techniques for gene prediction. *Analytical Biochemistry* 438.1, pp. 14–21.

Gokul, J. K., Hodson, A. J., Saetnan, E. R., Irvine-Fynn, T. D., Westall, P. J., Detheridge, A. P., Takeuchi, N., Bussell, J., Mur, L. A., and Edwards, A. (2016). Taxon interactions control the distributions of cryoconite bacteria colonizing a High Arctic ice cap. *Molecular Ecology* 25.15, pp. 3752–3767. arXiv: arXiv:1011.1669v3.

Goltsman, D. S., Comolli, L. R., Thomas, B. C., and Banfield, J. F. (2015). Community transcriptomics reveals unexpected high microbial diversity in acidophilic biofilm communities. *ISME Journal* 9, pp. 1014–1023.

González-Toril, E., Santofimia, E., Blanco, Y., López-Pamo, E., Gómez, M. J., Bobadilla, M., Cruz, R., Palomino, E. J., and Aguilera, Á. (2015). Pyrosequencing-Based Assessment of the Microbial Community Structure of Pastoruri Glacier Area (Huascarán National Park, Perú), a Natural Extreme Acidic Environment. *Microbial Ecology* 70.4, pp. 936–947.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17.6, pp. 333–351. arXiv: arXiv : 1011.1669v3.

Graham, E. D., Heidelberg, J. F., and Tully, B. J. (2017). Binsanity: Unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 2017.3, pp. 1–19.

Green, J. L., Bohannan, B. J., and Whitaker, R. J. (2008). Microbial biogeography: From taxonomy to traits. *Science* 320.5879, pp. 1039–1043.

Gryk, M. R., Jardetzky, O., Klig, L. S., and Yanofsky, C. (1996). Flexibility of DNA binding domain of trp repressor required for recognition of different operator sequences. *Protein Science* 5.6, pp. 1195–1197.

Grzesiak, J., Górniak, D., Światecki, A., Aleksandrzak-Piekarczyk, T., Szatraj, K., and Zdanowski, M. K. (2015). Microbial community development on the surface of Hans and Werenskiold Glaciers (Svalbard, Arctic): a comparison. *Extremophiles* 19.5, pp. 885–897.

Gtari, M., Ghodhbane-Gtari, F., Nouioui, I., Beauchemin, N., and Tisa, L. S. (2012). Phylogenetic perspectives of nitrogen-fixing actinobacteria. *Archives of Microbiology* 194.1, pp. 3–11.

Gutleben, J., Mares, M. C. D., Elsas, J. D. V., Overmann, J., and Sipkema, D. (2018). The multi-omics promise in context: from sequence to microbial isolate. *Critical Reviews in Microbiology* 44.2, pp. 212–229.

Hall, J. P., Brockhurst, M. A., and Harrison, E. (2017). Sampling the mobile gene pool: Innovation via horizontal gene transfer in bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1735, pp. 1–10.

Hamel, R., Levasseur, R., and Appanna, V. D. (1999). Oxalic acid production and aluminum tolerance in Pseudomonas fluorescens. *Journal of Inorganic Biochemistry* 76.2, pp. 99–104.

Hanna, E., Navarro, F. J., Pattyn, F., Domingues, C. M., Fettweis, X., Ivins, E. R., Nicholls, R. J., Ritz, C., Smith, B., Tulaczyk, S., Whitehouse, P. L., and Jay Zwally, H. (2013). Ice-sheet mass balance and climate change. *Nature* 498.7452, pp. 51–59.

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., and Martiny, J. B. (2012). Beyond biogeographic patterns: Processes shaping the microbial landscape. *Nature Reviews Microbiology* 10.7, pp. 497–506.

Haq, I. U., Zwahlen, R. D., Yang, P., and Elsas, J. D. van (2018). The response of Paraburkholderia terrae strains to two soil fungi and the potential role of oxalate. *Frontiers in Microbiology* 9.MAY, pp. 1–11.

Hara, S., Morikawa, T., Wasai, S., Kasahara, Y., Koshiba, T., Yamazaki, K., Fujiwara, T., Tokunaga, T., and Minamisawa, K. (2019). Identification of nitrogen-fixing bradyrhizobium associated with roots of field-grown sorghum by metagenome and proteome analyses. *Frontiers in Microbiology* 10.MAR, pp. 1–15.

Hawkings, J. R., Wadham, J. L., Tranter, M., Lawson, E., Sole, A., Cowton, T., Tedstone, A. J., Bartholomew, I., Nienow, P., Chandler, D., and Telling, J. (2015). The effect of warming climate on nutrient and solute export from the Greenland Ice Sheet. *Geochemical Perspectives Letters* 1.1, pp. 94–104.

Hayashi, M., Van Der Kamp, G., and Schmidt, R. (2003). Focused infiltration of snowmelt water in partially frozen soil under small depressions. *Journal of Hydrology* 270.3-4, pp. 214–229.

Heckmann, T., Mccoll, S., and Morche, D. (2016). Retreating ice: Research in pro-glacial areas matters. *Earth Surface Processes and Landforms* 41.2, pp. 271–276.

Hellweger, F. L., Van Sebille, E., and Fredrick, N. D. (2014). Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* 345.6202, pp. 1346–1349.

Hernández-Romero, D., Solano, F., and Sanchez-Amat, A. (2005). Polyphenol oxidase activity expression in Ralstonia solanacearum. *Applied and Environmental Microbiology* 71.11, pp. 6808–6815.

Hespen, C. W., Bruegger, J. J., Guo, Y., and Marletta, M. A. (2018). Native Alanine Substitution in the Glycine Hinge Modulates Conformational Flexibility of Heme Nitric Oxide/Oxygen (H-NOX) Sensing Proteins. *ACS Chemical Biology* 13.6, pp. 1631–1639.

Hibbett, D. (2016). The invisible dimension of fungal diversity. *Science* 351.6278, pp. 1150–1151.

Hillebrand, H., Dürselen, C. D., Kirschtel, D., Pollingher, U., and Zohary, T. (1999). Biovolume calculation for pelagic and benthic microalgae. *Journal of Phycology* 35.2, pp. 403–424.

Hodgkins, R., Tranter, M., and Dowdeswell, J. A. (2004). The characteristics and formation of a High-Arctic proglacial icing. *Geografiska Annaler, Series A: Physical Geography* 86.3, pp. 265–275.

Hoffman, B. M., Lukoyanov, D., Yang, Z. Y., Dean, D. R., and Seefeldt, L. C. (2014). Mechanism of nitrogen fixation by nitrogenase: The next stage. *Chemical Reviews* 114.8, pp. 4041–4062.

Hoham, R. W. and Remias, D. (2020). Snow and Glacial Algae: A Review1. *Journal of Phycology* 56.2, pp. 264–282.

Holmlund, P. and Eriksson, M. (1989). The cold surface layer on Storglaciären. *Geografiska Annaler, Series A* 71, pp. 241–244.

Hood, E., Battin, T. J., Fellman, J., O'neel, S., and Spencer, R. G. (2015). Storage and release of organic carbon from glaciers and ice sheets. *Nature Geoscience* 8, pp. 91–96.

Hooke, R. L. and Pohjola, V. A. (1994). Hydrology of a segment of a glacier situated in an overdeepening, Storglaciären, Sweden. *Journal of Glaciology* 40.134, pp. 140–148.

Hotaling, S., Hood, E., and Hamilton, T. L. (2017). Microbial ecology of mountain glacier ecosystems: biodiversity, ecological connections and implications of a warming climate. *Environmental Microbiology* 19.8, pp. 2935–2948.

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences of the United States of America* 111.13, pp. 4904–4909.

Hsieh, T. C., Ma, K. H., and Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* 7.12, pp. 1451–1456.

Hubbard, B., Tison, J. L., Janssens, L., and Spiro, B. (2000). Ice-core evidence of the thickness and character of clear-facies basal ice: Glacier de Tsanfleuron, Switzerland. *Journal of Glaciology* 46.152, pp. 140–150.

Hudleston, P. J. (2015). Structures and fabrics in glacial ice: A review. *Journal of Structural Geology* 81, pp. 1–27.

Hugerth, L. W., Wefer, H. A., Lundin, S., Jakobsson, H. E., Lindberg, M., Rodin, S., Engstrand, L., and Andersson, A. F. (2014). DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Applied and Environmental Microbiology* 80.16, pp. 5116–5123.

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., Harden, J., Turetsky, M. R., McGuire, A. D., Shah, M. B., VerBerkmoes, N. C., Lee, L. H., Mavrommatis, K., and Jansson, J. K. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* 521.7551, pp. 208–212. arXiv: arXiv:1011.1669v3.

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data MEGAN analysis of metagenomic data. *Genome research* 17.3, pp. 377–386. arXiv: arXiv: 1011.1669v3.

Huvet, M. and Stumpf, M. P. (2014). Overlapping genes: A window on gene evolvability. *BMC Genomics* 15.1, pp. 1–10.

Hyatt, D., Locascio, P. F., Hauser, L. J., and Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28.17, pp. 2223–2230.

in 't Zandt, M. H., Liebner, S., and Welte, C. U. (2020). Roles of Thermokarst Lakes in a Warming World. *Trends in Microbiology* 28.9, pp. 769–779.

Iserte, J. A., Stephan, B. I., Goñi, S. E., Borio, C. S., Ghiringhelli, P. D., and Lozano, M. E. (2013). Family-Specific Degenerate Primer Design: A Tool to Design Consensus Degenerated Oligonucleotides. *Biotechnology Research International* 2013, pp. 1–9.

Jansson, J. K. and Hofmockel, K. S. (2018). The soil microbiome — from metagenomics to metaphenomics. *Current Opinion in Microbiology* 43, pp. 162–168.

Jansson, J. K. and Prosser, J. I. (2013). Microbiology: The life beneath our feet. *Nature* 494.7435, pp. 40–41.

Jansson, J. K. and Taş, N. (2014). The microbial ecology of permafrost. *Nature Reviews Microbiology* 12.6, pp. 414–425.

Jansson, P. (1996). Dynamics and hydrology of a small polythermal valley glacier. *Geografiska Annaler, Series B: Human Geography* 78.2-3, pp. 171–180.

Jetten, M. S. (2008). The microbial nitrogen cycle. *Environmental Microbiology* 10.11, pp. 2903–2909.

Jiang, Y., Lei, Y., Qin, W., Korpelainen, H., and Li, C. (2019). Revealing microbial processes and nutrient limitation in soil through ecoenzymatic stoichiometry and glomalin-related soil proteins in a retreating glacier forefield. *Geoderma* 338, pp. 313–324.

Jiang, Y., Lei, Y., Yang, Y., Korpelainen, H., Niinemets, Ü., and Li, C. (2018). Divergent assemblage patterns and driving forces for bacterial and fungal communities along a glacier forefield chronosequence. *Soil Biology and Biochemistry* 118, pp. 207–216.

Johansson, E., Berglund, S., Lindborg, T., Petrone, J., Van As, D., Gustafsson, L. G., Näslund, J. O., and Laudon, H. (2015). Hydrological and meteorological investigations in a periglacial lake catchment near Kangerlussuaq, west Greenland - Presentation of a new multi-parameter data set. *Earth System Science Data* 7.1, pp. 93–108.

Johnes, P. J. and Heathwaite, A. L. (1992). A procedure for the simultaneous determination of total nitrogen and total phosphorus in freshwater samples using persulphate microwave digestion. *Water Research* 26.10, pp. 1281–1287.

Jørgensen, A. S. and Andreasen, F. (2007). Mapping of permafrost surface using ground-penetrating radar at Kangerlussuaq Airport, western Greenland. *Cold Regions Science and Technology* 48.1, pp. 64–72.

Ju, F. and Zhang, T. (2015). 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Applied Microbiology and Biotechnology* 99.10, pp. 4119–4129.

Juers, D. H., Matthews, B. W., and Huber, R. E. (2012). LacZ β-galactosidase: Structure and function of an enzyme of historical and molecular biological importance. *Protein Science* 21.12, pp. 1792–1807.

Jungblut, A. D. and Vincent, W. F. (2017). Cyanobacteria in polar and alpine ecosystems. *Psychrophiles: From Biodiversity to Biotechnology.* Second Edi. Springer, pp. 181–206.

Kalendar, R., Lee, D., and Schulman, A. (2009). FastPCR software for PCR primer and probe design and repeat search. *Genes, Genomes and Genomics* 3.1, pp. 1–14.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019.7, pp. 1–13.

Kang, H., Kim, H., Lee, B. I., Joung, Y., and Joh, K. (2014). Sediminibacterium goheungense sp. nov., isolated from a freshwater reservoir. *International Journal of Systematic and Evolutionary Microbiology* 64.4, pp. 1328–1333.

Kano, H., Taguchi, S., and Momose, H. (1997). Cold adaptation of a mesophilic serine protease, subtilisin, by in vitro random mutagenesis. *Applied Microbiology and Biotechnology* 47.1, pp. 46–51.

Karan, H., Funk, C., Grabert, M., Oey, M., and Hankamer, B. (2019). Green Bioplastics as Part of a Circular Bioeconomy. *Trends in Plant Science* 24.3, pp. 237–249.

Karigar, C. S. and Rao, S. S. (2011). Role of microbial enzymes in the bioremediation of pollutants: A review. *Enzyme Research* 2011.1, pp. 1–11.

Karimi, B., Terrat, S., Dequiedt, S., Saby, N. P., Horrigue, W., Lelièvre, M., Nowak, V., Jolivet, C., Arrouays, D., Wincker, P., Cruaud, C., Bispo, A., Maron, P. A., Prévost-Bouré, N. C., and Ranjard, L. (2018). Biogeography of soil bacteria and archaea across France. *Science Advances* 4.7, pp. 1–14.

Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., Tameda, M., Shiraki, K., Ito, M., Takei, Y., and Takase, K. (2015). Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterology* 15.1, pp. 1–10.

Kassambara, A. (2018). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.7. *https://CRAN.R-project.org/package=ggpubr.*

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30.4, pp. 772–780.

Kaur, H. and Gill, P. K. (2019). Microbial Enzymes in Food and Beverages Processing. Elsevier Inc., pp. 255–282.

Kayani, M. u. R., Doyle, S. M., Sangwan, N., Wang, G., Gilbert, J. A., Christner, B. C., and Zhu, T. F. (2018). Metagenomic analysis of basal ice from an Alaskan glacier. *Microbiome* 6.1, pp. 14–16.

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research* 40.1.

Kelly, E. F., Chadwick, O. A., and Hilinski, T. E. (1998). The effect of plants on mineral weathering. *Biogeochemistry* 42, pp. 21–53.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research* 12.6, pp. 996–1006.

Khan, S., Farooq, U., and Kurnikova, M. (2016). Exploring protein stability by comparative molecular dynamics simulations of homologous hyperthermophilic, mesophilic, and psychrophilic proteins. *Journal of Chemical Information and Modeling* 56.11, pp. 2129–2139.

Khrapunov, S., Chang, E., and Callender, R. H. (2017). Thermodynamic and Structural Adaptation Differences between the Mesophilic and Psychrophilic Lactate Dehydrogenases. *Biochemistry* 56.28, pp. 3587–3595.

Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., Veen, J. A. van, and Kuramae, E. E. (2016). The ecology of Acidobacteria: Moving beyond genes and genomes. *Frontiers in Microbiology* 7.744, pp. 1–16.

Kim, H., Park, A. K., Lee, J. H., Kim, H. W., and Shin, S. C. (2018). Complete genome sequence of Colwellia hornerae PAMC 20917, a cold-active enzyme-producing bacterium isolated from the Arctic Ocean sediment. *Marine Genomics* 41, pp. 54–56.

Kim, Y. J., Nguyen, N. L., Weon, H. Y., and Yang, D. C. (2013). Sediminibacterium ginsengisoli sp. nov., isolated from soil of a ginseng field, and emended descriptions of the genus Sediminibacterium and of Sediminibacterium salmoneum. *International Journal of Systematic and Evolutionary Microbiology* 63.3, pp. 905–912.

King, O., Bhattacharya, A., Bhambri, R., and Bolch, T. (2019). Glacial lakes exacerbate Himalayan glacier mass loss. *Scientific Reports*.

Kivisaar, M. (2003). Stationary phase mutagenesis: Mechanisms that accelerate adaptation of microbial populations under environmental stress. *Environmental Microbiology* 5.10, pp. 814–827.

Klassen, J. L. and Foght, J. M. (2011). Characterization of Hymenobacter isolates from Victoria Upper Glacier, Antarctica reveals five new species and substantial non-vertical evolution within this genus. *Extremophiles* 15.1, pp. 45–57.

Kmezik, C., Bonzom, C., Olsson, L., Mazurkewich, S., and Larsbrink, J. (2020). Multimodular fused acetyl-feruloyl esterases from soil and gut Bacteroidetes improve xylanase depolymerization of recalcitrant biomass. *Biotechnology for Biofuels* 13.1, pp. 1–14.

Knelman, J. E., Legg, T. M., O'Neill, S. P., Washenberger, C. L., González, A., Cleveland, C. C., and 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, pp. 172–180.

Knowlton, C., Veerapaneni, R., D'Elia, T., and Rogers, S. O. (2013). Microbial analyses of ancient ice core sections from Greenland and Antarctica. *Biology* 2.1, pp. 206–232.

Kokubo, H., Harris, R. C., Asthagiri, D., and Pettitt, B. M. (2013). Solvation free energies of alanine peptides: The effect of flexibility. *Journal of Physical Chemistry B* 117.51, pp. 16428–16435.

Kolmogorov, M., Rayko, M., Yuan, J., Polevikov, E., and Pevzner, P. (2019). metaFlye: scalable long-read metagenome assembly using repeat graphs. *bioRxiv*, p. 637637.

Kono, N. and Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. *Development Growth and Differentiation* 61.5, pp. 316–326.

Koshila Ravi, R., Anusuya, S., Balachandar, M., and Muthukumar, T. (2019). Microbial Interactions in Soil Formation and Nutrient Cycling. *Mycorrhizosphere and Pedogenesis.* February 2020, pp. 363–382.

Krewulak, K. D. and Vogel, H. J. (2008). Structural biology of bacterial iron uptake. *Biochimica et Biophysica Acta - Biomembranes* 1778.9, pp. 1781–1804.

Kryukova, M. V., Petrovskaya, L. E., Kryukova, E. A., Lomakina, G. Y., Yakimov, S. A., Maksimov, E. G., Boyko, K. M., Popov, V. O., Dolgikh, D. A., and Kirpichnikov, M. P. (2019). Thermal inactivation of a cold-active esterase PMGL3 isolated from the permafrost metagenomic library. *Biomolecules* 9.12, pp. 1–13.

Kujawinski, E. B. (2017). Cryospheric science: The power of glacial microbes. *Nature Geoscience* 10.5, pp. 329–330.

Kulkarni, G., Busset, N., Molinaro, A., Gargani, D., Chaintreuil, C., Silipo, A., and Giraud, E. (2015). Specific Hopanoid Classes Differentially Affect Free-Living and. 6.5, pp. 1–9.

Kumar, R., Kumar, P., and Giri, A. (2019). Regional impact of psychrophilic bacteria on bioremediation. *Smart Bioremediation Technologies: Microbial Enzymes.* Chap. 7, pp. 119–135.

Lan, S., Wu, L., Zhang, D., Hu, C., and Liu, Y. (2010). Effects of drought and salt stresses on man-made cyanobacterial crusts. *European Journal of Soil Biology* 46.6, pp. 381–386.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9.4, pp. 357–359.

Larkin, A. A. and Martiny, A. C. (2017). Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. *Environmental Microbiology Reports* 9.2, pp. 55–70.

Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* 32.1, pp. 11–16.

Laybourn-Parry, J., Tranter, M., and Hodson, A. J. (2012). An introduction to ice environments and their biology. *The Ecology of Snow and Ice Environments.* Chap. An introdu, pp. 1–36.

Lebre, P. H., De Maayer, P., and Cowan, D. A. (2017). Xerotolerant bacteria: Surviving through a dry spell. *Nature Reviews Microbiology* 15.5, pp. 285–296.

Lee, K. C., Morgan, X. C., Dunfield, P. F., Tamas, I., McDonald, I. R., and Stott, M. B. (2014). Genomic analysis of Chthonomonas calidirosea, the first sequenced isolate of the phylum Armatimonadetes. *ISME Journal* 8.7, pp. 1522–1533.

Legendre, P. and Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia* 129.2, pp. 271–280.

Lennon, J. T. and Jones, S. E. (2011). Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology* 9.2, pp. 119–130.

Lennon, J. T. and Locey, K. J. (2020). More support for Earth's massive microbiome. *Biology Direct* 15.1, pp. 1–6.

Lepleux, C., Turpault, M. P., Oger, P., Frey-Klett, P., and Uroz, S. (2012). Correlation of the abundance of betaproteobacteria on mineral surfaces with mineral weathering in forest soils. *Applied and Environmental Microbiology* 78.19, pp. 7114–7119.

Lesniewski, R. A., Jain, S., Anantharaman, K., Schloss, P. D., and Dick, G. J. (2012). The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME Journal* 6.12, pp. 2257–2268.

Lever, M. A. and Teske, A. P. (2015). Diversity of methane-cycling archaea in hydrothermal sediment investigated by general and group-specific PCR primers. *Applied and Environmental Microbiology* 81.4, pp. 1426–1441.

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31.10, pp. 1674–1676.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34.18, pp. 3094–3100. arXiv: 1708.01492.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26.5, pp. 589–595. arXiv: 1303.3997.

Li, Q., Liu, Y., Gu, Y., Guo, L., Huang, Y., Zhang, J., Xu, Z., Tan, B., Zhang, L., Chen, L., Xiao, J., and Zhu, P. (2020). Ecoenzymatic stoichiometry and microbial nutrient limitations in rhizosphere soil along the Hailuogou Glacier forefield chronosequence. *Science of the Total Environment* 704, p. 135413.

Li, X., Li, Z., Ding, Y., Liu, S., Zhao, Z., Luo, L., Pang, H., Li, C., Li, H., You, X., and Wang, F. (2007). Seasonal variations of pH and electrical conductivity in a snow-firn pack on Glacier No. 1, eastern Tianshan, China. *Cold Regions Science and Technology* 48.1, pp. 55–63.

Linhart, C. and Shamir, R. (2005). The degenerate primer design problem: Theory and applications. *Journal of Computational Biology* 12.4, pp. 431–456.

Liu, C., Song, Y., McTeague, M., Vu, A. W., Wexler, H., and Finegold, S. M. (2003). Rapid identification of the species of the Bacteroides fragilis group by multiplex PCR assays using group- and species-specific primers. *FEMS Microbiology Letters* 222.1, pp. 9–16.

Liu, D. F., Lian, B., and Wang, B. (2016). Solubilization of potassium containing minerals by high temperature resistant Streptomyces sp. isolated from earthworm's gut. *Acta Geochimica* 35.3, pp. 262–270.

Liu, K., Liu, Y., Han, B. P., Xu, B., Zhu, L., Ju, J., Jiao, N., and Xiong, J. (2019). Bacterial community changes in a glacial-fed Tibetan lake are correlated with glacial melting. *Science of the Total Environment* 651, pp. 2059–2067.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. arXiv: 25136.

Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., and Bai, Y. (2020). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein and Cell.*

Livermore, J. A. and Jones, S. E. (2015). Local-global overlap in diversity informs mechanisms of bacterial biogeography. *ISME Journal* 9.11, pp. 2413–2422.

Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America* 113.21, pp. 5970–5975.

Loladze, V. V., Ermolenko, D. N., and Makhatadze, G. I. (2002). Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior. *Journal of Molecular Biology* 320.2, pp. 343–357.

Loranty, M. M., Abbott, B. W., Blok, D., Douglas, T. A., Epstein, H. E., Forbes, B. C., Jones, B. M., Kholodov, A. L., Kropp, H., Malhotra, A., Mamet, S. D., Myers-Smith, I. H., Natali,

S. M., O'Donnell, J. A., Phoenix, G. K., Rocha, A. V., Sonnentag, O., Tape, K. D., and Walker, D. A. (2018). Reviews and syntheses: Changing ecosystem influences on soil thermal regimes in northern high-latitude permafrost regions. *Biogeosciences* 15.17, pp. 5287–5313.

Louca, S., Mazel, F., Doebeli, M., and Parfrey, L. W. (2019). A census-based estimate of earth's bacterial and archaeal diversity. *PLoS Biology* 17.2, pp. 1–30.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15.550, pp. 1–21.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, A., Buchner, A., Lai, T., Steppi, S., Jacob, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßbmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., and Schleifer, K. H. (2004). ARB: A software environment for sequence data. *Nucleic Acids Research* 32.4, pp. 1363–1371.

Luláková, P., Perez-Mon, C., Šantrůčková, H., Ruethi, J., and Frey, B. (2019). High-alpine permafrost and active-layer soil microbiomes differ in their response to elevated temperatures. *Frontiers in Microbiology* 10.APR, pp. 1–16.

Lutz, S., Anesio, A. M., Edwards, A., and Benning, L. G. (2015). Microbial diversity on icelandic glaciers and ice caps. *Frontiers in Microbiology* 6, p. 307.

– (2017). Linking microbial diversity and functionality of arctic glacial surface habitats. *Environmental Microbiology* 19.2, pp. 551–565.

Lutz, S., Anesio, A. M., Raiswell, R., Edwards, A., Newton, R. J., Gill, F., and Benning, L. G. (2016). The biogeography of red snow microbiomes and their role in melting arctic glaciers. *Nature Communications* 7.May, pp. 1–9.

MacDonell, S. and Fitzsimons, S. (2008). The formation and hydrological significance of cryoconite holes. *Progress in Physical Geography* 32.6, pp. 595–610.

MacKelprang, R., Burkert, A., Haw, M., Mahendrarajah, T., Conaway, C. H., Douglas, T. A., and Waldrop, M. P. (2017). Microbial survival strategies in ancient permafrost: Insights from metagenomics. *ISME Journal* 11.10, pp. 2305–2318.

MacKelprang, R., Waldrop, M. P., Deangelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., Rubin, E. M., and Jansson, J. K. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480.7377, pp. 368–371.

Mader, H. M. (1992). Observations of the water-vein system in polycrystalline ice. *Journal of Glaciology* 38.130, pp. 333–347.

Mader, H. M., Pettitt, M. E., Wadham, J. L., Wolff, E. W., and Parkes, R. J. (2006). Subsurface ice as a microbial habitat. *Geology* 34.3, pp. 169–172.

Madhavan, A., Sindhu, R., Parameswaran, B., Sukumaran, R. K., and Pandey, A. (2017). Metagenome Analysis: a Powerful Tool for Enzyme Bioprospecting. *Applied Biochemistry and Biotechnology* 183.2, pp. 636–651.

Makhalanyane, T. P., Valverde, A., Velázquez, D., Gunnigle, E., Van Goethem, M. W., Quesada, A., and Cowan, D. A. (2015). Ecology and biogeochemistry of cyanobacteria in soils, permafrost, aquatic and cryptic polar habitats. *Biodiversity and Conservation* 24.4, pp. 819–840.

Malard, L. A., Anwar, M. Z., Jacobsen, C. S., and Pearce, D. A. (2019). Biogeographical patterns in soil bacterial communities across the Arctic region. *FEMS Microbiology Ecology* 95.9, pp. 1–13.

Malard, L. A. and Pearce, D. A. (2018). Microbial diversity and biogeography in Arctic soils. *Environmental Microbiology Reports* 10.6, pp. 611–625.

Malisorn, K., Embaen, S., Sribun, A., Saeng-in, P., Phongsopitanun, W., and Tanasupawat, S. (2020). Identification and antimicrobial activities of Streptomyces, Micromonospora, and Kitasatospora strains from rhizosphere soils. *Journal of Applied Pharmaceutical Science* 10.2, pp. 123–128.

Mancabelli, L., Milani, C., Lugli, G. A., Turroni, F., Cocconi, D., Sinderen, D. van, and Ventura, M. (2017). Identification of universal gut microbial biomarkers of common human intestinal diseases by meta-analysis. *FEMS Microbiology Ecology* 93.12, pp. 1–10.

Mangiagalli, M., Brocca, S., Orlando, M., and Lotti, M. (2020). The “cold revolution”. Present and future applications of cold-active enzymes and ice-binding proteins. *New Biotechnology* 55.September 2019, pp. 5–11.

Margesin, R. and Collins, T. (2019). Microbial ecology of the cryosphere (glacial and permafrost habitats): current knowledge. *Applied Microbiology and Biotechnology* 103.6, pp. 2537–2549.

Markham, N. R. and Zuker, M. (2008). UNAFold. *Bioinformatics*. Ed. by K. J.M. Springer, pp. 3–31.

Marshall, S. J. (2005). Recent advances in understanding ice sheet dynamics. *Earth and Planetary Science Letters* 240.2, pp. 191–204.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17.1, p. 10. arXiv: ISSN2226-6089.

Martinez Arbizu, P. (2020). pairwiseAdonis: Pairwise multilevel comparison using adonis.

Martinez-Alonso, E., Pena-Perez, S., Serrano, S., Garcia-Lopez, E., Alcazar, A., and Cid, C. (2019). Taxonomic and functional characterization of a microbial community from a volcanic englacial ecosystem in Deception Island, Antarctica. *Scientific Reports* 9.1, pp. 1–14.

Marx, V. (2013). The big challenges of big data. *Nature* 498.7453, pp. 255–260.

Maurer, J. M., Schaefer, J. M., Rupper, S., and Corley, A. (2019). Acceleration of ice loss across the Himalayas over the past 40 years. *Science Advances* 5.6.

McAnulty, M. J., Poosarla, V. G., Kim, K. Y., Jasso-Chávez, R., Logan, B. E., and Wood, T. K. (2017). Electricity from methane by reversing methanogenesis. *Nature Communications* 8.May.

McCarthy, A., Chiang, E., Schmidt, M. L., and Denef, V. J. (2015). RNA preservation agents and nucleic acid extraction method bias perceived bacterial community composition. *PLoS ONE* 10.3, pp. 1–14.

McMurdie, P. J. and Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* 8.4, e61217.

Méheust, R., Castelle, C. J., Carnevali, P. B. M., Farag, I. F., He, C., Chen, L.-X., Amano, Y., Hug, L. A., and Banfield, J. F. (2019). Aquatic Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have novel nitrogenase-like gene clusters. *bioRxiv*.

Meier, M. F., Dyurgerov, M. B., Rick, U. K., O'Neel, S., Pfeffer, W. T., Anderson, R. S., Anderson, S. P., and Glazovsky, A. F. (2007). Glaciers dominate eustatic sea-level rise in the 21st century. *Science* 317.5841, pp. 1064–1067.

Miethke, M. (2013). Molecular strategies of microbial iron assimilation: From high-affinity complexes to cofactor assembly systems. *Metallomics* 5.1, pp. 15–28.

Miller, B. R. and Gulick, A. M. (2016). Structural biology of nonribosomal peptide synthetases. *Methods in Molecular Biology*.

Milner, A. M., Khamis, K., Battin, T. J., Brittain, J. E., Barrand, N. E., Füreder, L., Cauvy-Fraunié, S., Gíslason, G. M., Jacobsen, D., Hannah, D. M., Hodson, A. J., Hood, E., Lencioni, V., Ólafsson, J. S., Robinson, C. T., Tranter, M., and Brown, L. E. (2017). Glacier shrinkage driving global changes in downstream systems. *Proceedings of the National Academy of Sciences of the United States of America* 114.37, pp. 9770–9778.

Miri, S., Naghdi, M., Rouissi, T., Kaur Brar, S., and Martel, R. (2019). Recent biotechnological advances in petroleum hydrocarbons degradation under cold climate conditions: A review. *Critical Reviews in Environmental Science and Technology* 49.7, pp. 553–586.

Mitchell, A. C., Lafrenière, M. J., Skidmore, M. L., and Boyd, E. S. (2013). Influence of bedrock mineral composition on microbial diversity in a subglacial environment. *Geology* 41.8, pp. 855–858.

Miteva, V. (2008). Bacteria in snow and glacier ice. *Psychrophiles: From Biodiversity to Biotechnology*, pp. 31–50.

Mocali, S., Chiellini, C., Fabiani, A., Decuzzi, S., Pascale, D., Parrilli, E., Tutino, M. L., Perrin, E., Bosi, E., Fondi, M., Lo Giudice, A., and Fani, R. (2017). Ecology of cold environments: New insights of bacterial metabolic adaptation through an integrated genomic-phenomic approach. *Scientific Reports* 7.1, pp. 1–13.

Mondav, R., McCalley, C. K., Hodgkins, S. B., Frolking, S., Saleska, S. R., Rich, V. I., Chanton, J. P., and Crill, P. M. (2017). Microbial network, phylogenetic diversity and community membership in the active layer across a permafrost thaw gradient. *Environmental Microbiology* 19.8, pp. 3201–3218.

Moon, T., Ahlstrøm, A., Goelzer, H., Lipscomb, W., and Nowicki, S. (2018). Rising Oceans Guaranteed: Arctic Land Ice Loss and Sea Level Rise. *Current Climate Change Reports* 4.3, pp. 211–222.

Morales, S. E. and Holben, W. E. (2009). Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by in silico analysis. *Applied and Environmental Microbiology* 75.9, pp. 2677–2683.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L. K., Meng, J., Durham, B. P., Shen, C., Varaljay, V. A., Smith, C. B., Yager, P. L., and Hopkinson, B. M. (2013). Sizing up metatranscriptomics. *ISME Journal* 7.2, pp. 237–243.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., and Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics* 24.16, pp. 1757–1764.

Müller, A. L., De Rezende, J. R., Hubert, C. R., Kjeldsen, K. U., Lagkouvardos, I., Berry, D., Jørgensen, B. B., and Loy, A. (2014). Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. *ISME Journal* 8.6, pp. 1153–1165.

Müller, O., Bang-Andreasen, T., White, R. A., Elberling, B., Taş, N., Kneafsey, T., Jansson, J. K., and Øvreås, L. (2018). Disentangling the complexity of permafrost soil by using high resolution profiling of microbial community composition, key functions and respiration rates. *Environmental Microbiology* 20.12, pp. 4328–4342.

Musilova, M., Tranter, M., Bennett, S. A., Wadham, J., and Anesio, A. M. (2015). Stable microbial community composition on the Greenland Ice Sheet. *Frontiers in Microbiology* 6, pp. 1–10.

Muyzer, G., De Waal, E. C., and Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* 59.3, pp. 695–700.

Mykytczuk, N. C., Foote, S. J., Omelon, C. R., Southam, G., Greer, C. W., and Whyte, L. G. (2013). Bacterial growth at -15 °C; molecular insights from the permafrost bacterium Planococcus halocryophilus Or1. *ISME Journal* 7.6, pp. 1211–1226.

Nakata, P. A. (2011). The oxalic acid biosynthetic activity of Burkholderia mallei is encoded by a single locus. *Microbiological Research* 166.7, pp. 531–538.

Napieralski, S. A., Buss, H. L., Brantley, S. L., Lee, S., Xu, H., and Roden, E. E. (2019). Microbial chemolithotrophy mediates oxidative weathering of granitic bedrock. *Proceedings of the National Academy of Sciences of the United States of America* 116.52, pp. 26394–26401.

Nash, M. V., Anesio, A. M., Barker, G., Tranter, M., Varliero, G., Eloe-Fadrosh, E. A., Nielsen, T., Turpin-Jelfs, T., Benning, L. G., and Sánchez-Baracaldo, P. (2018). Metagenomic insights into diazotrophic communities across Arctic glacier forefields. *FEMS Microbiology Ecology* 94.9, pp. 1–12.

Naylor, D., Degraaf, S., Purdom, E., and Coleman-Derr, D. (2017). Drought and host selection influence bacterial community dynamics in the grass root microbiome. *ISME Journal* 11.12, pp. 2691–2704.

Nerem, R. S., Beckley, B. D., Fasullo, J. T., Hamlington, B. D., Masters, D., and Mitchum, G. T. (2018). Climate-change–driven accelerated sea-level rise detected in the altimeter era. *Proceedings of the National Academy of Sciences of the United States of America* 115.9, pp. 2022–2025.

Newman, D. J. and Cragg, G. M. (2016). Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* 79.3, pp. 629–661.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32.1, pp. 268–274.

Nicholls, S. M., Quick, J. C., Tang, S., and Loman, N. J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience* 8.5, pp. 1–9.

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., and Pettersson, S. (2012). Metabolic Interactions. *Science* 108.June, pp. 1262–1268.

Nienow, P. W., Sole, A. J., Slater, D. A., and Cowton, T. R. (2017). Recent Advances in Our Understanding of the Role of Meltwater in the Greenland Ice Sheet System. *Current Climate Change Reports* 3.4, pp. 330–344.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research* 27.5, pp. 824–834. arXiv: 1604.03071.

Nye, J. and Frank, F. (1973). Hydrology of the Intergranular Veins in a Temperate Glacier. *Symposium on the Hydrology of Glaciers.*

Oh, J., Goo, E., Hwang, I., and Rhee, S. (2014). Structural basis for bacterial quorum sensing-mediated oxalogenesis. *Journal of Biological Chemistry* 289.16, pp. 11465–11475.

Okoniewska, M., Tanaka, T., and Yada, R. Y. (2000). and Participates in Catalysis. *Society* 177, pp. 169–177.

Oksanen, J. (2017). Vegan: ecological diversity.

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., and Pop, M. (2018). Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics* 20.4, pp. 1140–1150.

Olsson-Francis, K., Boardman, C. P., Pearson, V. K., Schofield, P. F., Oliver, A., and Summers, S. (2015). A Culture-Independent and Culture-Dependent Study of the Bacterial Community from the Bedrock Soil Interface. *Advances in Microbiology* 05.13, pp. 842–857.

O'Malley, M. A. (2007). The nineteenth century roots of 'everything is everywhere'. *Nature Reviews Microbiology* 5, pp. 647–651.

O'Neel, S., Hood, E., Bidlack, A. L., Fleming, S. W., Arimitsu, M. L., Arendt, A., Burgess, E., Sergeant, C. J., Beaudreau, A. H., Timm, K., Hayward, G. D., Reynolds, J. H., and Pyare, S. (2015). Icefield-to-ocean linkages across the northern pacific coastal temperate rainforest ecosystem. *BioScience* 65.5, pp. 499–512.

Oshkin, I. Y., Kulichevskaya, I. S., Rijpstra, W. I. C., Damste, J. S., Rakitin, A. L., Ravin, N. V., and Dedysh, S. N. (2019). Granulicella sibirica sp. Nov., a psychrotolerant acidobacterium isolated from an organic soil layer in forested tundra, West Siberia. *International Journal of Systematic and Evolutionary Microbiology* 69.4, pp. 1195–1201.

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., and Iliopoulos, I. (2015). Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights* 9, pp. 75–88.

Owczarzy, R., Moreira, B. G., You, Y., Behlke, M. A., and Wälder, J. A. (2008). Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry* 47.19, pp. 5336–5353.

Owczarzy, R., You, Y., Moreira, B. G., Manthey, J. A., Huang, L., Behlke, M. A., and Walder, J. A. (2004). Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures. *Biochemistry* 43.12, pp. 3537–3554.

Palmieri, F., Estoppey, A., House, G. L., Lohberger, A., Bindschedler, S., Chain, P. S., and Junier, P. (2019). Oxalic acid, a molecule at the crossroads of bacterial-fungal interactions. 1st ed. Vol. 106. Elsevier Inc., pp. 49–77.

Panadare, D. and Rathod, V. K. (2018). Extraction and purification of polyphenol oxidase: A review. *Biocatalysis and Agricultural Biotechnology* 14.February, pp. 431–437.

Pankratov, T. A. and Dedysh, S. N. (2010). Granulicella paludicola gen. nov., sp. nov., Granulicella pectinivorans sp. nov., Granulicella aggregans sp. nov. and Granulicella rosea sp. nov., acidophilic, polymer-degrading acidobacteria from Sphagnum peat bogs. *International Journal of Systematic and Evolutionary Microbiology* 60.12, pp. 2951–2959.

Parada, A. E., Needham, D. M., and Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology* 18.5, pp. 1403–1414.

Park, M., Won, J., Choi, B. Y., and Lee, C. J. (2020). Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. *Experimental & molecular medicine* 52, pp. 963–977.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25.7, pp. 1043–1055.

Pasternak, K., Kocot, J., and Horecka, A. (2010). Biochemistry of magnesium. *Journal of Elemntology* 15.3/2010, pp. 601–616.

Paul, C., Filippidou, S., Jamil, I., Kooli, W., House, G. L., Estoppey, A., Hayoz, M., Junier, T., Palmieri, F., Wunderlin, T., Lehmann, A., Bindschedler, S., Vennemann, T., Chain, P. S., and Junier, P. (2019). Bacterial spores, from ecology to biotechnology. *Advances in Applied Microbiology*. 1st ed. Vol. 106. Elsevier Inc. Chap. 3, pp. 79–111.

Pavesi, A., Magiorkinis, G., and Karlin, D. G. (2013). Viral Proteins Originated De Novo by Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of Deltaretroviruses. *PLoS Computational Biology* 9.8.

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 35.13, pp. 2193–2198.

Perl, D., Mueller, U., Heinemann, U., and Schmid, F. X. (2000). Two exposed amino acid residues confer thermostability on a cold shock protein. *Nature Structural Biology* 7.5, pp. 380–383.

Pernice, M., Simpson, S. J., and Ponton, F. (2014). Towards an integrated understanding of gut microbiota using insects as model systems. *Journal of Insect Physiology* 69.C, pp. 12–18.

Perolo, P., Bakker, M., Gabbud, C., Moradi, G., Rennie, C., and Lane, S. N. (2019). Subglacial sediment production and snout marginal ice uplift during the late ablation season of a temperate valley glacier. *Earth Surface Processes and Landforms* 44.5, pp. 1117–1136.

Peyret, N., Seneviratne, P. A., Allawi, H. T., and SantaLucia, J. (1999). Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. *Biochemistry* 38.12, pp. 3468–3477.

Pinto, F., Tett, A., Armanini, F., Asnicar, F., Boscaini, A., Pasolli, E., Zolfo, M., Donati, C., Salmaso, N., and Segata, N. (2017). Draft genome sequence of the planktic cyanobacterium Tychonema bourrellyi, isolated from Alpine lentic freshwater. *Genome Announcements* 5.47, pp. 49–50.

Pochon, X., Zaiko, A., Fletcher, L. M., Laroche, O., and Wood, S. A. (2017). Wanted dead or alive? Using metabarcoding of environmental DNA and RNA to distinguish living assemblages for biosecurity applications. *PLoS ONE* 12.11, pp. 1–19.

Pohjola, V. A. (1996). Simulation of particle paths and deformation of ice structures along a flow-line on Storglaciären, Sweden. *Geografiska Annaler, Series B: Human Geography* 78.2-3, pp. 181–192.

Porder, S. (2019). How Plants Enhance Weathering and How Weathering is Important to Plants. *Elements*.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* 26.7, pp. 1641–1650.

Price, P. B. (2000). A habitat for psychrophiles in deep Antarctic ice. *Proceedings of the National Academy of Sciences* 97.3, pp. 1247–1251.

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., and Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE* 15.1, pp. 1–19.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35.9, pp. 833–844.

R Core Team 2019 (2019). : A language and environment for statistical computing.

Raiswell, R. and Canfield, D. E. (2012). The iron biogeochemical cycle past and present. *Geochemical Perspectives* 1.1, pp. 1–2.

Ram, Y. and Hadany, L. (2014). Stress-induced mutagenesis and complex adaptation. *Proceedings of the Royal Society B: Biological Sciences* 281.1792.

Rang, F. J., Kloosterman, W. P., and Ridder, J. de (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology* 19.1, pp. 1–11.

Rassoulzadegan, F. and Sheldon, R. W. (1986). Predator-prey interactions of nanozooplankton and bacteria in an oligotrophic marine environment. *Limnology and Oceanography* 31.5, pp. 1010–1029.

Razavi, B. S., Liu, S., and Kuzyakov, Y. (2017). Hot experience for cold-adapted microorganisms: Temperature sensitivity of soil enzymes. *Soil Biology and Biochemistry* 105, pp. 236–243.

Ren, B., Hu, Y., Chen, B., Zhang, Y., Thiele, J., Shi, R., Liu, M., and Bu, R. (2018). Soil pH and plant diversity shape soil bacterial community structure in the active layer across the latitudinal gradients in continuous permafrost region of Northeastern China. *Scientific Reports* 8.1, pp. 1–10.

Reynolds, H. L., Packer, A., Bever, J. D., and Clay, K. (2003). Grassroots ecology: Plant-microbe-soil interactions as drivers of plant community structure and dynamics. *Ecology* 84.9, pp. 2281–2291.

Rime, T., Hartmann, M., Brunner, I., Widmer, F., Zeyer, J., and Frey, B. (2015). Vertical distribution of the soil microbiota along a successional gradient in a glacier forefield. *Molecular Ecology* 24.5, pp. 1091–1108.

Rime, T., Hartmann, M., and Frey, B. (2016). Potential sources of microbial colonizers in an initial soil ecosystem after retreat of an alpine glacier. *ISME Journal* 10.7, pp. 1625–1641.

Rose, T. M., Henikoff, J. G., and Henikoff, S. (2003). CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Research* 31.13, pp. 3763–3766.

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., and Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in Genetics* 8, p. 23.

Rousk, J. and Bengtson, P. (2014). Microbial regulation of global biogeochemical cycles. *Frontiers in Microbiology* 5, p. 103.

Rubio, L. M. and Ludden, P. W. (2008). Biosynthesis of the Iron-Molybdenum Cofactor of Nitrogenase. *Annual Review of Microbiology* 62.1, pp. 93–111.

Rudnick, P., Meletzus, D., Green, A., He, L., and Kennedy, C. (1997). Regulation of nitrogen fixation by ammonium in diazotrophic species of proteobacteria. *Soil Biology and Biochemistry* 29.5-6, pp. 831–841.

Ruijter, J. M., Ramakers, C., Hoogaars, W. M., Karlen, Y., Bakker, O., Van den hoff, M. J., and Moorman, A. F. (2009). Amplification efficiency: Linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Research* 37.6.

Ryall, B., Eydallin, G., and Ferenci, T. (2012). Culture History and Population Heterogeneity as Determinants of Bacterial Adaptation: the Adaptomics of a Single Environmental Transition. *Microbiology and Molecular Biology Reviews* 76.3, pp. 597–625.

Rychlik, W. (2007). OLIGO 7 primer analysis software. *Methods in Molecular Biology* 402, pp. 35–59.

Salama, E. S., Jeon, B. H., Kurade, M. B., Patil, S. M., Usman, M., Li, X., and Lim, H. (2020). Enhanced anaerobic co-digestion of fat, oil, and grease by calcium addition: Boost of biomethane production and microbial community shift. *Bioresource Technology* 296.November 2019, p. 122353.

Salazar, G., Cornejo-Castillo, F. M., Benítez-Barrios, V., Fraile-Nuez, E., Álvarez-Salgado, X. A., Duarte, C. M., Gasol, J. M., and Acinas, S. G. (2016). Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME Journal* 10.3, pp. 596–608.

Salwoom, L., Rahman, R. N. Z. R. A., Salleh, A. B., Shariff, F. M., Convey, P., Pearce, D., and Ali, M. S. M. (2019). Isolation, characterisation, and lipase production of a cold-adapted bacterial strain Pseudomonas sp. LSK25 isolated from Signy Island, Antarctica. *Molecules* 24.4, pp. 1–14.

Samuels, T., Bryce, C., Landenmark, H., Marie-Loudon, C., Nicholson, N., Stevens, A. H., and Cockell, C. (2020). Microbial Weathering of Minerals and Rocks in Natural Environments. *Biogeochemical Cycles: Ecological Drivers and Environmental Impact*. Chap. 3.

Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4, pp. 1–11.

SantaLucia, J. and Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure* 33, pp. 415–440.

Santos, E. H. dos, Yamamoto, L., Domingues, W., Santi, S. M. di, Kanunfre, K. A., and Okay, T. S. (2020). A new Real Time PCR with species-specific primers from Plasmodium malariae/P. brasilianum mitochondrial cytochrome b gene. *Parasitology International* 76.January, p. 102069.

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., Funk, K., Ketter, A., Kim, S., Kimchi, A., Kitts, P. A., Kuznetsov, A., Lathrop, S., Lu, Z., McGarvey, K., Madden, T. L., Murphy, T. D., O'Leary, N., Phan, L., Schneider, V. A., Thibaud-Nissen, F., Trawick, B. W., Pruitt, K. D., and Ostell, J. (2020). Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 48.D1, pp. D9–D16.

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research* 47.D1, pp. D94–D99.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A. (2012). Fiji: An open-source platform for biological-image analysis. *Nature Methods* 9.7, pp. 676–682.

Schloss, P. D. and Handelsman, J. (2006). Toward a census of bacteria in soil. *PLoS Computational Biology* 2.7, pp. 0786–0793.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75.23, pp. 7537–7541.

Schmalenberger, A. and Noll, M. (2010). Shifts in desulfonating bacterial communities along a soil chronosequence in the forefield of a receding glacier. *FEMS Microbiology Ecology* 71.2, pp. 208–217.

Schmidt, S. K., Reed, S. C., Nemergut, D. R., Grandy, A. S., Cleveland, C. C., Weintraub, M. N., Hill, A. W., Costello, E. K., Meyer, A. F., Neff, J. C., and 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 B: Biological Sciences* 275.1653, pp. 2793–2802.

Schostag, M., Stibal, M., Jacobsen, C. S., Bælum, J., Tas, N., Elberling, B., Jansson, J. K., Semenchuk, P., and Priemé, A. (2015). Distinct summer and winter bacterial communities in the active layer of Svalbard permafrost revealed by DNA- and RNA-based analyses. *Frontiers in Microbiology* 6.APR, pp. 1–13.

Schuur, E. A., McGuire, A. D., Schädel, C., Grosse, G., Harden, J. W., Hayes, D. J., Hugelius, G., Koven, C. D., Kuhry, P., Lawrence, D. M., Natali, S. M., Olefeldt, D., Romanovsky, V. E., Schaefer, K., Turetsky, M. R., Treat, C. C., and Vonk, J. E. (2015). Climate change and the permafrost carbon feedback. arXiv: arXiv:1011.1669v3.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30.14, pp. 2068–2069.

Segawa, T., Takeuchi, N., Rivera, A., Yamada, A., Yoshimura, Y., Barcaza, G., Shinbori, K., Motoyama, H., Kohshima, S., and Ushida, K. (2013). Distribution of antibiotic resistance genes in glacier environments. *Environmental Microbiology Reports* 5.1, pp. 127–134.

Segawa, T., Yonezawa, T., Edwards, A., Akiyoshi, A., Tanaka, S., Uetake, J., Irvine-Fynn, T., Fukui, K., Li, Z., and Takeuchi, N. (2017). Biogeography of cryoconite forming cyanobacteria on polar and Asian glaciers. *Journal of Biogeography* 44.12, pp. 2849–2861.

Sellstedt, A. and Richau, K. H. (2013). Aspects of nitrogen-fixing actinobacteria, in particular free-living and symbiotic frankia. *FEMS Microbiology Letters* 342.2, pp. 179–186.

Selzer, P. M., Marhöfer, R. J., and Koch, O. (2018). Biological Databases. *Applied Bioinformatics: An Introduction.* Cham: Springer International Publishing, pp. 13–34.

Serebriiskii, I. G. and Golemis, E. A. (2000). Uses of lacZ to study gene function: Evaluation of β-galactosidase assays employed in the yeast two-hybrid system. *Analytical Biochemistry* 285.1, pp. 1–15.

Setlow, P. (2016). Spore Resistance Properties. *The Bacterial Spore: From Molecules to Systems.* Chap. 10.

Shafee, T. and Lowe, R. (2018). Eukaryotic and Prokaryotic Gene Structure. *SSRN Electronic Journal* 4.1, pp. 2002–4436.

Shah, V. and Subramaniam, S. (2018). Bradyrhizobium japonicum USDA110: A representative model organism for studying the impact of pollutants on soil microbiota. *Science of the Total Environment* 624, pp. 963–967.

Shaiber, A. and Eren, A. M. (2019). Composite metagenome-assembled genomes reduce the quality of public genome repositories. *mBio* 10.3, pp. 1–3.

Sharma, B., Dangi, A. K., and Shukla, P. (2018). Contemporary enzyme based technologies for bioremediation: A review. *Journal of Environmental Management* 210, pp. 10–22.

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science* 5. arXiv: 15334406.

Shcherbakova, V. and Troshina, O. (2018). Biotechnological perspectives of microorganisms isolated from the Polar Regions. *Microbiology Australia* 39.3, pp. 137–140.

Shen, B. (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology* 7.2, pp. 285–295.

Shin, S. C., Kim, S. J., Ahn, D. H., Lee, J. K., Lee, H., Lee, J., Hong, S. G., Lee, Y. M., and Park, H. (2012). Genome sequence of a Salinibacterium sp. Isolated from antarctic soil. *Journal of Bacteriology* 194.9, pp. 2404–2404.

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* 21.8, pp. 1794–1805.

Siddiqui, K. S. (2015). Some like it hot, some like it cold: Temperature dependent biotechnological applications and improvements in extremophilic enzymes. *Biotechnology Advances* 33.8, pp. 1912–1922.

Siddiqui, K. S. and Cavicchioli, R. (2006). Cold-Adapted Enzymes. *Annual Review of Biochemistry* 75.1, pp. 403–433. arXiv: arXiv:1011.1669v3.

Sieber, C. M., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., and Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* 3.7, pp. 836–843.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31.19, pp. 3210–3212.

Singh, H., Won, K., Ngo, H. T., Du, J., Kook, M., and Yi, T. H. (2015). Phycicoccus soli sp. nov., isolated from soil. *International Journal of Systematic and Evolutionary Microbiology* 65.8, pp. 2351–2356.

Singh, P., Singh, S. M., and Roy, U. (2016). Taxonomic characterization and the bio-potential of bacteria isolated from glacier ice cores in the High Arctic. *Journal of Basic Microbiology* 56.3, pp. 275–285.

Singh, S. P., Häder, D. P., and Sinha, R. P. (2010). Cyanobacteria and ultraviolet radiation (UVR) stress: Mitigation strategies. *Ageing Research Reviews* 9.2, pp. 79–90.

Singh, V. K., Mangalam, A. K., Dwivedi, S., and Naik, S. (1998). Primer Premier Program for Design of.pdf. 24.2, pp. 2–3.

Slaymaker, O. (2011). Criteria to distinguish between periglacial, proglacial and paraglacial environments. *Quaestiones Geographicae* 30.1, pp. 85–94.

Smith, D. J., Griffin, D. W., and Jaffe, D. A. (2011). The high life: Transport of microbes in the atmosphere. *Eos* 92.30, pp. 249–250.

Smith, H. J., Schmit, A., Foster, R., Littman, S., Kuypers, M. M., and Foreman, C. M. (2016). Biofilms on glacial surfaces: Hotspots for biological activity. *npj Biofilms and Microbiomes* 2, p. 16008.

Sohn, J. I. and Nam, J. W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics* 19.1, pp. 23–40.

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., Frey, J. E., and Ahrens, C. H. (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiology* 19.1, pp. 1–18.

Song, Y. L., Kato, N., Liu, C. X., Matsumiya, Y., Kato, H., and Watanabe, K. (2000). Rapid identification of 11 human intestinal Lactobacillus species by multiplex PCR assays using group- and species-specific primers derived from the 16S-23S rRNA intergenic spacer region and its flanking 23S rRNA. *FEMS Microbiology Letters* 187.2, pp. 167–173.

Sorek, R. and Cossart, P. (2010). Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics* 11.1, pp. 9–16.

Sriswasdi, S., Yang, C. C., and Iwasaki, W. (2017). Generalist species drive microbial dispersion and evolution. *Nature Communications* 8.1.

Stein, L. Y. and Klotz, M. G. (2016). The nitrogen cycle. *Current Biology* 26.3, R94–R98.

Steven, B., Léveillé, R., Pollard, W. H., and Whyte, L. G. (2006). Microbial ecology and biodiversity in permafrost. *Extremophiles* 10.4, pp. 259–267.

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology* 37.8, pp. 953–961.

Stibal, M., Hasan, F., Wadham, J. L., Sharp, M. J., and Anesio, A. M. (2012a). Prokaryotic diversity in sediments beneath two polar glaciers with contrasting organic carbon substrates. *Extremophiles* 16.2, pp. 255–265.

Stibal, M., Šabacká, M., and 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, pp. 644–654.

Stibal, M., Šabacká, M., and Žárský, J. (2012b). Biological processes on glacier and ice sheet surfaces. *Nature Geoscience* 5.11, pp. 771–774.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., and Nishijima, M. (2014). Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS ONE* 9.8.

Tamaki, H., Tanaka, Y., Matsuzawa, H., Muramatsu, M., Meng, X. Y., Hanada, S., Mori, K., and Kamagata, Y. (2011). Armatimonas rosea gen. nov., sp. nov., of a novel bacterial phylum, Armatimonadetes phyl. nov., formally called the candidate phylum OP10. *International Journal of Systematic and Evolutionary Microbiology* 61.6, pp. 1442–1447.

Tamames, J., Cobo-Simón, M., and Puente-Sánchez, F. (2019). Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics* 20.1, pp. 1–16.

Tanaka, T., Yamaguchi, H., Kato, H., Nishioka, T., Katsube, Y., and Oda, J. (1993). Flexibility Impaired by Mutations Revealed the Multifunctional Roles of the Loop in Glutathione Synthetase. *Biochemistry* 32.46, pp. 12398–12404.

Tang, Y., Horikoshi, M., and Li, W. (2016). Ggfortify: Unified interface to visualize statistical results of popular r packages. *R Journal* 8.2.

El-Tarabily, K. A., Nassar, A. H., and Sivasithamparam, K. (2008). Promotion of growth of bean (Phaseolus vulgaris L.) in a calcareous soil by a phosphate-solubilizing, rhizosphere-competent isolate of Micromonospora endolithica. *Applied Soil Ecology* 39.2, pp. 161–171.

Taylor, J. A., Sichel, S. R., and Salama, N. R. (2019). Bent Bacteria: A Comparison of Cell Shape Mechanisms in Proteobacteria. *Annual Review of Microbiology* 73.1, pp. 457–480.

Tedstone, A., Cook, J., Williamson, C., Hofer, S., McCutcheon, J., Irvine-Fynn, T., Gribbin, T., and Tranter, M. (2019). Algal growth and weathering crust structure drive variability in Greenland Ice Sheet ice albedo. *The Cryosphere Discussions*, pp. 1–24.

Telling, J., Boyd, E. S., Bone, N., Jones, E. L., Tranter, M., Macfarlane, J. W., Martin, P. G., Wadham, J. L., Lamarche-Gagnon, G., Skidmore, M. L., Hamilton, T. L., Hill, E., Jackson, M., and Hodgson, D. A. (2015). Rock comminution as a source of hydrogen for subglacial ecosystems. *Nature Geoscience* 8.11, pp. 851–855.

Thakur, M. P., Reich, P. B., Hobbie, S. E., Stefanski, A., Rich, R., Rice, K. E., Eddy, W. C., and Eisenhauer, N. (2018). Reduced feeding activity of soil detritivores under warmer and drier conditions. *Nature Climate Change* 8.1, pp. 75–78.

Thangaraj, B., Rajasekar, D. P., Vijayaraghavan, R., Garlapati, D., Devanesan, A. A., Lakshmanan, U., and Dharmar, P. (2017). Cytomorphological and nitrogen metabolic enzyme analysis of psychrophilic and mesophilic Nostoc sp.: a comparative outlook. *3 Biotech* 7.2, pp. 1–10.

Thawai, C., Kittiwongwattana, C., Thanaboripat, D., Laosinwattana, C., Koohakan, P., and Parinthawong, N. (2016). Micromonospora soli sp. nov., isolated from rice rhizosphere soil. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* 109.3, pp. 449–456.

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2.1, p. 3. arXiv: arXiv:1312.0570v2.

Thompson, L. R. et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551.7681, pp. 457–463.

Thursby, E. and Juge, N. (2017). Introduction to the human gut microbiota. *Biochemical Journal* 474.11, pp. 1823–1836.

Tomei, M. C. and Daugulis, A. J. (2013). Ex situ bioremediation of contaminated soils: An overview of conventional and innovative technologies. *Critical Reviews in Environmental Science and Technology* 43.20, pp. 2107–2139.

Torsvik, V., Sørheim, R., and Goksøyr, J. (1996). Total bacterial diversity in soil and sediment communities - A review. *Journal of Industrial Microbiology and Biotechnology* 17.3-4, pp. 170–178.

Tranter, M., Fountain, A. G., Fritsen, C. H., Lyons, W. B., Priscu, J. C., Statham, P. J., and Welch, K. A. (2004). Extreme hydrochemical conditions in natural microcosms entombed within Antarctic ice. *Hydrological Processes* 18.2, pp. 379–387.

Triadó-Margarit, X. and Casamayor, E. O. (2015). High protists diversity in the plankton of sulfurous lakes and lagoons examined by 18s rRNA gene sequence analyses. *Environmental Microbiology Reports* 7.6, pp. 908–917.

Tuorto, S. J., Darias, P., McGuinness, L. R., Panikov, N., Zhang, T., Häggblom, M. M., and Kerkhof, L. J. (2014). Bacterial genome replication at subzero temperatures in permafrost. *ISME Journal* 8.1, pp. 139–149.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, and JF, B. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428.6978, pp. 37–43.

Uetake, J., Naganuma, T., Hebsgaard, M. B., Kanda, H., and Kohshima, S. (2010). Communities of algae and cyanobacteria on glaciers in west Greenland. *Polar Science* 4.1, pp. 71–80.

Uetake, J., Nagatsuka, N., Onuma, Y., Takeuchi, N., Motoyama, H., and Aoki, T. (2019). Bacterial community changes with granule size in cryoconite and their susceptibility to exogenous nutrients on NW Greenland glaciers. *FEMS microbiology ecology* 95.7, pp. 1–8.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research* 40.15, pp. 1–12.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J. A. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research* 35.SUPPL.2, pp. 71–74.

Uroz, S., Calvaruso, C., Turpault, M. P., and Frey-Klett, P. (2009). Mineral weathering by bacteria: ecology, actors and mechanisms. *Trends in Microbiology* 17.8, pp. 378–387.

Uroz, S., Kelly, L. C., Turpault, M. P., Lepleux, C., and Frey-Klett, P. (2015). The Mineralosphere Concept: Mineralogical Control of the Distribution and Function of Mineral-associated Bacterial Communities. *Trends in Microbiology* 23.12, pp. 751–762.

Van Tatenhove, F. G. and Olesen, O. B. (1994). Ground temperature and related permafrost characteristics in west greenland. *Permafrost and Periglacial Processes* 5.4, pp. 199–215.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* 27.5, pp. 737–746.

Vigneron, A., Lovejoy, C., Cruaud, P., Kalenitchenko, D., Culley, A., and Vincent, W. F. (2019). Contrasting winter versus summer microbial communities and metabolic functions in a permafrost thaw lake. *Frontiers in Microbiology* 10.JULY, pp. 1–13.

Vincent, W. F. (2010). Microbial ecosystem responses to rapid climate change in the Arctic. *ISME Journal* 4.9, pp. 1089–1091.

Vishnivetskaya, T. A. and Kathariou, S. (2005). Putative transposases conserved in Exiguobacterium isolates from ancient Siberian permafrost and from contemporary surface habitats. *Applied and Environmental Microbiology* 71.11, pp. 6954–6962.

Vollmers, J., Wiegand, S., and Kaster, A. K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! *PLoS ONE* 12.1, pp. 1–31.

Von Ahsen, N., Wittwer, C. T., and Schütz, E. (2001). Oligonucleotide melting temperatures under PCR conditions: Nearest-neighbor corrections for MG2+, deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clinical Chemistry* 47.11, pp. 1956–1961.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9.11.

Wang, S., Liu, Y., Liu, G., Huang, Y., and Zhou, Y. (2017). A New Primer to Amplify pmoA Gene From NC10 Bacteria in the Sediments of Dongchang Lake and Dongping Lake. *Current Microbiology* 74.8, pp. 908–914.

Wang, Y., Dungait, J. A., Xing, K., Green, S. M., Hartley, I., Tu, C., Quine, T. A., Tian, J., and Kuzyakov, Y. (2020). Persistence of soil microbial function at the rock-soil interface in degraded karst topsoils. *Land Degradation and Development* 31.2, pp. 251–265.

Wang, Y. L., Wang, Q., Yuan, R., Sheng, X. F., and He, L. Y. (2019). Isolation and characterization of mineral-dissolving bacteria from different levels of altered mica schist surfaces and the adjacent soil. *World Journal of Microbiology and Biotechnology* 35.1, pp. 1–13.

Ward, B. A. (2019). Mixotroph ecology: More than the sum of its parts. *Proceedings of the National Academy of Sciences of the United States of America* 116.13, pp. 5846–5848.

Ward, C. P., Nalven, S. G., Crump, B. C., Kling, G. W., and Cory, R. M. (2017). Photochemical alteration of organic carbon draining permafrost soils shifts microbial metabolic pathways and stimulates respiration. *Nature Communications* 8.1, pp. 1–7.

Warnes, G. R. (2012). gplots: Various R programming tools for plotting data.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2019). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1.1. *http://CRAN.R-project.org/package=gplots.*

Watts, R. D. and England, A. W. (1976). Radio-echo Sounding of Temperate Glaciers: Ice Properties and Sounder Design Criteria. *Journal of Glaciology* 17.75, pp. 39–48.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Müller, R., Wohlleben, W., Breitling, R., Takano, E., and Medema, M. H. (2015). AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research* 43.W1, W237–W243.

Wei, S., Cui, H., Zhu, Y., Lu, Z., Pang, S., Zhang, S., Dong, H., and Su, X. (2018). Shifts of methanogenic communities in response to permafrost thaw results in rising methane emissions and soil property changes. *Extremophiles* 22.3, pp. 447–459.

Welch, S. A., Barker, W. W., and Banfield, J. F. (1999). Microbial extracellular polysaccharides and plagioclase dissolution. *Geochimica et Cosmochimica Acta* 63.9, pp. 1405–1419.

Weng, Y. Z., Chang, D. T., Huang, Y. F., and Lin, C. W. (2011). A study on the flexibility of enzyme active sites. *BMC Bioinformatics* 12.SUPPL. 1, S32.

Wernegreen, J. J. (2012). Endosymbiosis. *Current Biology* 22.14, pp. 555–561.

Whiteman, G., Hope, C., and Wadhams, P. (2013). Climate science: Vast costs of Arctic change.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* 95.12, pp. 6578–6583.

Wickham, H. and Henry, L. (2019). tidyr: Tidy Messy Data. *R package version 1.0.0.*

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software.*

– (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software.*

– (2016). ggplot2 Elegant Graphics for Data Analysis.

Wickham, H., Henry, L., and RStudio (2017). R: Package ‘tidyr’.

Wilhelm, L., Singer, G. A., Fasching, C., Battin, T. J., and Besemer, K. (2013). Microbial biodiversity in glacier-fed streams. *ISME Journal* 7.8, pp. 1651–1660.

Wilkinson, D. M., Koumoutsaris, S., Mitchell, E. A., and Bey, I. (2012). Modelling the effect of size on the aerial dispersal of microorganisms. *Journal of Biogeography* 39.1, pp. 89–97.

Williams, T. J., Wilkins, D., Long, E., Evans, F., Demaere, M. Z., Raftery, M. J., and Cavicchioli, R. (2013). The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environmental Microbiology* 15.5, pp. 1302–1317.

Williamson, C. J., Cameron, K. A., Cook, J. M., Zarsky, J. D., Stibal, M., and Edwards, A. (2019). Glacier algae: A dark past and a darker future. *Frontiers in Microbiology* 10.APR.

Wongfun, N., Plötze, M., Furrer, G., and Brandl, H. (2014). Weathering of Granite from the Damma Glacier Area: The Contribution of Cyanogenic Bacteria. *Geomicrobiology Journal* 31.2, pp. 93–100.

Woo, M. K., Kane, D. L., Carey, S. K., and Yang, D. (2008). Progress in permafrost hydrology in the new millennium. *Permafrost and Periglacial Processes* 19.2, pp. 237–254.

Woodhouse, J. N., Makower, A. K., Grossart, H. P., and Dittmann, E. (2017). Draft genome sequences of two uncultured Armatimonadetes associated with a Microcystis sp. (Cyanobacteria) Isolate. *Genome Announcements* 5.40, pp. 1–2.

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., and Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* 11.1, pp. 41–46.

Xi, J., Wei, M., and Tang, B. (2018). Differences in weathering pattern, stress resistance and community structure of culturable rock-weathering bacteria between altered rocks and soils. *RSC Advances* 8.26, pp. 14201–14211.

Xiao, R. and Zheng, Y. (2016). Overview of microalgal extracellular polymeric substances (EPS) and their applications. *Biotechnology Advances* 34.7, pp. 1225–1244.

Yallop, M. L., Anesio, A. M., Perkins, R. G., Cook, J., Telling, J., Fagan, D., MacFarlane, J., Stibal, M., Barker, G., Bellas, C., Hodson, A., Tranter, M., Wadham, J., and Roberts, N. W. (2012). Photophysiology and albedo-changing potential of the ice algal community on the surface of the Greenland ice sheet. *ISME Journal* 6.12, pp. 2302–2313.

Yang, J., Hassanpouryouzband, A., Tohidi, B., Chuvilin, E., Bukhanov, B., Istomin, V., and Cheremisin, A. (2019). Gas Hydrates in Permafrost: Distinctive Effect of Gas Hydrates and Ice on the Geomechanical Properties of Simulated Hydrate-Bearing Permafrost Sediments. *Journal of Geophysical Research: Solid Earth* 124, pp. 2551–2563.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). art%3A10.1186%2F1471-2105-13-134.

Yergeau, E., Bokhorst, S., Kang, S., Zhou, J., Greer, C. W., Aerts, R., and Kowalchuk, G. A. (2012). Shifts in soil microorganisms in response to warming are consistent across a range of Antarctic environments. *ISME Journal* 6.3, pp. 692–702.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F. O. (2014). The SILVA and ”all-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research* 42.D1, pp. 643–648.

Yoon, J. H., Lee, S. Y., Kang, S. J., and Oh, T. K. (2008). Phycicoccus dokdonensis sp. nov., isolated from soil. *International Journal of Systematic and Evolutionary Microbiology* 58.3, pp. 597–600.

You, I. and Kim, E. B. (2020). Genome-based species-specific primers for rapid identification of six species of Lactobacillus acidophilus group using multiplex PCR. *PLoS ONE* 15.3, pp. 1–9.

Yu, B. and Zhang, C. (2011). In silico PCR analysis. *Methods in Molecular Biology* 760, pp. 91–107.

Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnology and Bioengineering* 89.6, pp. 670–679.

Zamanzadeh, M., Hagen, L. H., Svensson, K., Linjordet, R., and Horn, S. J. (2016). Anaerobic digestion of food waste - Effect of recirculation and temperature on performance and microbiology. *Water Research* 96, pp. 246–254.

Zeng, Y. X., Yan, M., Yu, Y., Li, H. R., He, J. F., Sun, K., and Zhang, F. (2013). Diversity of bacteria in surface ice of Austre Lovénbreen glacier, Svalbard. *Archives of microbiology* 195.5, pp. 313–322.

Zhang, J. Y., Liu, X. Y., and Liu, S. J. (2011). Phycicoccus cremeus sp. nov., isolated from forest soil, and emended description of the genus Phycicoccus. *International Journal of Systematic and Evolutionary Microbiology* 61.1, pp. 71–75.

Zhang, S., Hou, S., Qin, X., Du, W., Liang, F., and Li, Z. (2015). Preliminary Study on Effects of Glacial Retreat on the Dominant Glacial Snow Bacteria in Laohugou Glacier No. 12. *Geomicrobiology Journal* 32.2, pp. 113–118.

Zhang, W., Zhang, F., Niu, Y., Li, Y. X., Jiang, Y., Bai, Y. N., Dai, K., and Zeng, R. J. (2020). Power to hydrogen-oxidizing bacteria: Effect of current density on bacterial activity and community spectra. *Journal of Cleaner Production* 263, p. 121596.

Zhang, X., Ma, X., Wang, N., and Yao, T. (2009). New subgroup of Bacteroidetes and diverse microorganisms in Tibetan plateau glacial ice provide a biological record of environmental conditions. *FEMS Microbiology Ecology* 67.1, pp. 21–29.

Zhao, C., Gupta, V. V., Degryse, F., and McLaughlin, M. J. (2017). Abundance and diversity of sulphur-oxidising bacteria and their role in oxidising elemental sulphur in cropping soils. *Biology and Fertility of Soils* 53.2, pp. 159–169.

Zhou, H. and Zhou, Y. (2004). Quantifying the Effect of Burial of Amino Acid Residues on Protein Stability. *Proteins: Structure, Function and Genetics* 54.2, pp. 315–322.

Zumsteg, A., Luster, J., Göransson, H., Smittenberg, R. H., Brunner, I., Bernasconi, S. M., Zeyer, J., and Frey, B. (2012). Bacterial, Archaeal and Fungal Succession in the Forefield of a Receding Glacier. *Microbial Ecology* 63.3, pp. 552–564.