DOCTOR OF PHILOSOPHY

Metagenomics to determine land use effects on soil ecosystem services

Jones, Briony

Award date:
2021

Awarding institution:
Bangor University

Metagenomics to determine land use effects on soil ecosystem services

Briony A Jones

Thesis submitted to Bangor University in

candidature for the degree Philosophiae Doctor

December (2020)

School of Natural Sciences, Bangor University,

Bangor, Gwynedd, LL57 2UW

PRIFYSGOL
BANGOR
UNIVERSITY

Canolfan Ecoleg
a Hydroleg y DU
UK Centre for
Ecology & Hydrology

Envision
Developing next generation
leaders in environmental science

Declaration

I hereby declare that this thesis is the results of my own investigations, except where otherwise stated. All other sources are acknowledged by bibliographic references. This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless, as agreed by the University, for approved dual awards.

Abstract

Agricultural intensification is exerting increasing pressure on our soils and there is growing recognition of the trade-offs in balancing human nutritional needs with longer term degradation of soil ecosystem services. It is well established that soil microorganisms play a significant role in driving numerous soil processes, including biogeochemical cycling, however we lack a detailed understanding of the biodiversity and functioning of these organisms in response to land use change. The development of molecular methodologies has provided new insights into microbial community change across different soils and land use systems, but advanced approaches are now needed to synthesise findings across studies and build a more predictive framework to specifically link land use change to change in microbial taxonomy and function. Initial work examined bacterial taxonomic responses to soil pH, as pH is both highly influential over soil microbial communities and heavily associated with land use practices such as fertilization and liming. I modelled pH responses of several thousand bacterial taxa from a large amplicon dataset consisting of > 1000 soils across Britain and developed a novel database/web application enabling querying of 16S sequences to obtain associated ecological information for novel soil taxa, in addition to taxonomy. Importantly this work also demonstrated that taxonomic shifts in soil bacterial communities can be predicted based solely on soil pH information.

In subsequent chapters I explored how such predictable changes in taxonomy link to functional change, through exploiting a large metagenomic dataset covering various land use intensification contrasts (grassland and arable) at a range of locations in the UK. Here I found that whilst both taxonomic and functional composition of microbial communities were largely driven by soil pH they also varied with land use intensity. The relationship between soil pH and specific microbial functions was explored through examining relative abundances of a key organic matter decomposition gene (β-glucosidase) within long term pH manipulated grassland plots. Here I found there were increased relative abundance of *Acidobacteria* β-glucosidase genes in more acidic soils alongside shifts in related glycoside hydrolase families in response to soil pH, demonstrating pH has not only an important influence on bacterial taxonomy, but also important soil functions relating to carbon cycling.

I then used novel genomic assembly and binning methods to demonstrate that numerous *Thaumarchaeota* (also known as *Thermoproteota*) bins were important discriminators of intensive arable soils based on random forest analyses. Within short read analyses I found that nitrate reductase subunits were important in distinguishing land use and were consistently in higher abundance within high intensity soils**.** Other denitrification genes were statistically significant indicators of arable soils specifically, including nitrite reductase and nitric oxide reductase genes. Significant grassland indicators included numerous nitrogen fixation genes. Both sulfur and phosphorus metabolism also demonstrated shifts in genes in response to land use with findings collectively indicating differential nutrient

acquisition strategies in grassland and arable soils, potentially due to a reliance on nutrients derived from fertilisers within arable soils. Coupling of short read functional indicators with indicators of taxon based on metagenomic bins enabled insights into how land use driven changes in microbial taxa relate to functional change at the community level; and further emphasised the power of a binning approach to link taxa to environmental responses and functions within a land use change context. Through using these novel bioinformatics approaches there is now the opportunity to enhance the database system I developed early in my research, to capture novel ecological information on whole bacterial genomes. This will further enable an improved predictive understanding of the resilience of soil microbial functioning both to land use and wider environmental change.

Acknowledgements

Thanks to my supervisor Rob Griffiths, it has been great to work with you and I very much appreciate the support you have given me, both in terms of my project and also my general training and development. Thank you to my co supervisor Davey Jones for your valuable input, encouragement and pep talks throughout.

A massive thank you to Tim Goodall, both for all your molecular work contributing to the datasets studied and for taking time out to let me shadow and work in the lab with you providing context to the reams of a, t, g, c's on my screen.

Thanks to Melanie Armbruster for being a lovely friend and former office mate. I enjoyed the occasions I joined you in the Gro-dome/lab and witnessed what the hardcore hands on PhD's do. Thanks also to the wider UKCEH molecular ecology group past/present/transient for some solid advice and memorable times.

Thanks to Andrew Everitt for the ace maintenance of the Huxley server where a large proportion of this work was conducted. Thank you also to the UKCEH student community both in Wallingford and Bangor, for being a lovely and inclusive bunch of people.

Thank you to Chris who's been there for me from Oxon to North wales, through a pandemic and beyond and for being incredibly supportive throughout. Thank you to my parents, my brother Dylan and lovely friends. Your support and faith in me got me to the point of starting a PhD in the first place and helped me reach the eventual finish line. I am lucky to have you.

Table of contents

List of Figures

List of Tables

Chapter 1

Literature review

1.1 Introduction

Soil provides numerous services vital for ecosystem functioning. Not only does soil provide the fundamental media to grow plants, but it also stores large amounts of global carbon, and contributes to the regulation of numerous biogeochemical cycles which can affect both air and water quality (Blum, 2005; Powlson *et al.*, 2011). Many of these services are orchestrated by microbial communities composed of bacteria and fungi (Gao *et al.*, 2015) living within microhabitats on the surface of plant roots, soil aggregates and inside aggregate pores (Foster, 1988). These organisms act as biogeochemical engines (Falkowski *et al*., 2008) driving the numerous soil processes, which together contribute to the overall functioning of soil. A gram of soil is commonly reported to contain up to 10^{10} bacterial cells (Raynaud and Nunan, 2014), though the large majority of bacterial taxa are uncharacterised. Our fundamental ecological knowledge of how soil microbes regulate soil ecosystem services is therefore limited, thus it's an immense challenge to predict how future environmental change could impact upon microbially mediated soil processes.
A significant factor exerting environmental pressure on soils is human land use. In order to provide both food and raw material to build human societies, we have drastically transformed landscapes and consequently the global terrestrial ecosystem. Human induced land use change can either involve broad habitat conversions such as forest clearance, or more subtle alterations in management such as different tillage, fertilizer and liming practices (Guo and Gifford, 2002). Land use effects on ecosystems are receiving increased attention due to the growing need to increase agricultural production, coupled with a wider desire to sustainably manage natural resources. Soils are receiving increased recognition as a natural asset in this regard; and since soil development takes time, it is now considered a resource which needs protecting and considering in policies surrounding protection of the natural environment (FAO and ITPS, 2018; Pennock, 2019).

Whilst the motivation for most land practices is to acquire natural resources of some kind, this often comes at the price of degradation of the environment (Foley *et al.*, 2005). In order for land to be utilized for agricultural purposes the land must be cleared of its natural vegetation which often results in a loss of soil organic matter and the release of CO_2 (West *et al.*, 2010; Vlek *et al.*, 2017). It has been estimated that 11% of Greenhouse gas (GHG)

emissions have been attributed to land use change whilst a further 13% has been attributed to agriculture (Vlek *et al.*, 2017). Whilst in their natural state these soils could store large volumes of carbon and play a key role in climate change mitigation (West *et al.*, 2010; Vlek *et al.*, 2017). The use of mineral fertilizers also has both positive and detrimental impacts on varying ecosystem services. For example, whilst mineral nitrogen has played a significant role in massively increasing post-wartime crop production, and meeting human nutritional needs, it also has numerous detrimental impacts on the environment (Galloway *et al.*, 2003, 2008; Zhang *et al.*, 2015) including pollution of ground water, eutrophication and nitrous oxide emissions (Zhang *et al.*, 2015). Given that land use practices can lead to degradation of soils and the environment there is an urgent need to adopt more sustainable agricultural practices, whilst also ensuring human nutritional needs are met. It has also been proposed that proactive policies should be implemented to actively encourage carbon sequestration in soils (Demenois *et al.*, 2020). Suggested methods to increase carbon sequestration in soils include agroforestry (Powlson *et al.*, 2011; Demenois *et al.*, 2020), reduced tillage (Govaerts *et al.*, 2009), intercropping with perennials and breeding crops with longer roots (Powlson *et al.*, 2011; Demenois *et al.*, 2020). Whilst careful management and monitoring of fertilizer has been suggested to ensure efficient use of nutrients and reduce the risk of leaching (Powlson *et al.*, 2011). Understanding the wider ES trade-offs (Bennett *et al.,* 2009) at stake when making land use decisions is therefore of fundamental importance, as an increase in one ES can often result in the decline of another. Whilst much is known about above-ground trade-offs, comparatively little is known about the trade-offs with respect to below-ground soil ecosystem services. This is due to a fundamental lack of understanding about the ecology of soil organisms and in particular a lack of understanding about how the vast diversity of soil microorganisms interacts with the environment to regulate soil functions and services. Previously, this lack of understanding has been due to the absence of appropriate methodologies to study both microbial diversity and function.

These issues have now lessened somewhat due to the development of molecular methodologies (summarised in **1.3-1.4)** that allow us to study both taxonomic and functional diversity, based on extraction of nucleic acids from soil. Numerous studies have used genomic approaches to study the taxonomic composition of microbial communities in response to land use and have observed taxonomic shifts in microbial communities in

response to varying land use treatments, both on a local scale (Hartmann *et al.*, 2015; Pershina *et al.*, 2015; Banerjee *et al.*, 2016; Francioli *et al.*, 2016; Schöps *et al.*, 2018; Huang *et al.*, 2019) and within globally distributed sites (Leff *et al.*, 2015). It has been reported that taxa associated with carbon decomposition are found in increased relative abundance within manure treated soils such as *Firmicutes*, *Proteobacteria* and *Zygomycota*, whilst more oligotrophic taxon such as *Acidobacteria* are found at sites without fertiliser treatment (Francioli *et al.*, 2016). Other work has reported that there was reduction in *Acidobacteria* at sites with nutrient additions coupled with an increase in copiotrophic organisms such as *Actinobacteria* and *Alphaproteobacteria* (Leff *et al.*, 2015). Whilst other findings have shown a reduction in the number of carbon (C), nitrogen (N) and phosphorus (P) cycling genes in response to N and P addition (Huang *et al.*, 2019). Novel metagenome assembly methods have also begun to be applied to study microbial responses to land use, with work finding that genes encoding nitrification enzymes were typically found in *Thaumarchaeota* and *Nitrospira* MAGs (Metagenome assembled Genomes) and that MAGs belonging to these taxa increased in abundance in response to N additions (Orellana *et al.*, 2018).

Now that we have the molecular tools (marker gene and metagenomics approaches) and computational methods (specifically related to genome assembly) there is a need for further research in this area specifically linking land use to taxonomy and function to meet two primary goals:

1. Provide new fundamental understanding of the ecology and functional potential of previously undiscovered microbes.

2. To determine how land management in interaction with natural environmental change affect the diversity and function of soil microbes across a range of scales.

These two challenges underpin the work presented in this thesis. In the following sections of this introductory chapter, I will expand on the concepts raised above, reviewing both the link with soil microbes and ecosystem services, as well as providing an up to date review of the methodologies which can now be used to assess the molecular basis of soil microbial function. Finally, I will introduce the specific aims of the thesis.

1.2 Microbial roles in soil function

1.2.1 Carbon Cycling

Soil is known to store three times as much carbon as is found in the atmosphere or within plants (Schmidt *et al.*, 2011). Microbial communities play a significant role in soil carbon cycling and storage and consequentially the regulation of GHG such as carbon dioxide (CO_2) and methane (CH_4) emissions (Schmidt *et al.*, 2011; Jansson and Hofmockel, 2020). Fungi and bacteria are important mediators of carbon decomposition (both above ground plant litter and below ground root litter and exudates) through secreting extracellular enzymes. The activity of these enzymes facilitate the release of vital nutrients, which can be utilized for bacterial and plant growth (whilst also resulting in the release of CO_2) (Gougoulias et al., 2014). Microbes therefore regulate C cycling through modulating rates of degradation of plant carbon inputs and also through converting carbon into more stable forms termed humus (Kallenbach *et al.*, 2016). Dead microbes are also an important constituent of hummus, alongside undegraded or biochemically transformed plant material. Microbes also contribute to the regulation of CH_4 emissions, which is considered the second most potent GHG after CO_2 and though it accounts for a much smaller proportion of GHG emissions in comparison to CO_2, it possesses 25 times the warming potential (Nazaries *et al.*, 2013). Microbes termed methanogens are capable of producing methane from methanogenic substrates such CO_2 , acetate, and methylated compounds (Nazaries *et al.*, 2013; Evans *et al.*, 2019). Previously methanogens were thought to be exclusively archaeal, however it has since been found a broader range of organisms are capable of methanogenesis including the bacterial phyla *Cyanobacteria* (Bižić et al., 2020). Methanotrophs (predominantly bacterial) are capable of consuming methane and therefore play a role in attenuating the CH_4 emissions from methanogenesis (Nazaries *et al.*, 2013).

1.2.2 Nitrogen cycling

Nitrogen transformations are another critical function of soil microorganisms and are of fundamental importance to both the global N cycle and plant growth. Nitrogen is an

essential nutrient used for amino acid and nucleic acid synthesis, but available nitrogen predominantly exists as dinitrogen gas (N_2) which cannot be used by plants.
Nitrogen fixing bacteria are capable of converting dinitrogen (N_2) to plant available forms such as ammonia (NH_3), using the enzyme nitrogenase (Zehr *et al.*, 2003; Kuypers *et al.,* 2018). Other microbes then orchestrate nitrification which has traditionally been considered a two-step process. The first step ammonia oxidation describes the oxidation of NH_3 or ammonium (NH_4^+) to nitrite (NO_2^-). Ammonia oxidation was previously thought to be solely conducted by bacteria (ammonia oxidising bacteria/ AOB), however we now know that this is also commonly conducted by archaea (ammonia oxidising archaea /AOA) (Martens-Habbena *et al.*, 2009; Lu *et al.,* 2016). In the second step of nitrification, nitrite oxidizing bacteria (NOB) oxidise NO_2^- to nitrate (NO_3^-) allowing nitrogen to be assimilated by plants (Han *et al.*, 2018). Recently an organism from the bacterial genus *Nitrospira* has been found to conduct both steps of nitrification in a process known as complete ammonia oxidation (commanox) (Daims *et al.*, 2015) .

Denitrification is also microbially orchestrated whereby NO_3^- is reduced to nitrite NO_2^- (Zumft, 1997), nitric oxide (NO) and nitrous oxide (N_2O) (two potent greenhouse gases) (Zumft, 1997) before being reduced to N_2 (Graf *et al.,* 2014). Microbes (predominantly *Planctomycetes*) also contribute to the anaerobic oxidation of NH_3 (anammox) whereby NH_3 is oxidised to N_2 with NO_2^- as an electron acceptor (Humbert *et al.*, 2010) without NO and N_2O as intermediates (van Niftrik and Jetten, 2012; Wang *et al.*, 2019).

1.2.3 Phosphorus cycling

Phosphorus is widely considered as the second most important nutrient for plant growth after nitrogen (Liang *et al.*, 2020) and contributes to many metabolic processes including photosynthesis, signal transduction, respiration and biosynthesis. Phosphorus is highly abundant in soils but largely exists in forms that are inaccessible to plant roots and therefore is also a limiting factor to plant growth (Sharma *et al.*, 2013). Soil phosphorus exists in both organic and inorganic forms, inorganic phosphorus predominantly exists in the form of insoluble mineral complexes (Rodríguez *et al.*, 2007), while organic phosphorus is either incorporated within biomass or associated with soil organic matter (Richardson and

Simpson, 2011). Microbes play an important role in making phosphorus available to plants by converting both organic and inorganic phosphorus into soluble forms such as orthophosphate (Rodríguez *et al.*, 2007). Microbes mineralise organic phosphorus through the release of extracellular enzymes (Sharma *et al.*, 2013) such as phosphatases (Hayat *et al.*, 2010; Sharma *et al.*, 2013). Phosphatases are secreted by bacteria such as *Bacillus* and *Pseudomonas*, as well as fungi including *Aspergillus*, *Pencillum*, *Mucor*, and mycorrhizal hyphae (Shrivastava *et al.,* 2018). Microbes solubilise inorganic phosphorus in a process linked to the release of organic acids, such as gluconic acid (Rodríguez *et al.*, 2007; Liang *et al.*, 2020). Microbial solubilization of inorganic phosphorus has been associated with bacteria (including *Actinomycetes*, *Pseudomonas* and *Bacillus* spp.) and fungi (including *Aspergillus* and *Penicillum* spp.) (Richardson and Simpson, 2011). Mycorrhizae also play a role in increasing the uptake of phosphorus by plants, by associating with the roots in the rhizosphere which effectively extends plant root systems (Browne *et al.*, 2009; Richardson and Simpson, 2011; Alori *et al*., 2017).

1.2.4 Sulfur cycling

Sulfur is another essential element for plant growth and is present in amino acids such as cysteine (Kertesz and Mirleau, 2004; Gahan and Schmalenberger, 2014) and methionine (Gahan and Schmalenberger, 2014) and is also required for the synthesis of coenzyme A (Eriksen, 2009; Lucheta and Lambais, 2012). Sulfur is also present in numerous redox and electron-transfer proteins. However ~95% of soil sulfur is found in organic forms (predominantly sulfate esters and C-bonded sulfur (Scherer, 2009)), which cannot be utilised by plants (Kertesz and Mirleau, 2004; Gahan and Schmalenberger, 2014). Microbes help mediate the mineralisation and immobilisation of sulfur as well as numerous oxidation and reduction reactions (Lucheta and Lambais, 2012). A significant contribution of both bacteria and fungi to sulfur cycling is the secretion of sulfatases which hydrolyse sulfate esters (which make up ~30-70% of organic sulfur in soils) (Klose *et al.*, 2015) to inorganic sulfates (SO_4^{2-}) (the predominant plant available form).

1.2.5 Broader role of soil microbes of relevance for soil ecosystem services

In addition to their significant role in a number of biogeochemical cycles microbial communities also provide other broader ecosystem services. Microbial biodiversity can itself be considered a highly valuable soil function, providing a wealth of novel gene products with a wide range of applications. These natural products are now more easily accessible due to the development of culture-independent genomic approaches(Handelsman *et al.*, 1998; Lee and Lee, 2013; Katz *et al.,* 2016). For example, novel microbial enzymes derived from soils have been utilized in a range of sectors including pharmacy, food, the production of detergents, textiles, leather and pulp and paper (Demain and Adrio, 2008; Berini *et al.*, 2017; Castillo Villamizar *et al.*, 2019). Microbes with specific enzymatic capacities can also be utilised for the purpose of bioremediation, whereby microbes (typically originating from polluted environments) are used to degrade toxic pollutants into less toxic/ non-toxic forms (Techtmann and Hazen, 2016). Soil microbial communities are also a major source of antibiotics; it has previously been reported that 80% of antibiotics in clinical use are derived from soil bacteria (D'Costa et al., 2007; Joseph et al., 2009; Woappi, 2013).Thus, some soil microbes are naturally resistant to a range of antibiotics (van Elsas *et al.*, 2008), therefore soils are also a reservoir of antibiotic resistance genes (Torres-Cortés *et al.*, 2011) and can provide insights into novel mechanisms of antibiotic resistance which may emerge within a clinical setting (Martínez, 2008; Torres-Cortés *et al.*, 2011).

Microbes also contribute to wider ecosystem functioning through their interactions with other organisms. Plant growth promoting rhizobacteria (PGPRs) aid plant growth through mechanisms such as the release of siderophores inhibiting phythopathogen growth (Schroth and Hancock, 1982) as well as the release of phytohormones including cytokines, auxins, abscisic acid and gibberellic acid (Ortíz-Castro *et al.*, 2009; Fahad *et al.*, 2015). Other microbes act as plant pathogens , whereby they colonize the plant surface or shoots and secrete "effectors" including degradative enzymes or toxins (Kannan, Bastas and Devi, 2015; Martins *et al.*, 2018). Microbes also contribute to soil functioning through their role in the wider food web. Indeed higher level microorganisms such as protozoa and nematodes predate bacteria and fungi (Griffiths, 1994; Raynaud, Lata and Leadley, 2006) resulting in the release of mineral N into the soil which can be up taken by plants. Viruses specifically

phages (viruses which infect bacteria) are also contributors to microbial mortality and have been hypothesised to also release mineral N from microbial biomass through infecting and lysing microbial cells (Emerson, 2019).

1.3 Methods for studying microbes in soils

Early analyses of soil microorganisms involved characterising microbes in terms of morphology using light microscopy and staining methods. Pure culture methods were later developed to isolate specific microbes of interest on media such as potato slices, gelatine and agar (Stackebrandt, 2006; Escobar-Zepeda, De León and Sanchez-Flores, 2015). Microbes were later cultured on nutrient enriched media to gain insight into their physiologies (Stackebrandt, 2006). Through using these methods it became apparent there was a substantial difference in the number of cells growing/dividing in culture and the number observed using direct microscopy, a phenomenon referred to as "the great plate anomaly" (Staley and Konopka, 1985). Improved insights into these unculturable organisms came with the advent of fatty acid methods such as phospholipid fatty acids (PFLA), a culture independent method enabling broad insights into taxonomic composition of microbial communities (Nannipieri *et al.*, 2003; Liu *et al.*, 2006). Since fatty acids constitute a relatively constant proportion of cell membranes, the basic principle of this approach is to measure changes in phospholipids in order to infer shifts in microbial biomass and detect broad changes in taxonomy (Zelles, 1999) with specific fatty acid signatures being indicative of the presence of broad taxonomic groupings (Liu *et al.*, 2006).

Later, Carl Woese proposal to use universally conserved rRNA genes to infer phylogeny and the development of Sanger sequencing enabled us to gain insights into unculturable microbes using genomic approaches (Escobar-Zepeda *et al.,* 2015). Early genomic methods used to study these microbes included polymerase chain reaction (PCR) and rRNA 'clone sequencing' (Fromin *et al.*, 2002; Liu *et al.*, 2006; Escobar-Zepeda *et al.,* 2015). While cloning PCR products was able to yield phylogenetic information, it was laborious and expensive and therefore not suitable for studying multiple samples. This lead to the development of fast DNA fingerprinting methods (temperature gradient gel electrophoresis/TGGE, denaturing gradient gel electrophoresis/DGGE and terminal-restriction fragment length/TRFLP)

whereby PCR amplified DNA fragments are separated by electrophoresis to provide broad insights into microbial composition(Forbes *et al.*, 1991; Fromin *et al.*, 2002; Sabale *et al*., 2020).

More recently, high-throughput sequencing and subsequently modern marker gene analyses have provided a sensitive and efficient method to detect taxa present in an environment (Bharti and Grimm, 2019), revolutionising our knowledge of microbial communities (Caporaso *et al.*, 2011). These methods have uncovered a wealth of previously undetectable microbial diversity (Buée *et al.*, 2009; Orgiazzi *et al.*, 2015) and enabled new insights into taxonomic diversity and composition within varying soil environments. Within these methods, specific marker genes are amplified by PCR and then sequenced (Orgiazzi *et al.*, 2015). Marker genes need to be conserved enough to be able to profile taxonomic communities through a common primer binding site, while being simultaneously variable enough to distinguish between taxa (Bharti and Grimm, 2019). Typically, rRNA genes are used for these purposes as they are both highly conserved whilst also containing hypervariable regions. Most commonly, ITS (internal transcribed spacer) is used to taxonomically profile fungi, whilst 16S and 18S genes (small subunit rRNA) are used to profile bacteria and eukaryotes respectively (Lindahl *et al.*, 2013).

Despite the wealth of knowledge we have derived from these approaches regarding biodiversity and taxonomic responses, these methods do not provide insight into how taxonomic diversity relates to microbial functioning (van Elsas and Boersma, 2011). Methods such as qPCR can be used to amplify and quantify particular functional genes of interest, however other approaches are needed when wishing to understand collective functioning of microbial communities. Microarrays have previously been used to gain broader functional insights into microbial communities. Microarrays are composed of a matrix of immobilized DNA fragments, sat upon a substrate, with each fragment (referred to as a probe) measuring the expression of a specific gene (Plomin and Schalkwyk, 2007). Microarrays are referred to as a “closed” system technology, meaning the range of experimental results that are possible to attain (i.e. the selection of probes bound to the substrate) is defined prior to analyses. Such formats, whilst useful, do not allow for novel gene discovery and in the context of soils where the majority of bacterial taxa have yet to be

cultured, an “open” system (which doesn’t require prior knowledge of a community) is arguably more useful. Indeed as NGS technologies have continued to develop there has been a shift away from the use of microarrays to study community functioning (Ledford, 2008; Roh *et al.*, 2010) and a move toward metagenomics approaches (van Elsas and Boersma, 2011).

1.4 Metagenomics data production and analysis

Metagenomics is a methodology whereby DNA is sequenced directly from the sample, enabling us to study the “collective genome” of a microbial community, with the potential to gain genomic and functional information of unculturable microorganisms (Handelsman *et al.*, 1998; Daniel, 2005; Wooley and Ye, 2010). Applying metagenomic methods to soils has the potential to improve understanding of the genetic regulation of vital soil microbial processes (described in section **1.2**). Whilst also enabling better insights into how land management affects soil function (section **1.1**) and indeed a better understanding of the relationship between function and microbial (taxonomic) diversity. Such understanding is critical not only to enhance wider scientific understanding, but also for environmental monitoring purposes with respect to policy implementation, where there is a clear need for biotic indicators of the enhancement or impairment of soil functionality caused by land management change and indeed climate change. The following sections will focus on the key stages involved in soil metagenomics (summarised in **Fig.1.1**), whilst also highlighting the challenges of implementing these steps in soils specifically.

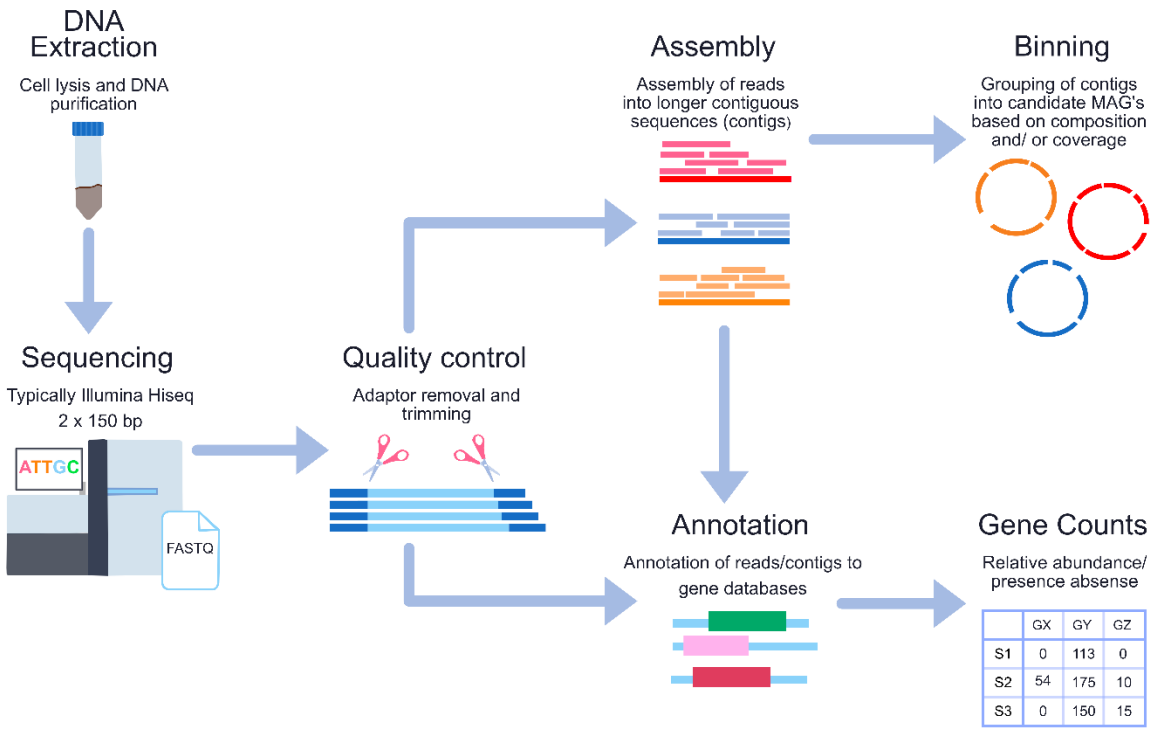


Fig.1.1. Summary of commonly implemented metagenomics methods.

1.4.1 DNA extraction

The first step in metagenomics analyses is DNA extraction, which is a highly critical process as large amounts of high quality DNA are required to ensure successful library preparation and sufficient representation of the microbial community in the sample (Thomas *et al.,* 2012). DNA extraction from soils is a particular challenge, largely due to humic acids (Steffan *et al.*, 1988; Tebbe and Vahjen, 1993) which are often co-extracted with DNA as they share similar physiochemical properties to nucleic acids (Lakay *et al.,* 2007) meaning suitable purification methods are required (Harry *et al.*, 1999; Robe *et al.*, 2003; Lakay *et al.,* 2007). All DNA extraction methods involve some form of cell lysis, which aims to disrupt the cell wall to release DNA. There are three methods of cell lysis, which can be used alone, but more commonly are used in combination: physical (e.g. freeze-thawing, freeze-boiling, bead beating), chemical (e.g. use of detergents) and enzymatic (e.g. lysozyme treatment) (Robe *et al.*, 2003). DNA extraction methods are broadly described as either being 'direct' where cells lysis is conducted within the soil or 'indirect' where microbial cells are separated from the soil particles prior to cell lysis (Robe *et al.*, 2003). Direct extraction methods are typically

characterised by higher yields and shorter fragment length, whilst indirect methods typically have lower yields, longer fragment size and less humic contaminants (Steffan *et al.*, 1988; Leff *et al.*, 1995; Lombard *et al.*, 2011). Direct extraction methods are most commonly used (Delmont *et al.*, 2011; Lombard *et al.*, 2011; Xie *et al.*, 2018) due to higher DNA yields. Different extraction methods have also been found to impact upon soil microbial composition (Wüst *et al.*, 2016; Zielińska *et al.*, 2017) and diversity (Gupta *et al.*, 2017). This is likely to be related to the fact that soil microbes vary in susceptibility to cell lysis depending on the method used (Daniel, 2005).

1.4.2 Sequencing technologies

For the past 40 years, sequencing methods have been developing and advancing, whilst not all methods are appropriate for modern day metagenomic analyses (specifically first generation sequencing), all commonly used sequencing methods (first generation through to third generation) shall be briefly described in the following section.

Sanger sequencing is regarded as the 'first generation' of sequencing, and since it emerged in 1977, it has played a vital role in early bacterial community studies (Escobar-Zepeda *et al.*, 2015). Sanger sequencing uses a chain termination method, initially normal elongation occurs whereby a polymerase is used to incorporate dideoxy-nucleotides (dNTP's) into a template strand. Modified dideoxy-nucleotides (ddNTP's) are then added which lack a hydroxyl group on the 3' end which is required for extension by polymerase, thus when ddNTP's are incorporated into the template strand elongation is terminated. Traditional 'manual' Sanger sequencing was conducted using four separate reactions each containing a different ddNTP (ddATP, ddGTP, ddCTP, or ddTTP), resulting in fragments of various lengths derived from each of the modified bases reactions. These fragments were then visualised on four lanes of a polyacrylamide gel, enabling the sequence to be inferred (Sanger *et al.,* 1977; Kchouk *et al.,* 2017). Since the 1990's Sanger sequencing has been automated using fluorescently dye labelled ddNTPs and capillary based electrophoresis (Swerdlow and Gesteland, 1990, Heather and Chain, 2016).

Though Sanger sequencing dominated for three decades, the time and costs required were a significant drawback (Kchouk *et al.,* 2017). In the past 10-20 years there has been a massive shift away from traditional Sanger sequencing towards Next Generation Sequencing (NGS) technologies. NGS platforms are able to sequence millions of reads in parallel distinguishing them from Sanger sequencing technologies as much more efficient technologies (Escobar-Zepeda *et al.,* 2015) and also as a cheaper sequencing method (Kchouk e*t al.,* 2017). The introduction of the first widely used NGS sequencing technology Roche 454 was a huge leap forward in sequencing throughput (**Table.1.1**) (Heather and Chain, 2016) and reduced the need for laborious fragment cloning used in Sanger sequencing (Kulski, 2016). In 454 sequencing, DNA is first randomly fragmented (Hall, 2007) and then attached to beads using adaptor sequences. These beads are then amplified by a water in oil emulsion PCR (emPCR), generating approximately a million copies of each fragment. These beads are then washed over a PicoTiterPlates, containing a huge number of wells enabling thousands of pyrosequencing reactions to be conducted simultaneously. The base incorporated is determined by light emissions which are captured by a sensor beneath the plates wells (Heather and Chain, 2016; Kchouk *et al.,* 2017) .It is of note that 454 sequencers are no longer supported and Roche no longer supplies 454 sequencers or reagents (Kulski, 2016). An alternate second generation sequencing technology, Ion torrent is conceptually similar to Roche 454 (Reuter et al., 2015), whereby DNA fragments are replicated using emPCR before being washed over a picowell plate. However, unlike Roche 454 nucleotides are determined based upon a change in pH in the solution opposed to luminescence. This change in pH is caused by a proton release which is detected by a sensor (Rothberg *et al.*, 2011; Kchouk, et al., 2017).

Today short read sequencing is largely dominated by Illumina sequencing. In Illumina sequencing, DNA is fragmented before adaptors are ligated to each end (Kchouk *et al.,* 2017). DNA is then replicated using PCR bridge amplification where DNA is washed over a flow cell of complementary oligonucleotides (Heather and Chain, 2016) creating “clusters” of sequences containing ~1000 copies (Kchouk *et al.,* 2017). The sequence is then established using fluorescent reversible terminator dNTP, which have a fluorophore in the 3’ region which prevents further nucleotides binding. The fluorophores are excited by lasers,

creating light signals that are detected by a coupled charge device camera (CCD). The fluorophore is then cleaved in order to enable sequencing to continue (Reuter *et al.,* 2015).

Whilst second generation sequencers have undoubtedly revolutionised sequencing, their short read length (Pollard *et al.*, 2018; van Dijk *et al.*, 2018) make both annotation and assembly challenging. Since the 2010's, "third generation" sequencing technologies have emerged (Giani *et al.*, 2020), characterised by longer read lengths, reduced sequencing costs (Kulski, 2016), and no explicit need for DNA amplification (Kchouk *et al.,* 2017). These technologies however, have significantly higher error rates as shown in **Table.1.1** (Kchouk et al., 2017). The first widely used third generation sequencing technology was the single molecule real time (SMRT) technology from PacBio (Pollard *et al.*, 2018). SMRT conducts real time sequencing, whereby signals are emitted as incorporations occur (Rhoads and Au, 2015; Kchouk *et al.,* 2017). During library preparation, hairpin adaptors are ligated to template DNA creating a structure known as a SMRT bell. SMRT bells are then loaded into a chip termed a SMRT cell, made up of hundreds of thousands of zero-mode waveguide nanowell arrays (ZMW) containing immobilised DNA polymerases. These polymerases are able to bind to the hairpin adaptors and start replication. Fluorescently labelled nucleotides are then incorporated and generate distinct light emissions, which are captured by a camera. This movie of light pulses, enables the sequence to be inferred (Rhoads and Au, 2015; Giani *et al.*, 2020). Nanopore is another widely used third generation sequencer, which utilizes numerous protein pores, these protein pores sit within a lipid bilayer (Wang *et al.,* 2014*)* which separates two chambers containing aqueous electrolytes (Branton *et al.*, 2008). A voltage is applied across the membrane, DNA then enters pores one base at a time, which temporarily obstructs the lumen and alters the current. The different bases result in different levels of flow disruption, enabling the base sequence to be inferred (Dear, 2003; Wang *et al.,* 2014).

	Sequencing technology	~ No of reads per run	~ Read length	~ Error rate	References
First gen	**Sanger**	6144 (16 x 384)	600 – 1000 bp	0.3%	(Wang *et al.*, 2012; Escobar-Zepeda, De León and Sanchez-Flores, 2015; Heather and Chain, 2016)
Second gen	**Roche 454 (454 GS FLX+)**	1 x 10^6	700 bp	1%	(Kulski, 2016)
	Ion torrent	5 x 10^6 (Ion PGM) 6 x 10^7 (Ion Proton)	200 bp	1.71 %	(Quail *et al.*, 2012; Kulski, 2016)
	Illumina	5 x 10^9 (Hiseq) 3 x 10^8 (Miseq)	2 x 150 bp (HiSeq) 2 x 300 bp (Miseq)	0.26% (Hiseq) 0.8% (Miseq)	(Quail *et al.*, 2012; Kulski, 2016)
Third gen	**PacBio**	1 × 10^6	On average 10–15 kbp (max read length > 80 kbp)	1-14%	(Kulski, 2016; Pollard *et al.*, 2018; van Dijk *et al.*, 2018)
	Nanopore	6 x 10^4	>20 kbp	5-20 %	(Kulski, 2016; Rang, Kloosterman and de Ridder, 2018; Loit *et al.*, 2019)

Table.1.1. Summary of sequencing technologies statistics.

1.4.3 Quality Control

The first stage in metagenomics data pre-processing is quality control. This is necessary to remove sequencing artefacts including sequences of low quality, contaminating reads (Zhou *et al.*, 2014) and adaptor sequences (Bolger *et al.,* 2014). Sequencing platforms themselves can introduce biases based on their base calling methods (Escobar-Zepeda *et al*., 2015) and sequence properties such as GC rich or poor regions can make the sequencing process more error prone (particularly in the case of Illumina sequencing) (Benjamini and Speed, 2012; Chen *et al.*, 2013; Ladoukakis *et al.*, 2014). Sequence quality is quantified by phred scores (recorded in FASTQ files (Cock et al., 2009)) which denote the probability that an incorrect base has been incorporated with higher phred scores denoting lower error probabilities and lower scores signifying higher error probabilities (Ewing and Green, 1998). Typical quality control methods involve the removal of adaptor/partial adaptor sequences (Martin, 2011; Bolger *et al.*, 2014) prior to sequence trimming. Sequence trimming is often informed by phred scores, with many trimming tools employing a "sliding window" approach (Joshi et al., 2011; Bolger *et al.,* 2014). In this approach a window signifies a specific length of consecutive bases, this 'window' slides along the sequence, one base at a time until the quality of the bases within the grouping falls below a set threshold, prompting the sequence to be trimmed (Bolger *et al.*, 2014).

1.4.4 Functional annotation

Once sequences have been quality controlled, they may be functionally annotated directly or after being assembled into longer contiguous sequences (discussed in greater detail in 1.5.4). There are a number of methods to annotate sequences, including alignments, mapping, kmers or hidden Markov models (HMM). Local alignments offer a slow but precise method capable of finding short stretches of overlap between two sequences, providing a detailed placement (base positions) of where the query and reference overlap. Mappers or read recruiters in contrast are a method which tell us the approximate origin of a sequence opposed to the precise placement of the query sequence within the reference and therefore provides a fast annotation method (De Filippo *et al.*, 2012). A downside of both of these methods is that they assume the genome is linear. This assumption is not always

appropriate, due to recombination events, deletions, insertions and duplications (Vinga and Almeida, 2003; Zielezinski *et al.*, 2017). An alternate approach employs kmer searching to annotate sequences. A kmer is defined as a short stretch of nucleotide sequence of "k" bases long (e.g. the 4-mer AACT), the frequencies of which can be indexed for known functional genes, allowing rapid annotation of query sequences. This approach overcomes many of the issues identified above as it focuses on the sub sequences quantity opposed to their order (Vinga and Almeida, 2003; Zielezinski *et al.*, 2017). Kmer searching is also based on exact matches of strings which is computationally more efficient than similarity searching resulting in less computationally intensive routines with dramatic decreases in runtimes (Wood and Salzberg, 2014). Another approach to functional annotation uses hidden markov models (HMM's). HMMs provide a form of homology search (i.e. it can be used to detect sequences with a common evolutionary history) which is sensitive to the overall structure of a gene rather than possessing the greater level of precision that aligners do. This approach relies upon the principle that some components of a gene are more likely to be conserved than others, and uses state transition probability to assess sequence similarity (Eddy, 2011).

It is worth noting that all of these annotation approaches are limited by the number of reference genomes available within reference databases which are often biased to highly studied organisms including medically relevant taxa. Indeed we know that the majority of microbes in natural environments are yet to be characterised, meaning there is inadequate representation of many dominant soil taxa in genetic reference databases and so we are reliant largely on matches which are far from exact (Alneberg *et al.*, 2014).

Once genes are annotated it is desirable to understand the broader functions and biochemical pathways that they are likely to be contributing to. Indeed most functions cannot be attributed to a single gene and are instead related to a larger group of interacting gene products (Ogata *et al.*, 1999). This has led to the development of functional ontology classification systems linking individual genes to pathways, processes and structural complexes. There are numerous functional ontology frameworks including KEGG (Kyoto Encyclopedia of Genes and Genomes) (Ogata *et al.*, 1999), SEED (Overbeek *et al.*, 2005), MetaCyc (Caspi *et al.*, 2016), COG (Clusters of Orthologous Groups) (Kristensen *et al.*, 2010), GO (Gene Ontology) (Gene Ontology Consortium 2000) and FOAM (Functional Ontology

Assignments for Metagenomes) (Prestat *et al.*, 2014). With the exception of the latter system, these have largely been created for biochemical applications and challenges remain in utilising and developing these approaches for ecological interrogation of soil biogeochemical functions.

1.4.5 Assembly

Whilst short reads are often annotated and analysed directly, there are limitations in the amount of information we can derive from short sequences. Indeed many microbial genes are considerably longer (~1000 bp) than the reads produced by current commonly used sequencing platforms (Illumina hiseq 2 x150 bp) making many genes difficult to identify (van der Walt *et al.*, 2017). Short reads together with database limitations also presents challenges in terms of taxonomic classification of functionally annotated reads. Therefore an alternate approach is often taken whereby reads are assembled into longer contiguous sequences (contigs) (Ayling *et al.,* 2020). These methods both increase the chance of identifying genes and genomic signatures, as well as providing the opportunity to assemble novel genomes from uncultured taxa. Assembly also removes most random sequencing errors, although in turn it can introduce new errors which arise from the assembly process (Howe *et al.*, 2014; Ayling *et al.,* 2020). Assembly also makes downstream analyses much less computationally intensive due to the significant reduction in data when processing contigs opposed to reads, though there are large computational challenges in the first step of assembling reads from hyper diverse soil systems (Howe *et al.*, 2014) (described in the next section).There are two broad assembly methods, reference based assembly (requiring a taxonomic reference) and de novo assembly (which does not require a taxonomic reference) (Ayling et al., 2020). As the large majority of microbes are unculturable (including most dominant soil taxa) (Alneberg *et al.*, 2014) and we therefore lack reference genomes for them, the focus on the proceeding section is de-novo assembly.

1.4.5.1 De-novo Assembly methods

Assembly approaches have developed alongside the sequencing technologies used, taking into account features such as the sequencing technologies throughput and read length. The

assembly of Sanger sequencing has predominantly been conducted using overlap layout consensus methods (OLC) (Ayling *et al.,* 2020). OLC first conducts pair wise sequence alignments (Miller *et al.,* 2010) before constructing a graph whereby every node is a read and every edge indicates two nodes with sequence overlap, before eventually producing a consensus sequence (Li *et al.*, 2012; Nagarajan and Pop, 2013). As OLC compares overlaps of all reads it is a computationally intensive approach and therefore considered less practical for the second generation sequencing technologies (which are frequently used today) given they have considerably larger sequencing outputs (Ayling *et al.,* 2020). Short reads are instead typically assembled using de Bruijn graph assembly whereby reads are first split into all possible kmers (Compeau *et al*., 2011). A graph is then constructed whereby the nodes depict kmers and every edge represents overlaps by k -1 (Nagarajan and Pop, 2013). This algorithm is much more efficient in comparison to OLC, as the use of kmers means it circumvents the need to compare the overlaps of all reads (Pevzner et al., 2001; Howe and Chain, 2015).

Assembling large volumes of metagenome data is also a computational challenge requiring considerable resources (Howe *et al.*, 2014). These issues can be addressed directly by assembly algorithms, for example MEGAHIT reduces the memory needed to assemble by using a succinct de Bruijn algorithm (Li *et al.*, 2015), whilst other assembly algorithms can conduct graph construction on multiple nodes on a high performance computing cluster (Mahadik *et al.*, 2019). Pre-processing methods can also be used prior to assembly, to reduce the computational resources need to assemble the dataset. These methods include digital normalisation, which reduces the size of a dataset and results in more uniform levels of species sequence coverage and partitioning whereby reads are grouped by sequence overlap (Howe *et al.*, 2014).

The hyperdiverse nature of soils coupled with short read lengths, means assembling soil metagenomes is highly challenging, as it's difficult to obtain the amount of sequence representation needed to assemble complete or near complete genomes. It has previously been estimated that 50 Tbp of sequencing data is required to sample a gram of soil sufficiently (Howe *et al.*, 2014). The variation in the relative abundance of taxa, also results in the genomes of dominant taxa being covered by sequencing thousands of times, whilst

rarer taxa may be covered by just a few sequencing reads or not at all (Miller *et al*., 2014, Koren and Sutton, 2010; Howe and Chain, 2015). It is also an immense challenge to distinguish between different taxa in soils, as given the hyper diverse nature of soil microbial communities it is likely various closely related taxa may be present as well as strains/sub-species of the same taxa (Ayling *et al.,* 2020). An additional challenge in both soils and wider applications is that assembled contigs are typically still considerably shorter than the length of entire genomes and therefore often require "metagenomic binning" methods to retrieve more complete genomes (Kang *et al.,* 2016). Binning methods will be discussed in greater detail in the next section.

1.4.6 Binning assembled contigs

To overcome the challenges of retrieving genomes using assembly methods alone, binning is often implemented to group assembled sequences that are likely to be derived from the same genome. Binning methods can be broadly described as taxonomy-dependent and taxonomy-independent approaches. Taxonomy dependent methods rely on comparing the contigs to sequences within reference databases (or models derived from them) to classify contigs into bins. This can be done both by sequence alignments and through comparing GC content and codon usage of contigs to sequences/models in reference databases (Mande *et al.,* 2012*)*. These approaches are however both time consuming and assume that suitable reference genomes exist for your taxa of interest (Sedlar *et al.,* 2017) (which is often not the case when studying soil microbes) (Alneberg *et al.*, 2014).

A taxonomy independent approach is therefore more realistic for soil taxa, enabling insights into the genomes of otherwise inaccessible taxa. Taxonomy independent approaches work by employing clustering algorithms to group contigs into bins based on contig characteristics. One method is to use composition referring to the contigs sequence motifs such as tetranucleotide frequency (i.e. kmers, k=4) or GC content. This is based on the assumption that different phylotypes have different nucleotide compositions and therefore possess different relative frequencies of kmers (Graham *et al.,* 2016; Sedlar *et al.*, 2017). Of course, there are limitations to this approach, as closely related phylotypes may share similar kmer ratios, leading to contigs being grouped into the wrong bin (Breitwieser *et al.*,

2018). Likewise, genes recently acquired by a taxon through horizontal gene transfer (HGT), may also be misclassified (Dick *et al.*, 2009; Graham *et al.,* 2016). This has led many binning tools to also use coverage as an additional variable to ensure more accurate binning (Alneberg *et al.*, 2014; Wu *et al.*, 2014; Graham et al., 2016; Lin and Liao, 2016). These methods are based on the expectation that the coverage profiles of contigs derived from the same genome will be highly correlated across samples (Sedlar *et al.*, 2017).

1.5 Challenges: microbes into soil ecosystem service frameworks

1.5.1 General issues of microbial "Big Data"

Since the development of readily implementable molecular approaches to examine soil microbial biodiversity, the last 20 years has seen an explosion of studies evaluating soil microbial community responses to either natural or anthropogenic drivers. These studies have been conducted on the local geographic level (Buée *et al.*, 2009) but also on the landscape (Fierer and Jackson, 2006; Griffiths *et al.*, 2011) and even global scales (Leff *et al.*, 2015). The continual progress in technology has often meant the studies have become larger, as have the range and numbers of taxonomic units evaluated. At the most basic level this presents a number of challenges with respect to the ways in which ecological knowledge is inferred from appropriate statistical analyses of large multispecies datasets. The general aim of most ecological metagenomic analyses is to study how the collective microbial community varies in response to natural drivers, or other perturbations in experimental settings. Regardless of whether taxonomic or functional metagenomic approaches are used, the ultimate aim is typically to produce and analyse a table of gene counts across samples. As the number of reads (or indeed contigs, or MAGs) will vary across samples through non-biological methodology related mechanisms, the data often needs to first be normalised. This is typically conducted through either calculating the proportional abundance of each genomic feature per sample, or rarefaction which involves resampling datasets to standardise equivalent read numbers across samples. Both approaches suffer in that they only provide information on relative abundances, not absolute biomass, and so observed increases in one genomic feature could simply be indicative of decreases in another unrelated genomic feature. Rarefaction methods have also been criticised because

they implicitly involve removing large quantities of valid data (McMurdie and Holmes, 2014).

Analysing metagenome data also requires careful consideration due to its high dimensional nature, typically featuring thousands of genes, and a much smaller number of samples (commonly referred to as the large p small n problem) (Prosser, 2010; Touw *et al.*, 2012). This problem is likely exacerbated in soil metagenomes, because soil is a complex environment with a much higher species richness than that found in biomedical samples or non-terrestrial environmental systems (Daniel, 2005). Most studies begin with an exploratory analysis often involving ordination, whereby samples or taxonomic /functional features are "ordered" in multidimensional space with respect to similarity. There are a large number of ordination methodologies, but approaches typically can be classified as either "unconstrained" or "constrained", depending on whether associated environmental information is used to constrain the analyses to only that variance in composition which relates to specific environmental gradients. Commonly used ordination methods in microbial ecology include "unconstrained" multidimensional scaling (MDS), principal component analysis (PCA); and "constrained" canonical correspondence analysis (CCA) (Ramette, 2007). Statistical inference as to the strength of effects of either environmental gradients or imposed treatments can be gleaned from these types of analyses, either through relating associated explanatory variables to ordination axes (or dissimilarity metrics underlying the ordination) in the case of unconstrained ordination; or evaluating the proportion of variance explained by environmental variables in the case of constrained ordination.

Alongside the quantification of broad patterns or treatment effects on community similarity, a major goal of most contemporary soil microbial ecology studies is to actually identify specific taxa or genes responsive to change. Identifying specific taxa or functional genes relevant to a specific treatment or environmental gradient amongst a large noisy dataset is also highly challenging (Jonsson *et al.*, 2016), though many multivariate techniques are currently used. Approaches such as indicator species analyses (Dufrêne and Legendre, 1997) or simper (Warton *et al.,* 2012) are inherited from the field of broader ecology, and are typically used to identify taxa responsive to paired treatments; whereas taxa responsive to

environmental gradients can be extracted from the outputs of ordination analyses. Newer methods such as deseq2 (Love *et al.*, 2014) have been developed specifically for molecular applications, in particular differential expression, and are claimed to work better for unnormalized sparse molecular data.

1.5.2 The need for synthesis

Despite the methodological challenges outlined above, thousands of studies across the globe utilising principally amplicon approaches are generating much needed information, both on the breadth of microbial diversity existing in different soil systems and also their sensitivity to change. However, we are still some way from synthesising this new knowledge on the ecology of these novel organisms, as there is currently no formalised way of capturing this information other than in journal articles. Journal articles simply do not provide the space required to report on the responses of the large amounts of microbial taxa typically undergoing change in soil systems. Furthermore, whilst we have excellent digital tools for the storage and taxonomic characterisation of novel recovered sequences, such as Genbank (Benson *et al.*, 2013) or EBI sequence repository and resources (Amid et al., 2020); these databases include limited information on ecological attributes (for example a habitat classification). Synthesising relationships between soil microbial taxa and environmental parameters is now necessary to progress ecological understanding of soil microbes beyond those few organisms that are readily cultivated. This is particularly true in relation to building an understanding of how soil management affects soil ecosystem services. For example, the diversity of different soils that exist in different climates means that it is likely that different taxa will respond to future change (be it from climate, land use, or some other perturbation). Therefore, understanding how changes in biodiversity affects ecosystem services in different soil contexts requires a fundamental predictive understanding of native biodiversity distributions in the first instance.

1.5.3 Linking taxonomic change to functional change

Metagenomic methodologies have permitted numerous microbial biogeographical studies across the globe. As a result, we are beginning to understand and identify the various environmental factors which strongly influence soil microbial communities. We know for

example that soil pH is extremely influential on soil bacterial biodiversity (Fierer and Jackson, 2006; Griffiths *et al.*, 2011), with certain broad groups of newly discovered taxa responding similarly to soil pH irrespective of other geographic factors. These findings corroborate to some degree the long standing theory put forward by Lourens Gerhard Marinus Baas Becking that "everything is everywhere, but the environment selects" meaning that microbes are globally distributed but only capable of proliferating in environmentally favourable conditions. We do not however know what these findings regarding microbial distributions mean in terms of microbial functioning (Green *et al.,* 2008). This alongside the previously mentioned difficulties in synthesising how specific taxa or "phylotypes" respond to change means we are far from being able to predict change in soil microbes and consequences for soil functioning and derived services.

Previously it has been assumed that microbes possess a large degree of functional redundancy, as they have the ability to exchange genomic elements through lateral gene transfer (Cohan and Koeppel, 2008; Martiny *et al.*, 2015). However it has been difficult to empirically demonstrate this given past technological challenges and the large diversity of uncharacterised soil organisms (Alneberg *et al.*, 2014). Past questioning in this area has centred on determining whether diversity is important to specific soil functions or rather whether the presence or absence of specific taxa is more critical. Studies which have specifically manipulated soil microbial diversity, do appear to show evidence for redundancy at least for broad soil functions such as soil respiratory activity (Stres *et al.*, 2010). However more specific processes such as those involved in pollution tolerance (Brandt *et al.*, 2010) and N cycling (Steenwerth *et al.*, 2005) are more sensitive to changes in diversity. Coupling synthesised knowledge on taxonomic responses to environmental change with detailed genetic information on the functional capabilities of responsive taxa therefore offers the potential to better understanding the importance of microbial diversity for soil ecosystem functioning; alongside building a more mechanistic and predictive framework for understanding functional resistance and resilience.

Theoretical frameworks already exist describing relationships between taxonomic responses to environmental change and the functional capabilities of responders. Functional trait theory is a framework commonly applied to plants and has proven useful in predicting

ecosystem functioning. The framework is broadly based on the concept of “response groups” to refer to organisms that respond similarly to environmental stressors or gradients, and “effect groups” to refer to organisms that affect ecosystem functioning (Suding *et al.*, 2008). The degree of coupling between these two groups can be used to assess the underpinning mechanisms determining how biodiversity regulates ecosystem functioning with environmental change scenarios. Quantification of the degree of coupling between response and effect traits can help assess the resilience of soils to environmental change including land use transitions. For example, if response groups (e.g. pH affected bacteria) and effect groups (e.g. N fixers) comprise of the same organisms then a change in soil conditions such as pH could lead to change in the soil function. Conversely, if response groups and effect groups have little taxonomic overlap, then changes in the soil environment would have little effect on soil functioning via the so called “insurance hypothesis” (**Fig**.**1.2**).

Though there has been recent discussion of implementing trait-based frameworks for microbes, these approaches are largely underdeveloped. Microbial ecology arguably has extensive opportunity to apply such trait based frameworks given the wealth of genomic data we are able to obtain providing insights both into phylogeny and function (through amplicon and metagenome studies respectively) (Martiny *et al.*, 2015). More generally there is a broader fundamental need to actually identify which genes underpin microbial responses to environmental stressors and novel genes responsible for soil functions. It is noteworthy that these response effect trait frameworks can equally be applied to specific traits at the genetic level.

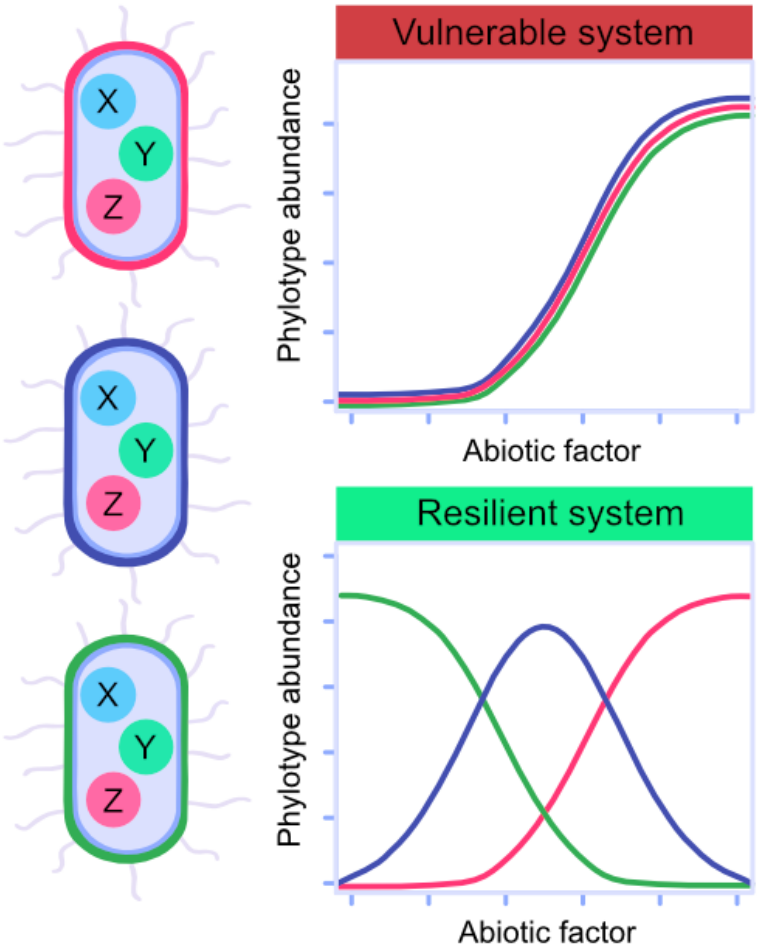


Fig.1.2. Representation of how coupling of functional and response traits could impact on soil system resilience. A community of microbes possessing the same functions (x, y, z) and responding similarly to environmental driver would cause the soil system to have greater vulnerability to environmental change. If these phylotypes vary in terms of their response to the abiotic factor, the soil system may possess greater resilience.

1.6 Project aims

This thesis aims to use landscape scale genomic datasets to develop tools and models for predicting and understanding soil microbial response and effect traits. The approaches will also assess how resilient microbial taxonomic and functional biodiversity is to environmental change both in terms of edaphic properties and land use factors.

The specific objectives are:

1) To determine pH responses of bacteria phylotypes across a nationwide soil survey.

Given that microbial responses to environmental variables are typically reported at a broad taxonomic scale (leading to a loss of finer scale trait information), **chapter 2** will model pH responses of taxa at the OTU level using 16S rRNA gene amplicon inventories from a large survey of GB. In addition, this chapter will present a tool to disseminate ecological information on the ecological response traits of a large number of bacterial phylotypes.

2) To examine broad functional changes in microbial communities in response to land use change.

In **chapter 3** l will investigate broad changes in functional genetic profiles in response to land use, through analysing a large metagenome dataset encompassing paired land use contrasts, distributed across the UK.

3) To link specific taxa with environmental responses and functional capacities.

My final data chapters (chapters **4-5**) will aim to build on my previous chapters by examining both the functional capabilities of novel soil bacterial taxa and their responses to soil change. **Chapter 4** will focus on examining soil organic matter decomposition through assessing the taxonomy of secreted Glycoside Hydrolase genes in assembled contigs obtained from a long term pH manipulation field experiment. **Chapter 5** will attempt to assemble and bin metagenomes from all paired land use sites to examine both functional composition and environmental distributions of novel metagenome assembled genomes (MAGs).

1.7 Bibliography

Adhikari, K. and Hartemink, A. E. (2016) 'Linking soils to ecosystem services - A global review', *Geoderma*. Elsevier B.V., pp. 101–111. doi: 10.1016/j.geoderma.2015.08.009.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., *et al.* (2014) 'Binning metagenomic contigs by coverage and composition', *Nature Methods*, 11(11), pp. 1144–1146. doi: 10.1038/nmeth.3103.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., *et al.* (2014) 'Binning metagenomic contigs by coverage and composition', *Nature Methods*, 11(11), pp. 1144–1146. doi: 10.1038/nmeth.3103.

Alori, E. T., Glick, B. R. and Babalola, O. O. (2017) 'Microbial phosphorus solubilization and its potential for use in sustainable agriculture', *Frontiers in Microbiology*. Frontiers Media S.A., p. 971. doi: 10.3389/fmicb.2017.00971.

Amid, C. *et al.* (2020) 'The European Nucleotide Archive in 2019', *Nucleic Acids Research*. Oxford University Press, 48(D1), pp. D70–D76. doi: 10.1093/nar/gkz1063.

Ayling, M., Clark, M. D. and Leggett, R. M. (2020) 'New approaches for metagenome assembly with short reads', *Briefings in Bioinformatics*. Oxford University Press, pp. 584–594. doi: 10.1093/bib/bbz020.

Banerjee, S. *et al.* (2016) 'Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil', *Soil Biology and Biochemistry*, 97, pp. 188–198. doi: 10.1016/j.soilbio.2016.03.017.

Benjamini, Y. and Speed, T. P. (2012) 'Summarizing and correcting the GC content bias in high-throughput sequencing', *Nucleic Acids Research*, 40(10). doi: 10.1093/nar/gks001.

Bennett, E. M., Peterson, G. D. and Gordon, L. J. (2009) 'Understanding relationships among multiple ecosystem services', *Ecology Letters*, 12(12), pp. 1394–1404. doi: 10.1111/j.1461-0248.2009.01387.x.

Benson, D. A. et al. (2013) ‘GenBank’, Nucleic Acids Research. Nucleic Acids Res, 41(D1). doi: 10.1093/nar/gks1195.

Berini, F. *et al.* (2017) ‘Metagenomics: Novel enzymes from non-culturable microbes’, *FEMS Microbiology Letters*. Oxford University Press, p. 211. doi: 10.1093/femsle/fnx211.

Bharti, R. and Grimm, D. G. (2019) ‘Current challenges and best-practice protocols for microbiome analysis’, *Briefings in Bioinformatics*, 2019(00), pp. 1–16. doi: 10.1093/bib/bbz155.

Blum, W. E. H. (2005) ‘Functions of soil for society and the environment’, *Reviews in Environmental Science and Biotechnology*. Springer, pp. 75–79. doi: 10.1007/s11157-005-2236-x.

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

Brandt, K. K. *et al.* (2010) ‘Development of pollution-induced community tolerance is linked to structural and functional resilience of a soil bacterial community following a five-year field exposure to copper’, *Soil Biology and Biochemistry*, 42(5), pp. 748–757. doi: 10.1016/j.soilbio.2010.01.008.

Branton, D. *et al.* (2008) ‘The potential and challenges of nanopore sequencing’, *Nature Biotechnology*, 26(10), pp. 1146–1153. doi: 10.1038/nbt.1495.The.

Breitwieser, F. P., Lu, J. and Salzberg, S. L. (2018) ‘A review of methods and databases for metagenomic classification and assembly’, *Briefings in Bioinformatics*, 20(4), pp. 1125–1139. doi: 10.1093/bib/bbx120.

Browne, P. *et al.* (2009) ‘Superior inorganic phosphate solubilization is linked to phylogeny within the Pseudomonas fluorescens complex’, *Applied Soil Ecology*, 43(1), pp. 131–138. doi: 10.1016/j.apsoil.2009.06.010.

Buée, M. *et al.* (2009) ‘454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity’, *New Phytologist*, 184(2), pp. 449–456. doi: 10.1111/j.1469-8137.2009.03003.x.

Bižić, M. et al. (2020) ‘Aquatic and terrestrial cyanobacteria produce methane’, Science Advances. American Association for the Advancement of Science, 6(3), p. eaax5343. doi: 10.1126/sciadv.aax5343.

Caporaso, J. G. *et al.* (2011) ‘Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(SUPPL. 1), pp. 4516–4522. doi: 10.1073/pnas.1000080107.

Caspi, R. *et al.* (2016) ‘The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases’, *Nucleic Acids Research*, 44(D1), pp. D471–D480. doi: 10.1093/nar/gkv1164.

Castillo Villamizar, G. A. *et al.* (2019) ‘Functional metagenomics reveals an overlooked diversity and novel features of soil-derived bacterial phosphatases and phytases’, *mBio*, 10(1). doi: 10.1128/mBio.01966-18.

Chen, Y. C. *et al.* (2013) ‘Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly’, *PLoS ONE*, 8(4), p. 62856. doi: 10.1371/journal.pone.0062856.

Cohan, F. M. and Koeppel, A. F. (2008) ‘The Origins of Ecological Diversity in Prokaryotes’, *Current Biology*. Cell Press, pp. R1024–R1034. doi: 10.1016/j.cub.2008.09.014.

Compeau, P. E. C., Pevzner, P. A. and Tesler, G. (2011) ‘How to apply de Bruijn graphs to genome assembly’, *Nature Biotechnology*, 29(11), pp. 987–991. doi: 10.1038/nbt.2023.

Consortium, T. G. O. (2000) ‘Gene ontology: Tool for the identification of biology.’, *Natural Genetics*, 25(may), pp. 25–29.

D’Costa, V. M., Griffiths, E. and Wright, G. D. (2007) ‘Expanding the soil antibiotic resistome: exploring environmental diversity’, *Current Opinion in Microbiology*. Elsevier Current Trends, pp. 481–489. doi: 10.1016/j.mib.2007.08.009.

Daims, H. et al. (2015) ‘Complete nitrification by Nitrospira bacteria’, Nature. Nature Publishing Group, 528(7583), pp. 504–509. doi: 10.1038/nature16461.

Daniel, R. (2005) ‘The metagenomics of soil’, *Nature Reviews Microbiology*, pp. 470–478. doi: 10.1038/nrmicro1160.

Dear, P. H. (2003) 'One by one: Single molecule tools for genomics', *Briefings in Functional Genomics and Proteomics*, 1(4), pp. 397–416. doi: 10.1093/bfgp/1.4.397.

Delmont, T. O. *et al.* (2011) 'Metagenomic comparison of direct and indirect soil DNA extraction approaches', *Journal of Microbiological Methods*, 86(3), pp. 397–400. doi: 10.1016/j.mimet.2011.06.013.

Demain, A. L. and Adrio, J. L. (2008) 'Contributions of microorganisms to industrial biology', *Molecular Biotechnology*, 38(1), pp. 41–55. doi: 10.1007/s12033-007-0035-z.

Demenois, J. *et al.* (2020) 'Barriers and Strategies to Boost Soil Carbon Sequestration in Agriculture', *Frontiers in Sustainable Food Systems*, 4, p. 37. doi: 10.3389/fsufs.2020.00037.

Dick, G. J. *et al.* (2009) 'Community-wide analysis of microbial genome sequence signatures.', *Genome biology*, 10(8), p. R85. doi: 10.1186/gb-2009-10-8-r85.

van Dijk, E. L. *et al.* (2018) 'The Third Revolution in Sequencing Technology', *Trends in Genetics*, pp. 666–681. doi: 10.1016/j.tig.2018.05.008.

Dufrêne, M. and Legendre, P. (1997) 'Species assemblages and indicator species: The need for a flexible asymmetrical approach', *Ecological Monographs*, 67(3), pp. 345–366. doi: 10.2307/2963459.

Eddy, S. R. (2011) 'Accelerated profile HMM searches', *PLoS Computational Biology*, 7(10). doi: 10.1371/journal.pcbi.1002195.

van Elsas, J. D. *et al.* (2008) 'The metagenomics of disease-suppressive soils - experiences from the METACONTROL project', *Trends in Biotechnology*. Trends Biotechnol, pp. 591–601. doi: 10.1016/j.tibtech.2008.07.004.

van Elsas, J. D. and Boersma, F. G. H. (2011) 'A review of molecular methods to study the microbiota of soil and the mycosphere', *European Journal of Soil Biology*, 47(2), pp. 77–87. doi: 10.1016/j.ejsobi.2010.11.010.

Emerson, J. B. (2019) 'Soil Viruses: A New Hope', *mSystems*, 4(3). doi: 10.1128/msystems.00120-19.

Eriksen, J. (2009) 'Chapter 2 Soil Sulfur Cycling in Temperate Agricultural Systems', *Advances in Agronomy*. Academic Press Inc., pp. 55–89. doi: 10.1016/S0065-2113(09)01002-5.

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. doi: 10.3389/fgene.2015.00348.

Evans, P. N. *et al.* (2019) 'An evolving view of methane metabolism in the Archaea', *Nature Reviews Microbiology*. Nature Publishing Group, pp. 219–232. doi: 10.1038/s41579-018-0136-7.

Ewing, B. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. II. Error probabilities', *Genome Research*, 8(3), pp. 186–194. doi: 10.1101/gr.8.3.186.

Fahad, S. *et al.* (2015) 'Potential role of phytohormones and plant growth-promoting rhizobacteria in abiotic stresses: consequences for changing environment', *Environmental Science and Pollution Research*, 22(7), pp. 4907–4921. doi: 10.1007/s11356-014-3754-2.

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. doi: 10.1126/science.1153213.

FAO and ITPS (2018) *Status of the World's Soil Resources.*, *Intergovernmental Technical Panel on Soils*. Available at: www.fao.org/publications%0Ahttp://www.fao.org/3/a-i5199e.pdf.

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. doi: 10.1073/pnas.0507535103.

De Filippo, C. *et al.* (2012) 'Bioinformatic approaches for functional annotation and pathway inference in metagenomics data', *Briefings in Bioinformatics*, 13(6), pp. 696–710. doi: 10.1093/bib/bbs070.

Foley, J. a *et al.* (2005) 'Global consequences of land use.', *Science (New York, N.Y.)*, 309(5734), pp. 570–4. doi: 10.1126/science.1111772.

Forbes, K. J. *et al.* (1991) 'Rapid methods in bacterial DNA fingerprinting', *Journal of General Microbiology*, 137(9), pp. 2051–2058. doi: 10.1099/00221287-137-9-2051.

Foster, R. C. (1988) 'Microenvironments of soil microorganisms', *Biology and Fertility of Soils*, 6(3), pp. 189–203. doi: 10.1007/BF00260816.

Francioli, D. *et al.* (2016) 'Mineral vs. organic amendments: Microbial community structure, activity and abundance of agriculturally relevant microbes are driven by long-term fertilization strategies', *Frontiers in Microbiology*, 7(SEP). doi: 10.3389/fmicb.2016.01446.

Fromin, N. *et al.* (2002) 'Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns', *Environmental Microbiology*, pp. 634–643. doi: 10.1046/j.1462-2920.2002.00358.x.

Gahan, J. and Schmalenberger, A. (2014) 'The role of bacteria and mycorrhiza in plant sulfur supply', *Frontiers in Plant Science*, 5(DEC). doi: 10.3389/fpls.2014.00723.

Galloway, J. N. *et al.* (2003) 'The nitrogen cascade', *BioScience*. American Institute of Biological Sciences, pp. 341–356. doi: 10.1641/0006-3568(2003)053[0341:TNC]2.0.CO;2.

Galloway, J. N. *et al.* (2008) 'Transformation of the nitrogen cycle: Recent trends, questions, and potential solutions', *Science*. American Association for the Advancement of Science, pp. 889–892. doi: 10.1126/science.1136674.

Gao, J. *et al.* (2015) 'The impact of land-use change on water-related ecosystem services: A study of the Guishui River Basin, Beijing, China', *Journal of Cleaner Production*, 163, pp. S148–S155. doi: 10.1016/j.jclepro.2016.01.049.

Giani, A. M. *et al.* (2020) 'Long walk to genomics: History and current approaches to genome sequencing and assembly', *Computational and Structural Biotechnology Journal*. Elsevier B.V., pp. 9–19. doi: 10.1016/j.csbj.2019.11.002.

Gougoulias, C., Clark, J. M. and Shaw, L. J. (2014) 'The role of soil microbes in the global carbon cycle: Tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems', *Journal of the Science of Food and Agriculture*, 94(12), pp. 2362–2371. doi: 10.1002/jsfa.6577.

Govaerts, B. *et al.* (2009) 'Conservation agriculture and soil carbon sequestration: Between myth and farmer reality', *Critical Reviews in Plant Sciences*, 28(3), pp. 97–122. doi: 10.1080/07352680902776358.

Graf, D. R. H., Jones, C. M. and Hallin, S. (2014) 'Intergenomic comparisons highlight modularity of the denitrification pathway and underpin the importance of community

structure for N2O emissions', *PLoS ONE*. Edited by V. de Crécy-Lagard, 9(12), p. e114118. doi: 10.1371/journal.pone.0114118.

Graham, E., Hiedelberg, J. and Tully, B. (2016) 'BinSanity: Unsupervised Clustering of Environmental Microbial Assemblies Using Coverage and Affinity Propagation'.

Green, J. L., Bohannan, B. J. M. and Whitaker, R. J. (2008) 'Microbial biogeography: From taxonomy to traits', *Science*. American Association for the Advancement of Science, pp. 1039–1043. doi: 10.1126/science.1153475.

Griffiths, B. S. (1994) 'Microbial-feeding nematodes and protozoa in soil: Their effectson microbial activity and nitrogen mineralization in decomposition hotspots and the rhizosphere', *Plant and Soil*, 164(1), pp. 25–33. doi: 10.1007/BF00010107.

Griffiths, R. I. *et al.* (2011) 'The bacterial biogeography of British soils', *Environmental Microbiology*, 13(6), pp. 1642–1654. doi: 10.1111/j.1462-2920.2011.02480.x.

Guo, L. B. and Gifford, R. M. (2002) 'Soil carbon stocks and land use change: A meta analysis', *Global Change Biology*, 8(4), pp. 345–360. doi: 10.1046/j.1354-1013.2002.00486.x.

Gupta, P. *et al.* (2017) 'Comparison of Metagenomic DNA Extraction Methods for Soil Sediments of High Elevation Puga Hot Spring in Ladakh, India to Explore Bacterial Diversity', *Geomicrobiology Journal*, 34(4), pp. 289–299. doi: 10.1080/01490451.2015.1128995.

Hall, N. (2007) 'Advanced sequencing technologies and their wider impact in microbiology', *Journal of Experimental Biology*. The Company of Biologists Ltd, pp. 1518–1525. doi: 10.1242/jeb.001370.

Han, S. *et al.* (2018) 'Nitrite-oxidizing bacteria community composition and diversity are influenced by fertilizer regimes, but are independent of the soil aggregate in acidic subtropical red soil', *Frontiers in Microbiology*, 9(MAY), p. 885. doi: 10.3389/fmicb.2018.00885.

Handelsman, J. *et al.* (1998) 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.', *Chemistry & biology*, 5(10), pp. R245–R249. doi: 10.1016/S1074-5521(98)90108-9.

Harry, M. *et al.* (1999) 'Evaluation of purification procedures for DNA extracted from organic rich samples: Interference with humic substances', *Analusis*, 27(5), pp. 439–442. doi: 10.1051/analusis:1999270439.

Hartmann, M. *et al.* (2015) 'Distinct soil microbial diversity under long-term organic and conventional farming', *ISME Journal*, 9(5), pp. 1177–1194. doi: 10.1038/ismej.2014.210.

Hayat, R. *et al.* (2010) 'Soil beneficial bacteria and their role in plant growth promotion: A review', *Annals of Microbiology*, 60(4), pp. 579–598. doi: 10.1007/s13213-010-0117-1.

Heather, J. M. and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*. Academic Press Inc., pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Howe, A. C., Jansson, Janet K., *et al.* (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. doi: 10.1073/pnas.1402564111.

Howe, A. C., Jansson, Janet K, *et al.* (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. doi: 10.1073/pnas.1402564111.

Howe, A. and Chain, P. S. G. (2015) 'Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial)', *Frontiers in Microbiology*, 6(JUL), pp. 10–13. doi: 10.3389/fmicb.2015.00678.

Huang, R. *et al.* (2019) 'Plant–microbe networks in soil are weakened by century-long use of inorganic fertilizers', *Microbial Biotechnology*, 12(6), pp. 1464–1475. doi: 10.1111/1751-7915.13487.

Humbert, S. *et al.* (2010) 'Molecular detection of anammox bacteria in terrestrial ecosystems: Distribution and diversity', *ISME Journal*, 4(3), pp. 450–454. doi: 10.1038/ismej.2009.125.

Jansson, J. K. and Hofmockel, K. S. (2020) 'Soil microbiomes and climate change', *Nature Reviews Microbiology*, 18(1), pp. 35–46. doi: 10.1038/s41579-019-0265-7.

Jonsson, V. *et al.* (2016) 'Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics', *BMC Genomics*, 17(1), p. 78. doi: 10.1186/s12864-016-2386-y.

Falkinham, J. O. *et al.* (2009) 'Proliferation of antibiotic-producing bacteria and concomitant antibiotic production as the basis for the antibiotic activity of Jordan's red soils', *Applied and Environmental Microbiology*, 75(9), pp. 2735–2741. doi: 10.1128/AEM.00104-09.

Joshi, N. A., Fass, J. N. and others (2011) 'Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software]'.

Kallenbach, C. M., Frey, S. D. and Grandy, A. S. (2016) 'Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls', *Nature Communications*, 7. doi: 10.1038/ncomms13630.

Kang, D. D., Rubin, E. M. and Wang, Z. (2016) 'Reconstructing single genomes from complex microbial communities', *it - Information Technology*, 58(3), pp. 133–139. doi: 10.1515/itit-2016-0011.

Kannan, V., Bastas, K. and Devi, R. (2015) 'Scientific and Economic Impact of Plant Pathogenic Bacteria', in *Sustainable Approaches to Controlling Plant Pathogenic Bacteria*. CRC Press, pp. 369–392. doi: 10.1201/b18892-21.

Katz, M., Hover, B. M. and Brady, S. F. (2016) 'Culture-independent discovery of natural products from soil metagenomes', *Journal of Industrial Microbiology and Biotechnology*, 43(2–3), pp. 129–141. doi: 10.1007/s10295-015-1706-6.

Kchouk, M., Gibrat, J. F. and Elloumi, M. (2017) 'Generations of Sequencing Technologies: From First to Next Generation', *Biology and Medicine*, 09(03). doi: 10.4172/0974-8369.1000395.

Kertesz, M. A. and Mirleau, P. (2004) 'The role of soil microbes in plant sulphur nutrition', in *Journal of Experimental Botany*. Oxford Academic, pp. 1939–1945. doi: 10.1093/jxb/erh176.

Klose, S. *et al.* (2015) 'Sulfur Cycle Enzymes', in. John Wiley & Sons, Ltd, pp. 125–159. doi: 10.2136/sssabookser9.c7.

Kristensen, D. M. *et al.* (2010) 'A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches', *Bioinformatics*, 26(12), pp. 1481–1487. doi: 10.1093/bioinformatics/btq229.

Kulski, J. K. (2016) 'Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications', in *Next Generation Sequencing - Advances, Applications and Challenges*. InTech. doi: 10.5772/61964.

Kuypers, M. M. M., Marchant, H. K. and Kartal, B. (2018) 'The microbial nitrogen-cycling network', *Nature Reviews Microbiology*, 16(5), pp. 263–276. doi: 10.1038/nrmicro.2018.9.

Ladoukakis, E., Kolisis, F. N. and Chatziioannou, A. A. (2014) 'Integrative workflows for metagenomic analysis.', *Frontiers in cell and developmental biology*, 2(November), p. 70. doi: 10.3389/fcell.2014.00070.

Lakay, F. M., Botha, A. and Prior, B. A. (2007) 'Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils', *Journal of Applied Microbiology*, 102(1), pp. 265–273. doi: 10.1111/j.1365-2672.2006.03052.x.

Ledford, H. (2008) 'The death of microarrays?', *Nature*. Nature Publishing Group, p. 847. doi: 10.1038/455847a.

Lee, M. H. and Lee, S.-W. (2013) 'Bioprospecting Potential of the Soil Metagenome: Novel Enzymes and Bioactivities', *Genomics & Informatics*, 11(3), p. 114. doi: 10.5808/gi.2013.11.3.114.

Leff, J. W. *et al.* (2015) 'Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe', *Proceedings of the National Academy of Sciences of the United States of America*, 112(35), pp. 10967–10972. doi: 10.1073/pnas.1508382112.

Leff, L. G. *et al.* (1995) 'Comparison of methods of DNA extraction from stream sediments', *Applied and Environmental Microbiology*, 61(3), pp. 1141–1143. doi: 10.1128/aem.61.3.1141-1143.1995.

Li, D. *et al.* (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. doi: 10.1093/bioinformatics/btv033.

Li, Z. *et al.* (2012) 'Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph', *Briefings in Functional Genomics*, 11(1), pp. 25–37. doi: 10.1093/bfgp/elr035.

Liang, J. L. *et al.* (2020) 'Novel phosphate-solubilizing bacteria enhance soil phosphorus cycling following ecological restoration of land degraded by mining', *ISME Journal*, pp. 1600–1613. doi: 10.1038/s41396-020-0632-4.

Lin, H.-H. and Liao, Y.-C. (2016) 'Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes.', *Scientific reports*, 6(October 2015), p. 24175. doi: 10.1038/srep24175.

Lindahl, B. D. *et al.* (2013) 'Fungal community analysis by high-throughput sequencing of amplified markers - a user's guide', *New Phytologist*, 199(1), pp. 288–299. doi: 10.1111/nph.12243.

Liu, B. R. *et al.* (2006) 'A review of methods for studying microbial diversity in soils', *Pedosphere*, 16(1), pp. 18–24. doi: 10.1016/S1002-0160(06)60021-0.

Loit, K. *et al.* (2019) 'Relative performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) thirdgeneration sequencing instruments in identification of agricultural and forest fungal pathogens', *Applied and Environmental Microbiology*, 85(21). doi: 10.1128/AEM.01368-19.

Lombard, N. *et al.* (2011) 'Soil-specific limitations for access and analysis of soil microbial communities by metagenomics', *FEMS Microbiology Ecology*. Oxford Academic, pp. 31–49. doi: 10.1111/j.1574-6941.2011.01140.x.

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(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lu, X., Seuradge, B. J. and Neufeld, J. D. (2016) 'Biogeography of soil Thaumarchaeota in relation to soil depth and land usage', *FEMS Microbiology Ecology*, 93(2). doi: 10.1093/femsec/fiw246.

Lucheta, A. R. and Lambais, M. R. (2012) 'Enxofre na Agricultura', *Revista Brasileira de Ciencia do Solo*, 36(5), pp. 1369–1379. doi: 10.1590/S0100-06832012000500001.

Mahadik, K. *et al.* (2019) 'Scalable Genome Assembly through Parallel de Bruijn Graph Construction for Multiple k-mers', *Scientific Reports*, 9(1), pp. 1–15. doi: 10.1038/s41598-019-51284-9.

Mande, S. S., Mohammed, M. H. and Ghosh, T. S. (2012) 'Classification of metagenomic sequences: Methods and challenges', *Briefings in Bioinformatics*, 13(6), pp. 669–681. doi: 10.1093/bib/bbs054.

Martens-Habbena, W. *et al.* (2009) 'Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria', *Nature*, 461(7266), pp. 976–979. doi: 10.1038/nature08465.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10–12. doi: 10.14806/ej.17.1.200.

Martínez, J. L. (2008) 'Antibiotics and antibiotic resistance genes in natural environments', *Science*. American Association for the Advancement of Science, pp. 365–367. doi: 10.1126/science.1159483.

Martins, P. M. M. *et al.* (2018) 'Persistence in phytopathogenic bacteria: Do we know enough?', *Frontiers in Microbiology*. Frontiers Media S.A., p. 1099. doi: 10.3389/fmicb.2018.01099.

Martiny, J. B. H. *et al.* (2015) 'Microbiomes in light of traits: A phylogenetic perspective', *Science*, 350(6261). doi: 10.1126/science.aac9323.

McMurdie, P. J. and Holmes, S. (2014) 'Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible', *PLoS Computational Biology*, 10(4). doi: 10.1371/journal.pcbi.1003531.

Miller, J. R., Koren, S. and Sutton, G. (2010) 'Assembly algorithms for next-generation sequencing data', *Genomics*. Academic Press, pp. 315–327. doi: 10.1016/j.ygeno.2010.03.001.

Sabale, S. *et al*. (2020) 'Soil Metagenomics: Concepts and Applications', in *Metagenomics - Basics, Methods and Applications*. IntechOpen. doi: 10.5772/intechopen.88958.

Nagarajan, N. and Pop, M. (2013) 'Sequence assembly demystified.', *Nature reviews. Genetics*, 14(3), pp. 157–67. doi: 10.1038/nrg3367.

Nannipieri, P. *et al.* (2003) 'Microbial diversity and soil functions', *European Journal of Soil Science*, (54), pp. 655–670. doi: 10.1046/j.1365-2389.2003.00556.x.

Nazaries, L. *et al.* (2013) 'Methane, microbes and models: Fundamental understanding of the soil methane cycle for future predictions', *Environmental Microbiology*. John Wiley & Sons, Ltd, pp. 2395–2417. doi: 10.1111/1462-2920.12149.

van Niftrik, L. and Jetten, M. S. M. (2012) 'Anaerobic Ammonium-Oxidizing Bacteria: Unique Microorganisms with Exceptional Properties', *Microbiology and Molecular Biology Reviews*, 76(3), pp. 585–596. doi: 10.1128/mmbr.05025-11.

Ogata, H. *et al.* (1999) 'KEGG: Kyoto encyclopedia of genes and genomes', *Nucleic Acids Research*, 27(1), pp. 29–34. doi: 10.1093/nar/27.1.29.

Orellana, L. H. *et al.* (2018) 'Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization', *Applied and Environmental Microbiology*, 84(2). doi: 10.1128/AEM.01646-17.

Orgiazzi, A. *et al.* (2015) 'Soil biodiversity and DNA barcodes: Opportunities and challenges', *Soil Biology and Biochemistry*, 80, pp. 244–250. doi: 10.1016/j.soilbio.2014.10.014.

Ortíz-Castro, R. *et al.* (2009) 'The role of microbial signals in plant growth and development', *Plant Signaling and Behavior*. Landes Bioscience, pp. 701–712. doi: 10.4161/psb.4.8.9047.

Overbeek, R. *et al.* (2005) 'The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes', *Nucleic Acids Research*, 33(17), pp. 5691–5702. doi: 10.1093/nar/gki866.

Pennock, D. (2019) *Soil erosion: the greatest challenge to sustainable soil management*, *Food and Agriculture Organization of the United Nations*. Food and Agriculture Organization of the United Nations Rome, 2019. Available at: http://www.fao.org/3/ca4395en/ca4395en.pdf.

Pershina, E. *et al.* (2015) 'Comparative analysis of prokaryotic communities associated with organic and conventional farming systems', *PLoS ONE*. Edited by A. M. Ibekwe, 10(12), p. e0145072. doi: 10.1371/journal.pone.0145072.

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) 'An Eulerian path approach to DNA fragment assembly.', *Proceedings of the National Academy of Sciences*, 98(17), pp. 9748–53. doi: 10.1073/pnas.171285098.

Plomin, R. and Schalkwyk, L. C. (2007) 'Microarrays', *Developmental Science*, pp. 19–23. doi: 10.1111/j.1467-7687.2007.00558.x.

Pollard, M. O. *et al.* (2018) 'Long reads: their purpose and place', *Human molecular genetics*. NLM (Medline), pp. R234–R241. doi: 10.1093/hmg/ddy177.

Powlson, D. S. *et al.* (2011) 'Soil management in relation to sustainable agriculture and ecosystem services', *Food Policy*, 36(SUPPL. 1), pp. S72–S87. doi: 10.1016/j.foodpol.2010.11.025.

Prestat, E. *et al.* (2014) 'FOAM (Functional Ontology Assignments for Metagenomes): A Hidden Markov Model (HMM) database with environmental focus', *Nucleic Acids Research*, 42(19), pp. 1–7. doi: 10.1093/nar/gku702.

Prosser, J. I. (2010) 'Replicate or lie', *Environmental Microbiology*, 12(7), pp. 1806–1810. doi: 10.1111/j.1462-2920.2010.02201.x.

Quail, M. A. *et al.* (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers', *BMC Genomics*, 13(1), p. 341. doi: 10.1186/1471-2164-13-341.

Ramette, A. (2007) 'Multivariate analyses in microbial ecology', *FEMS Microbiology Ecology*. Wiley-Blackwell, pp. 142–160. doi: 10.1111/j.1574-6941.2007.00375.x.

Rang, F. J., Kloosterman, W. P. and de Ridder, J. (2018) 'From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy', *Genome Biology*. BioMed Central Ltd., p. 90. doi: 10.1186/s13059-018-1462-9.

Raynaud, X., Lata, J. C. and Leadley, P. W. (2006) 'Soil microbial loop and nutrient uptake by plants: A test using a coupled C:N model of plant-microbial interactions', *Plant and Soil*, 287(1–2), pp. 95–116. doi: 10.1007/s11104-006-9003-9.

Raynaud, X. and Nunan, N. (2014) 'Spatial ecology of bacteria at the microscale in soil', *PLoS ONE*, 9(1). doi: 10.1371/journal.pone.0087217.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015) 'High-Throughput Sequencing Technologies', *Molecular Cell*. Cell Press, pp. 586–597. doi: 10.1016/j.molcel.2015.05.004.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics and Bioinformatics*. Beijing Genomics Institute, pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Richardson, A. E. and Simpson, R. J. (2011) 'Soil microorganisms mediating phosphorus availability', *Plant Physiology*, 156(3), pp. 989–996. doi: 10.1104/pp.111.175448.

Robe, P. *et al.* (2003) 'Extraction of DNA from soil', *European Journal of Soil Biology*, 39(4), pp. 183–190. doi: 10.1016/S1164-5563(03)00033-5.

Rodríguez, H. *et al.* (2007) 'Genetics of phosphate solubilization and its potential applications for improving plant growth-promoting bacteria', in *First International Meeting on Microbial Phosphate Solubilization*. Springer Netherlands, pp. 15–21. doi: 10.1007/978-1-4020-5765-6_2.

Roh, S. W. *et al.* (2010) 'Comparing microarrays and next-generation sequencing technologies for microbial ecology research', *Trends in Biotechnology*. Elsevier Current Trends, pp. 291–299. doi: 10.1016/j.tibtech.2010.03.001.

Rothberg, J. M. *et al.* (2011) 'An integrated semiconductor device enabling non-optical genome sequencing', *Nature*, 475(7356), pp. 348–352. doi: 10.1038/nature10242.

Sanger, F., Nicklen, S. and Coulson, a R. (1977) 'DNA sequencing with chain-terminating inhibitors.', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–7. doi: 10.1073/pnas.74.12.5463.

Scherer, H. W. (2009) 'Sulfur in soils', Journal of Plant Nutrition and Soil Science. John Wiley & Sons, Ltd, pp. 326–335. doi: 10.1002/jpln.200900037.

Schmidt, M. W. I. *et al.* (2011) 'Persistence of soil organic matter as an ecosystem property', *Nature*, 478, pp. 49–56. doi: 10.1038/nature10386.

Schöps, R. *et al.* (2018) 'Land-use intensity rather than plant functional identity shapes bacterial and fungal rhizosphere communities', *Frontiers in Microbiology*, 9(NOV), p. 2711. doi: 10.3389/fmicb.2018.02711.

Schroth, M. N. and Hancock, J. G. (1982) 'Disease-Suppressive Soil and Root-Colonizing Bacteria', 216(June).

Sedlar, K., Kupkova, K. and Provaznik, I. (2017) 'Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics', *Computational and Structural Biotechnology Journal*. Elsevier B.V., pp. 48–55. doi: 10.1016/j.csbj.2016.11.005.

Sharma, S. B. *et al.* (2013) 'Phosphate solubilizing microbes: sustainable approach for managing phosphorus deficiency in agricultural soils', *SpringerPlus*, 2, p. 587. doi: 10.1186/2193-1801-2-587.

Shrivastava, M., Srivastava, P. C. and D'Souza, S. F. (2018) 'Phosphate-solubilizing microbes: Diversity and phosphates solubilization mechanism', in *Role of Rhizospheric Microbes in Soil: Volume 2: Nutrient Management and Crop Improvement*. Springer Singapore, pp. 137–165. doi: 10.1007/978-981-13-0044-8_5.

Stackebrandt, E. (2006) *Molecular identification, systematics, and population structure of prokaryotes*, *Molecular Identification, Systematics, and Population Structure of Prokaryotes*. Springer Berlin Heidelberg. doi: 10.1007/978-3-540-31292-5.

Staley, J. T. and Konopka, A. (1985) 'Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats.', Annual review of microbiology. doi: 10.1146/annurev.mi.39.100185.001541.

Steenwerth, K. L. *et al.* (2005) 'Response of microbial community composition and activity in agricultural and grassland soils after a simulated rainfall', *Soil Biology and Biochemistry*, 37(12), pp. 2249–2262. doi: 10.1016/j.soilbio.2005.02.038.

Steffan, R. J. *et al.* (1988) 'Recovery of DNA from soils and sediments.', *Applied and environmental microbiology*, 54(12), pp. 2908–2915. doi: 10.1128/aem.54.12.2908-2915.1988.

Stres, B. *et al.* (2010) 'Frequent freeze-thaw cycles yield diminished yet resistant and responsive microbial communities in two temperate soils: A laboratory experiment', *FEMS Microbiology Ecology*, 74(2), pp. 323–335. doi: 10.1111/j.1574-6941.2010.00951.x.

Suding, K. N. *et al.* (2008) 'Scaling environmental change through the community-level: A trait-based response-and-effect framework for plants', *Global Change Biology*, 14(5), pp. 1125–1140. doi: 10.1111/j.1365-2486.2008.01557.x.

Swerdlow, H. and Gesteland, R. (1990) 'Capillary gel electrophoresis for rapid, high resolution DNA sequencing', Nucleic Acids Research, 18(6), pp. 1415–1419. doi: 10.1093/nar/18.6.1415.

Tebbe, C. C. and Vahjen, W. (1993) 'Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast', *Applied and Environmental Microbiology*, 59(8), pp. 2657–2665. doi: 10.1128/aem.59.8.2657-2665.1993.

Techtmann, S. M. and Hazen, T. C. (2016) 'Metagenomic applications in environmental monitoring and bioremediation', *Journal of Industrial Microbiology and Biotechnology*, pp. 1345–1354. doi: 10.1007/s10295-016-1809-8.

Thomas, T., Gilbert, J. and Meyer, F. (2012) 'Metagenomics - a guide from sampling to data analysis', *Microbial Informatics and Experimentation*, 2(1), p. 3. doi: 10.1186/2042-5783-2-3.

Torres-Cortés, G. *et al.* (2011) 'Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples', *Environmental Microbiology*, 13(4), pp. 1101–1114. doi: 10.1111/j.1462-2920.2010.02422.x.

Touw, W. G. *et al.* (2012) 'Data mining in the Life Sciences with Random Forest : a walk in the park or lost in the jungle ?', 14(3). doi: 10.1093/bib/bbs034.

Victoria Wang, X. *et al.* (2012) 'Estimation of sequencing error rates in short reads', *BMC Bioinformatics*, 13(1), p. 185. doi: 10.1186/1471-2105-13-185.

Vinga, S. and Almeida, J. (2003) 'Alignment-free sequence comparison - A review', *Bioinformatics*, 19(4), pp. 513–523. doi: 10.1093/bioinformatics/btg005.

Vlek, P. L. G. *et al.* (2017) 'Trade-offs in multi-purpose land use under land degradation', *Sustainability (Switzerland)*. MDPI AG, p. 2196. doi: 10.3390/su9122196.

van der Walt, A. J. *et al.* (2017) 'Assembling metagenomes, one community at a time', *BMC Genomics*, 18(1), p. 521. doi: 10.1186/s12864-017-3918-9.

Wang, Y. *et al.* (2019) 'Global Distribution of Anaerobic Ammonia Oxidation (Anammox) Bacteria – Field Surveys in Wetland, Dryland, Groundwater Aquifer and Snow', *Frontiers in Microbiology*, 10, p. 2583. doi: 10.3389/fmicb.2019.02583.

Wang, Y., Yang, Q. and Wang, Z. (2014) 'The evolution of nanopore sequencing', *Frontiers in Genetics*, 5(DEC), p. 449. doi: 10.3389/fgene.2014.00449.

Warton, D. I., Wright, S. T. and Wang, Y. (2012) 'Distance-based multivariate analyses confound location and dispersion effects', *Methods in Ecology and Evolution*, 3(1), pp. 89–101. doi: 10.1111/j.2041-210X.2011.00127.x.

Wertz, S. *et al.* (2007) 'Decline of soil microbial diversity does not influence the resistance and resilience of key soil microbial functional groups following a model disturbance', *Environmental Microbiology*, 9(9), pp. 2211–2219. doi: 10.1111/j.1462-2920.2007.01335.x.

West, P. C. *et al.* (2010) 'Trading carbon for food: Global comparison of carbon stocks vs. crop yields on agricultural land', *Proceedings of the National Academy of Sciences of the United States of America*, 107(46), pp. 19645–19648. doi: 10.1073/pnas.1011078107.

Woappi, Y. (2013) 'Emergence of Antibiotic-Producing Microorganisms in Residential Versus Recreational Microenvironments', *British Microbiology Research Journal*, 3(3), pp. 280–294. doi: 10.9734/bmrj/2013/3205.

Wood, D. E. and Salzberg, S. L. (2014) 'Kraken: Ultrafast metagenomic sequence classification using exact alignments', *Genome Biology*, 15(3). doi: 10.1186/gb-2014-15-3-r46.

Wooley, J. C. and Ye, Y. (2010) 'Metagenomics : Facts and Artifacts , and Computational Challenges', 25(1), pp. 71–81.

Wu, Y.-W. *et al.* (2014) 'MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.', *Microbiome*, 2(1), p. 26. doi: 10.1186/2049-2618-2-26.

Wüst, P. K. *et al.* (2016) 'Estimates of soil bacterial ribosome content and diversity are significantly affected by the nucleic acid extraction method employed', *Applied and Environmental Microbiology*, 82(9), pp. 2595–2607. doi: 10.1128/AEM.00019-16.

Xie, K. *et al.* (2018) 'Biases in Prokaryotic community amplicon sequencing affected by DNA extraction methods in both saline and non-saline soil', *Frontiers in Microbiology*, 9(AUG), p. 1796. doi: 10.3389/fmicb.2018.01796.

Zehr, J. P. *et al.* (2003) 'Nitrogenase gene diversity and microbial community structure: A cross-system comparison', *Environmental Microbiology*, 5(7), pp. 539–554. doi: 10.1046/j.1462-2920.2003.00451.x.

Zelles, L. (1999) 'Fatty acid patterns of phospholipids and lipopolysaccharides in the characterisation of microbial communities in soil: A review', *Biology and Fertility of Soils*. Springer, pp. 111–129. doi: 10.1007/s003740050533.

Zhang, X. *et al.* (2015) 'Managing nitrogen for sustainable development', *Nature*. Nature Publishing Group, pp. 51–59. doi: 10.1038/nature15743.

Zhou, Q., Su, X. and Ning, K. (2014) 'Assessment of quality control approaches for metagenomic data analysis.', *Scientific reports*, 4(4), p. 6957. doi: 10.1038/srep06957.

Zielezinski, A. *et al.* (2017) 'Alignment-free sequence comparison: Benefits, applications, and tools', *Genome Biology*, 18(1), pp. 1–17. doi: 10.1186/s13059-017-1319-7.

Zielińska, S. *et al.* (2017) 'The choice of the DNA extraction method may influence the outcome of the soil microbial community structure analysis', *MicrobiologyOpen*, 6(4), p. e00453. doi: 10.1002/mbo3.453.

Zumft, W. G. (1997) 'Cell biology and molecular basis of denitrification.', *Microbiology and molecular biology reviews : MMBR*, 61(4), pp. 533–616. doi: 10.1128/.61.4.533-616.1997.

Chapter 2

Beyond taxonomic identification: integration of ecological responses to a soil bacterial 16S rRNA gene database.

Abstract

High-throughput sequencing 16S rRNA gene surveys have enabled new insights into the diversity of soil bacteria, and furthered understanding of the ecological drivers of abundances across landscapes. However, current analytical approaches are of limited use in formalising syntheses of the ecological attributes of taxa discovered, because derived taxonomic units are typically unique to individual studies and sequence identification databases only characterise taxonomy. To address this, we used sequences obtained from a large nationwide soil survey (GB Countryside Survey, henceforth "CS") to create a comprehensive soil specific 16S reference database, with coupled ecological information derived from the survey metadata. Specifically, we modelled taxon responses to soil pH at the OTU level using hierarchical logistic regression (HOF) models, to provide information on putative landscape scale pH-abundance responses. We identify that most of the soil OTUs examined exhibit predictable abundance responses across soil pH gradients, though with the exception of known acidophilic lineages, the pH optima of OTU relative abundance was variable and could not be generalised by broad taxonomy. This highlights the need for tools and databases to predict ecological traits at finer taxonomic resolution. We further demonstrate the utility of the database by testing against geographically dispersed query 16S datasets; evaluating efficacy by quantifying matches, and accuracy in predicting pH responses of query sequences from a separate large soil survey. We found that the CS database provided good coverage of dominant taxa; and that the taxa indicative of soil pH in a query dataset corresponded with the pH classifications of top matches in the CS database. Furthermore, we were able to predict query dataset community structure, using predicted abundances of dominant taxa based on query soil pH data and the HOF models of matched CS database taxa. The database with associated HOF model outputs is released as an online portal for querying single sequences of interest (https://shiny-apps.ceh.ac.uk/ID-TaxER), and as a dada2 database for use in bioinformatics pipelines. The further development of advanced informatics infrastructures incorporating modelled ecological attributes along with new functional genomic information will likely facilitate large scale exploration and prediction of soil microbial functional biodiversity under current and future environmental change scenarios.

2.1 Introduction

Soil bacteria are highly diverse (Gans, Wolinsky, & Dunbar, 2005; Roesch et al., 2010) and are significant contributors to soil functionality. Sequencing of 16S rRNA genes has enabled a wealth of new insights into the taxonomic diversity of soil prokaryotic communities, revealing the ecological controls on a vast diversity of yet to be cultured taxa with unknown functional potential (Fierer, 2017). However, despite thousands of studies across the globe, we are still some way from synthesising the new knowledge on the ecology of these novel organisms recovered through local and distributed soil surveillance. This is because there is currently no formalised way of retrieving ecological information on reference sequences which match user discovered taxa (either clustered operational taxonomic units or amplicon sequence variants). Whilst we have a wealth of databases and tools for characterising the taxonomy of matched sequences (McDonald et al., 2012; Quast et al., 2013; Wang, Garrity, Tiedje, & Cole, 2007), databases do not include any associated ecological information on sequences matches. Whilst new software has recently become available that uses text mining to return some ecological data on matched sequences to NCBI, this information is currently limited to descriptions of sequence associated habitat (Sinclair et al., 2016).

Synthesising relationships between soil amplicon abundances and environmental parameters is now necessary to progress ecological understanding of soil microbes beyond those few organisms that are readily cultivated. Determining microbial responses across environmental gradients can inform on the realised niche widths of discrete taxa, and may indicate the presence of shared functional traits across taxa (Martiny et al., 2015). This information is now urgently needed for microbes as we move into a period of increasing genomic data availability for uncultivated taxa. Coupling data on taxon responses across environmental gradients with functional trait information potentially allows a mechanistic and predictive understanding of both biodiversity and ecosystem level responses to environmental change. For example, a large body of theory exists describing how species responses to environmental change affects ecosystem functioning (Diaz et al., 2013; Lavorel & Garnier, 2002; Suding et al., 2008). Here functional “response” groups are defined as species sharing a similar response to an environmental driver; and functional “effect” groups refer to species that have similar effects on one or more ecosystem processes. The

degree of coupling between response and effect groups can then allow prediction of functional effects under change. For instance, if certain phylogenetic groups of taxa decrease due to environmental change, and these taxa also represent an effect group (e.g. these taxa possess a unique functional gene) then we can expect the function to also decrease. Conversely with uncoupled effect groups (e.g. responsive taxa all possess a ubiquitous functional gene), the system is likely to be more functionally resistant to change (Diaz et al., 2013). Applying such concepts to microbial ecology is a realistic ambition given the extensive availability of amplicon datasets coupled to environmental information, and the increasing feasibility of uncultivated microbial genome assembly from metagenomes or single cell genomics (Choi et al., 2017).

The fast evolution of microbial taxa coupled with potential horizontal gene transfer has led to assumptions that microbial diversity may be largely functionally redundant (Martiny et al., 2006). However we know from large-scale amplicon surveys that there are distinct differences in soil bacterial composition across environmental gradients, with soil pH frequently observed as a primary correlate (Fierer & Jackson, 2006; Griffiths et al., 2011). This implies that different microbial phylogenetic lineages possess adaptations conferring altered competitiveness in soils of different pH; paving the way for future studies into the genomic basis, and thereby elucidating specific genetic "response traits". There is also evidence that specific microbial functional capacities are less common e.g. pesticide degradation (Griffiths and Philippot, 2013; Jia and Whalen, 2020). Determining the degree of functional redundancy in taxa which respond across soil pH gradients, will permit new insight into the microbial biodiversity mechanisms underpinning soil functionality and resilience to change. Since soil pH is largely predictable from geo-climatic (Slessarev et al., 2016) and land use features (Wamelink et al., 2019); prediction of the abundances of individual bacterial taxa under environmental change scenarios is likely to be feasible. The immediate challenge is therefore to establish predictive frameworks for many soil bacterial taxa, which can be populated with genomic information as it becomes available; to ultimately facilitate predictions of microbial functional distributions.

We believe that attempts to progress understanding of the ecological attributes of environmentally retrieved bacterial taxa can be streamlined immediately by making better

use of the extensive amplicon datasets that exist, which already provide much useful information on taxa-environment responses. Indeed it has recently been shown that many prokaryotic taxa are distributed globally (particularly dominant OTUs (Delgado-Baquerizo et al., 2018)), yet there is currently no way to formally capture their ecological attributes in databases for further microbiological and ecological enquiry other than in supplementary material spreadsheets. Here we seek to address this by making available a database of representative sequences from a large 16S rRNA amplicon dataset from over 1000 soil samples collected across Britain. In addition to providing standard taxonomic annotation, we also seek to add ecological response information to each representative sequence. We focus here on soil pH responses as bacterial communities are known to respond strongly across soil pH gradients (Griffiths et al., 2011).

We will firstly model OTU abundances in response to soil pH using hierarchical logistic regression (HOF) (Jansen & Oksanen, 2013), a commonly used approach to examine vegetation responses across ecological gradients which has yet to be widely applied to microbial datasets. We will use model outputs to assign each OTU to a specific pH response group based on abundance optima, and in addition demonstrate the utility of the database in determining the phylogenetic relationships in ecological responses. The utility of the database will be further tested on 16S datasets to compare both the percentage of hits and modelled responses. The OTU database with associated HOF model outputs is released both as an online portal for visualising individual queries and as flat files for integration into existing bioinformatics pipelines.

2.2 Methods

Samples were collected as part of the UK Centre for Ecology and Hydrology Countryside survey (CS) between June and July 2007 covering sites throughout Great Britain. Samples were chosen through a stratified random sample of 1 km squares using a 15 km grid, implementing the institute of Terrestrial Ecology (ITE) land classification to ensure incorporation of different land classes, with up to 5 randomly sampled cores taken within each square. Surface litter was removed from soil cores. Metadata for each soil sample were collated including soil organic matter, soil organic carbon, bulk density, pH, indicator of phosphorus availability using methodologies detailed elsewhere (Griffiths et al., 2011;

Reynolds et al., 2013). Soil cores were homogenised wet without sieving prior to subsampling for DNA extraction.

DNA was extracted from 0.3g of soil using the MoBIO PowerSoil-htp 96 Well DNA Isolation kit (Carlsbad, CA) according to manufacturer protocols. Amplicon libraries were constructed according to the dual indexing strategy of Kozich et al (Kozich *et al*.,2013), using primers 341F (Muyzer, de Waal, & Uitterlinden, 1993) and 806R (Caporaso *et al*., 2011). Amplicons were generated using a high fidelity DNA polymerase (Q5 Taq, New England Biolabs) on 20 ng of template DNA employing an initial denaturation of 30 seconds at 95 ºC, followed by (25 for 16S and 30 cycles for ITS and 18S) of 30 seconds at 95 ºC, 30 seconds at 52 ºC and 2 minutes at 72 ºC. A final extension of 10 minutes at 72 ºC was also included to complete the reaction. Amplicon sizes were determined using an Agilent 2200 TapeStation system (~550bp) and libraries normalized using SequalPrep Normalization Plate Kit (Thermo Fisher Scientific). Library concentration was calculated using a SYBR green quantitative PCR (qPCR) assay with primers specific to the Illumina adapters (Kappa, Anachem). Libraries were sequenced at a concentration of 5.4 pM with a 0.6 pM addition of an Illumina generated PhiX control library. Sequencing runs, generating 2 x 300 bp reads were performed on an Illumina MiSeq using V3 chemistry.

Sequenced paired-end reads were joined using PEAR (Zhang *et al*., 2013), quality filtered using FASTX tools (hannonlab.cshl.edu), length filtered with the minimum length of 300bp. The presence of PhiX and adapters were checked and removed with BBTools (jgi.doe.gov/data-and-tools/bbtools/), and chimeras were identified and removed with VSEARCH_UCHIME_REF (Rognes *et al.*, 2016) using Greengenes Release 13_5 (at 97%). Singletons were removed and the resulting sequences were clustered into operational taxonomic units (OTUs) with VSEARCH_CLUSTER at 97% sequence identity. Representative sequences for each OTU were taxonomically assigned by RDP Classifier with the bootstrap threshold of 0.8 or greater using the Greengenes Release 13_5 (full) as the reference. Taxonomic groupings will be referred to as those assigned by the Greengenes release used, though we acknowledge these names may vary to those used in the Genome Taxonomy Database (GTDB) (Parks et al., 2018, 2020).

All statistical analyses and visualisations were conducted within the R package, predominantly using the vegan (Oksanen *et al.*, 2018) and ggplot2 (Wickham, 2016) packages unless otherwise indicated.

2.3 Results and discussion

2.3.1 Database Coverage

The database was constructed from sequences obtained from the 2007 Countryside Survey (CS), a random stratified sampling of most soil types and habitats across Great Britain, full details of which are provided elsewhere (Griffiths et al., 2011; Reynolds et al., 2013). Sequencing of 1113 soils using the universal 341f/806r (Takahashi *et al.*, 2014) primers targeting the V3 and V4 regions of the 16S rRNA gene yielded a total of 39952 reference sequence OTUs, after clustering at 97% sequence similarity and singleton removal. Coverage was assessed on a filtered dataset of 1006 samples which had at least 5000 reads per sample, using sample based species accumulation curves calculated per habitat class and pooled across all habitats (**Fig.2.1**). The curves for individual habitats, whilst not reaching saturation, reveal some interesting trends with grasslands exhibiting highest biodiversity at the landscape scale, which is likely attributable to the broad range of soil conditions they encompass. The pooled curves across all habitats however appear to begin to level off, which importantly reveals that in total the reference sequence dataset provides good coverage of the non-singleton 97% OTUs found across this landscape.

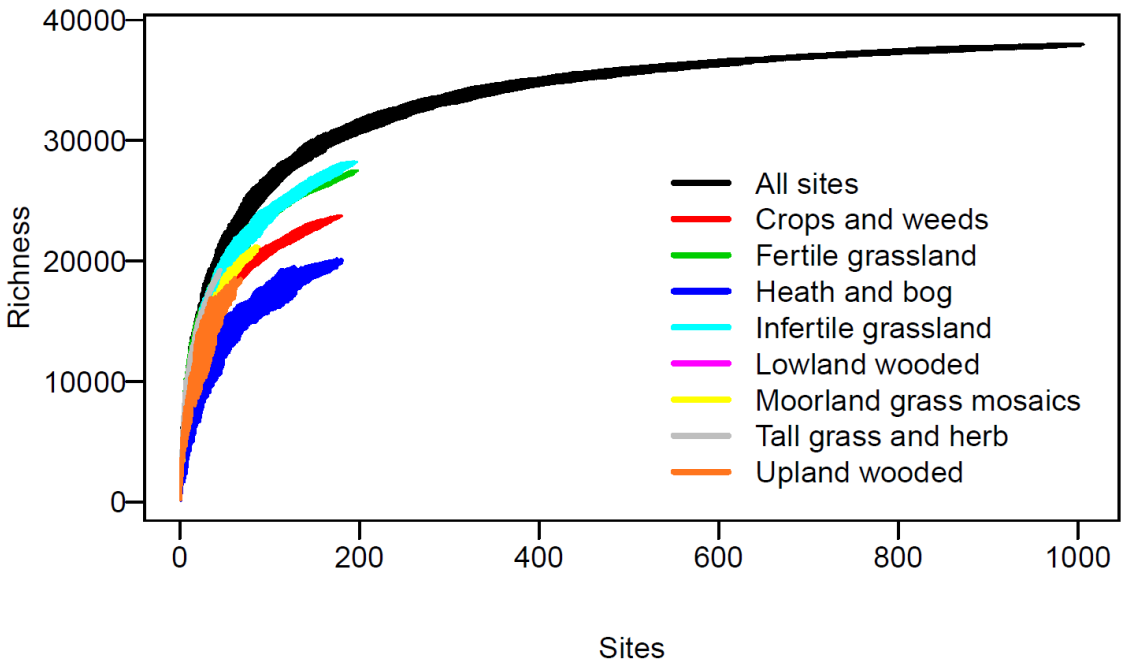


Fig.2.1. Coverage of bacterial 97% OTUs within the Countryside Survey (CS) dataset. Sample based richness accumulation curves were calculated across 1006 CS soil samples ("All sites") and within specific habitats. Standard deviations are calculated from random permutations of the data.

2.3.2 Performance of database against independent datasets

The coverage of this dataset was further assessed through blasting representative sequences from independent 16S datasets from various locations and habitats, against all 39952 CS representative sequences (**Table.2.1**). Here we defined an OTU 'hit' as a query OTU that shared 97% identity with a CS OTU and had an e value <0.001. We subsequently calculated the percentage of OTUs within the independent dataset meeting this criteria to gain insights into coverage.

For the two soil datasets, we found over 50% of the OTUs in each study had a hit within the CS database. Expectedly, this was in stark contrast to a fresh water dataset which exhibited much less overlap with the CS database with 33.2% having CS hits. 16S sequences from dataset 1 (**Table.2.1**), a study of land use change across the UK (Malik et al., 2018), also sequenced with the same 341f/806r primer set, had the highest percentage of hits against the CS representative sequences (67.26%). Wider assessment of our own unpublished datasets using the exact same methodologies yield percentages of hits of 62% and 56% for soils from UK calcareous grasslands and tropical rainforests respectively. A separate survey

of Welsh soils (George et al., 2019) was also queried against the CS database, which used the commonly used Earth Microbiome primer set exclusively targeting the V4 region (as opposed to V3 and V4 targeted region used for the CS dataset). This dataset had a percentage of hits of 58.49% providing evidence that datasets amplified with other primer sets can be matched to the CS database with only marginal loss of coverage.

Query Dataset	Habitat Description	Query out percentage of hits	Primer	Citation
1	Grassland and arable soils, Britain	67.26%	341f/806r V3-V4	Malik *et al.*, 2018
2	All habitat soils survey, Wales	58.49%	515f/806rB V4	George *et al.*, 2019
3	Thames River, Britain	33.2%	341f/806r V3-V4	Unpublished but see Read et al, 2015

Table.2.1. Validating the use of the CS OTU sequences as a database, through querying with independent datasets. Reference sequences from independent datasets were BLAST searched against countryside survey representative sequences, and the proportion of OTUs matched at over 97% similarity reported. British soil query datasets had highest percentage of hits irrespective of methodologies, with a set of riverine samples showing lowest proportion of OTUs matching the CS soil reference database.

We next wanted to explore possible reasons for obtaining less than 100% coverage from query soil datasets, given the good coverage of the CS reference sequence database evident from the rarefaction curve (**Fig.2.1**). We predicted this discrepancy was caused by rare OTUs being unique to specific studies and tested this by classifying the UGRASS OTUs (Query dataset 1) into 1000 discrete abundance based quantiles (1 being the most abundant quantile and 1000 being the least). Plotting the proportion of query OTUs which matched to the CS database by query OTU abundance class, confirmed that less abundant query OTUs had less matches to the CS database (**Fig.2.2**). This adds weight to arguments that much of the rare taxa detected through amplicon sequencing could be spurious artefacts of the PCR amplification process (Edgar, 2017). Regardless of these issues, the high proportion of hits for dominant taxa in the query dataset validates the use of the large CS dataset as a comprehensive reference database.

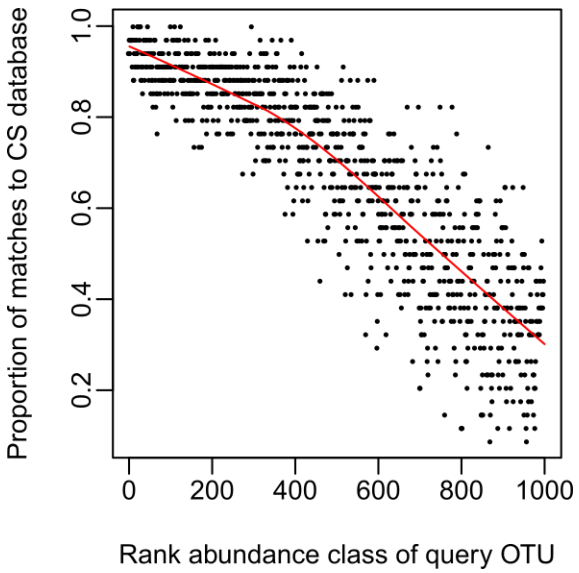


Fig.2.2. The CS database provides good coverage of dominant taxa within a query dataset. Query OTU reference sequences (dataset 1, **Table.2.1**) were grouped into 1000 bins by decreasing rank (e.g. the 1000th bin contains the least abundant OTUs); and the proportion of each bin matching the CS dataset calculated and displayed on the y axis. The proportion of matches to the CS database (> 97% similarity) declines as query taxa become rarer, despite the comprehensive nature of the CS database.

2.3.3 Modelling OTU responses to soil pH

Since the majority of the 39952 reference OTUs obtained across all CS samples likely derive from rare taxa with intrinsically little value for predictive modelling (low within-sample abundance, and occurrence across samples), we opted to only model taxa-pH relationships for those taxa which occurred in at least 30 samples. These taxa were selected from a cleaned dataset of 1006 samples which had at least 5000 reads per sample. Further examination of the species accumulation by sample curves for the resulting 13781 OTUs, revealed saturation implying that this dataset had complete coverage of common OTUs, defined by being present in at least 30 samples across Britain. Huisman-Olff-Fresco models were then applied to determine individual bacterial taxa responses to pH using the R package eHOF using a poisson error distribution (Gao et al., 2017 ; Jansen & Oksanen, 2013). Model choice was determined using AIC and bootstrapping methods implemented with the eHOF package (Jansen & Oksanen, 2013), whereby the model with the lowest AIC was initially chosen and its robustness then tested by rerunning models on 100 bootstrapped datasets (created by resampling with replacement). If the most frequently chosen model in the bootstrap runs was different to the initial model choice, the most common bootstrap choice was selected. The resultant pH-taxa response curves classified by the HOF models

include I: no significant change in abundance in response to pH, II: an increasing or decreasing trend, III: increasing or decreasing trend which plateaus, IV: Increase and decrease by same rate (unimodal) and V: Increase and decrease by different rates causing skew (**Fig.2.3**).

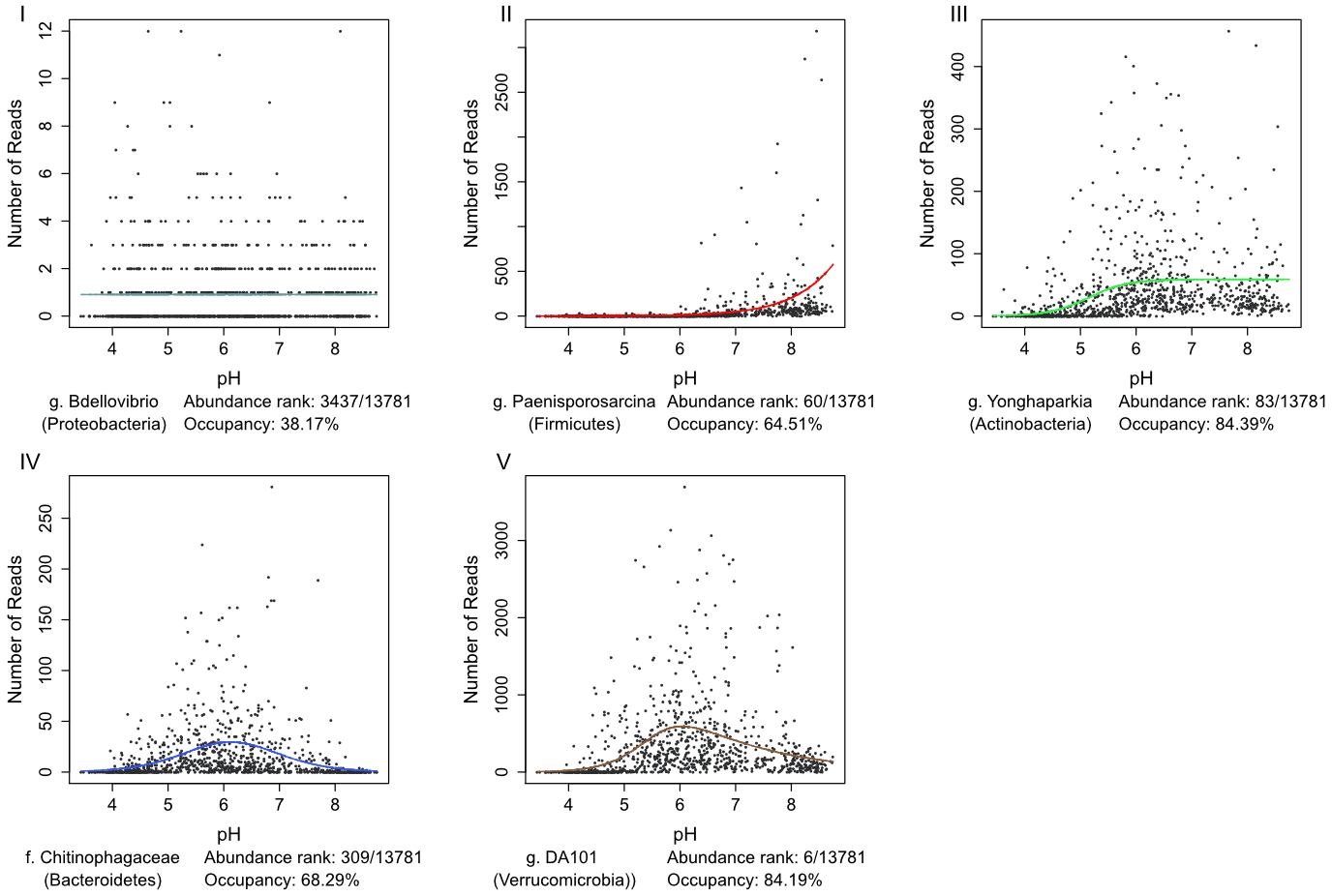


Fig.2.3. Examples of the five HOF model types. HOF models were generated through fitting countryside survey OTU abundances to soil pH (a pH range from 3.63 to 8.75). The five HOF models used were: I: no change in abundance across pH gradient, II: montonic an increase or decrease in abundance along pH gradient, III: plateau an increase or decrease in abundance along pH gradient that plateaus, IV: symmetrical unimodal, abundance increases and decreases across gradient at an equal rate, V: skewed unimodal, abundance increases and decreases across gradient at unequal rates. Abundance rank (out of all 13781 taxa modelled, 1 being the most abundant and 13781 being the least) and occupancy (percentage of samples taxon is found in) are shown for each example model taxon.

The proportion of OTUs assigned to each model is shown in **Table.2.2** and reveals that most of the soil OTUs exhibited some trend with soil pH, and with the unimodal skewed model (V) being the most commonly fitted model type (45.76%). OTUs were then assigned to pH response groups based on the fitted pH optima. We classified OTUs demonstrating an acidic preference if the fitted optima was below pH 5.2, based on previous data showing this represented a critical threshold for bacterial communities (Griffiths et al., 2011), which was further confirmed by a similar regression tree analyses of this sequence dataset (not shown). This pH value also represents a critical threshold in microbial functioning (Jones *et al.,* 2019). Similarly, a second threshold was designated at pH 7, with OTUs exhibiting an optima above this being classed as neutral, and those between 5.2 and 7 classed as "mid". Plateau model shapes (model III), were sometimes more difficult to classify, since two optima are provided which span the plateau, and in some cases these crossed the pH 5.2 and 7 thresholds. Whilst OTUs exhibiting this response were in the minority, we opted to assign a separate designation representing this range, for instance "acid to mid" for an OTU with two optima above and below pH 5.2. The proportion of taxa classified to each pH response group are shown in **Table.2.3**. This reveals that OTUs with acidic preference are in the minority, consistent with reduced bacterial biodiversity being frequently observed in acidic soils (Griffiths et al., 2011).

Model fit	Percentage of Countryside survey OTUs
V (Skewed Unimodal)	45.76%
III (Plateau)	24.13%
IV (Unimodal)	23.52%
II (Monotonic)	6.11%
I (No trend)	0.49%

Table.2.2. Percentage of 13781 CS OTUs fitted to each HOF model. Each OTU was classified to one of five HOF model types according to fitted relationships with soil pH. The different model response shapes are shown in **Fig.2.3.**

pH Response group	Percentage of Countryside survey OTUs
Mid (5.2 < Optima < 7)	34.8%
Neutral (Optima > 7)	31.62%
Acid (Optima < 5.2)	23.08%
Mid to Neutral (5.2 < Optimum1 < 7 and Optimum 2 > 7)	7.41%
Acid to Neutral (Optimum1 <5.2 and Optimum2 >7)	1.52 %
Acid to Mid (Optimum1 <5.2 and 5.2 < Optimum2 < 7)	1.14%

Table.2.3. Percentage of 13781 CS OTUs classified to different pH response groups. Each OTU was assigned to a pH response classification based on the modelled pH optima. The model outputs with one optima (II, IV, V) were classified as acidic, mid, or neutral based on pH thresholds identified above. Plateau shaped models with 2 optima (model III), which spanned the pH thresholds were labelled as either mid to neutral, acid to neutral, or acid to mid.

Representative sequences of all 13781 OTUs were aligned with Clustal Omega 1.2.1 (http://www.clustal.org/), and used to construct a Phylogenetic tree with FastTree 2.1.7 (Price, Dehal, & Arkin, 2010) using neighbour-joining (NJ) with the generalized time-reversible (GTR) model of nucleotide evolution. The tree is shown in **Fig.2.4** together with the pH classification derived from the HOF models. Distinct phylogenetic clustering is apparent for phyla with representatives known to have acidophilic preferences such as the *Acidobacteria* (Martiny et al., 2006). Additionally, other phyla such as the *Verrucomicrobia* appear to possess clades with a distinct pH preference. However, the overall impression across other taxonomic groups is that the pH abundance optima can vary substantially amongst closely related taxa. This emphasises the need to move beyond the association of traits with broad phylogenetic lineages; and identifies the need to determine traits at finer levels of taxonomic resolution.

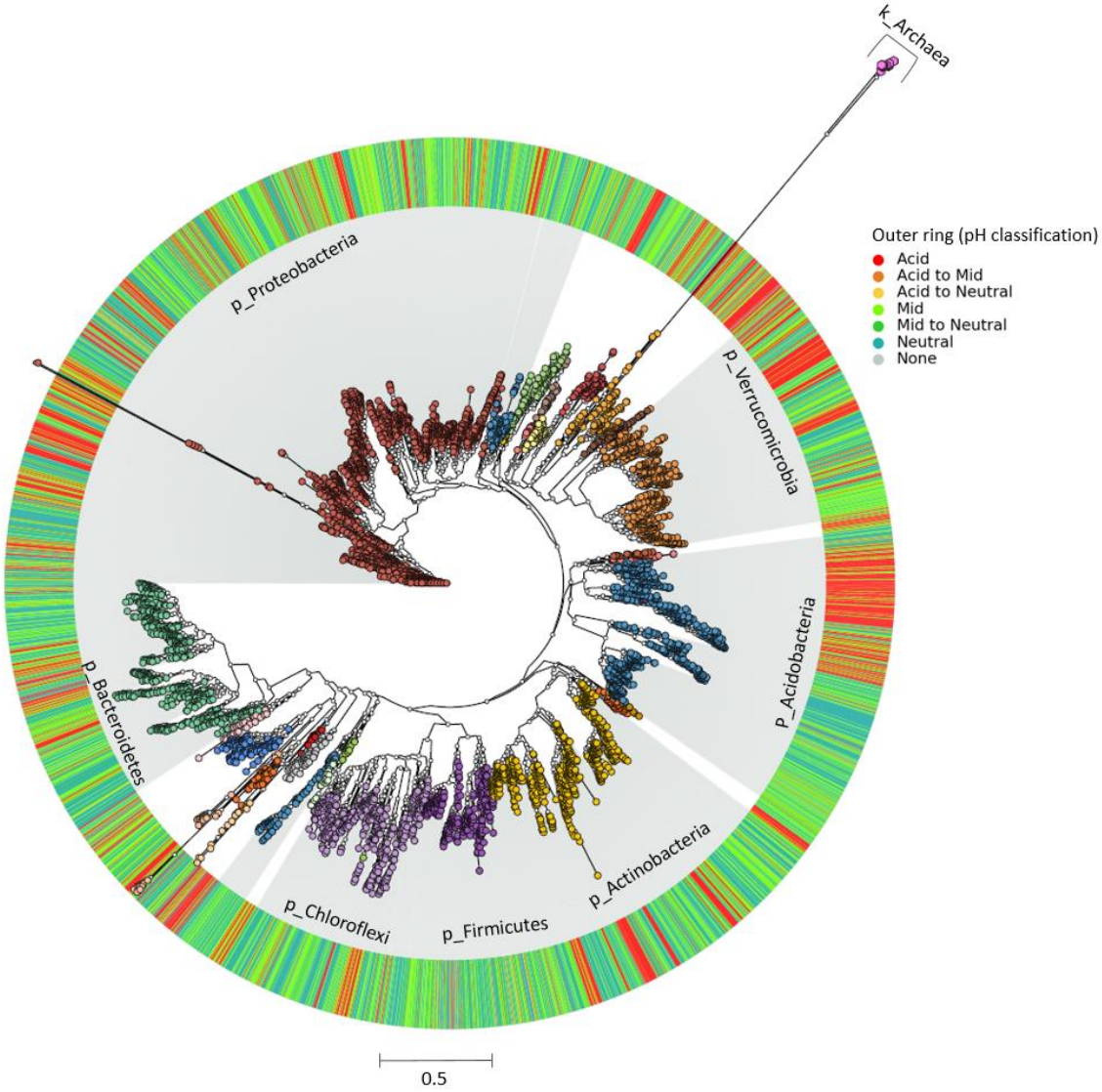


Fig.2.4. The phylogenetic distribution of bacterial pH optima. A phylogenetic tree of all OTUs present in >100 samples (totalling 6385 OTUs), with each OTU annotated according to pH classification based on HOF model optima (outer ring).

2.3.4 Incorporating CS data and pH responses into a sequence identification tool

A web application was developed using the Shiny package (https://shiny.rstudio.com/) which enables users to BLAST a 16S query sequence against the countryside survey representative sequences, subsequently allowing visualization of key environmental information including HOF model outputs, relevant to individual matched sequences. The Graphic User Interface was implemented in R (3.4.1) using the Shiny package (https://shiny.rstudio.com/) alongside ShinyJS to execute JavaScript functions from R (https://cran.r-project.org/web/packages/shinyjs/). BLASTn commands are executed from R using the users query sequence, e value of 0.01, and the reference sequence database of CS representative sequences. eHOF model objects were converted to binary using the Rbase serialize function and stored in a PostgreSQL (9.3.17) database (https://www.postgresql.org/) alongside model and other environmental metadata (**Fig.2.5**). BLAST results are displayed as an interactive table of hits, each hit linking to a plot of the pH model fit (based upon raw read number), a LOESS fit (based on relative abundance), a box plot of habitat associations and a simple interpolated map showing relative abundance distribution across Britain (**Fig.2.6**). Additionally, we provide a text box which can be populated with user submitted trait related information on matched OTUs. The application is available at shiny-apps.ceh.ac.uk/ID-TaxER/ and to facilitate batch processing of query sequences the sequence database, taxonomy and trait matrix are released via github (github.com/brijon/ID-TaxER-flat-files) for integration into bioinformatics pipelines.

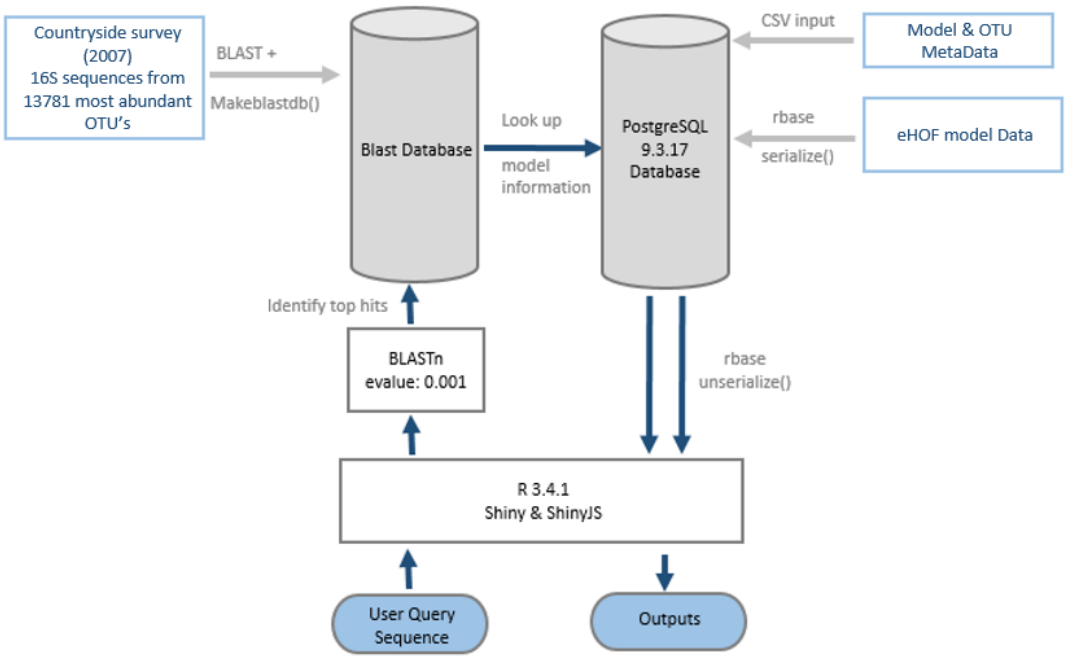


Fig. 2.5. ID-TaxER database Infrastructure 16S sequences are queried over the web via the R Shiny interface. A BLAST search is then performed against a blast database containing representative 16S sequences from the 2007 Countryside survey. Model information and associated metadata for match hits are located in a PostgreSQL database of OTU taxonomy/ model data, (model objects are stored as binary and retrieved for the user) and results displayed via the shiny interface.

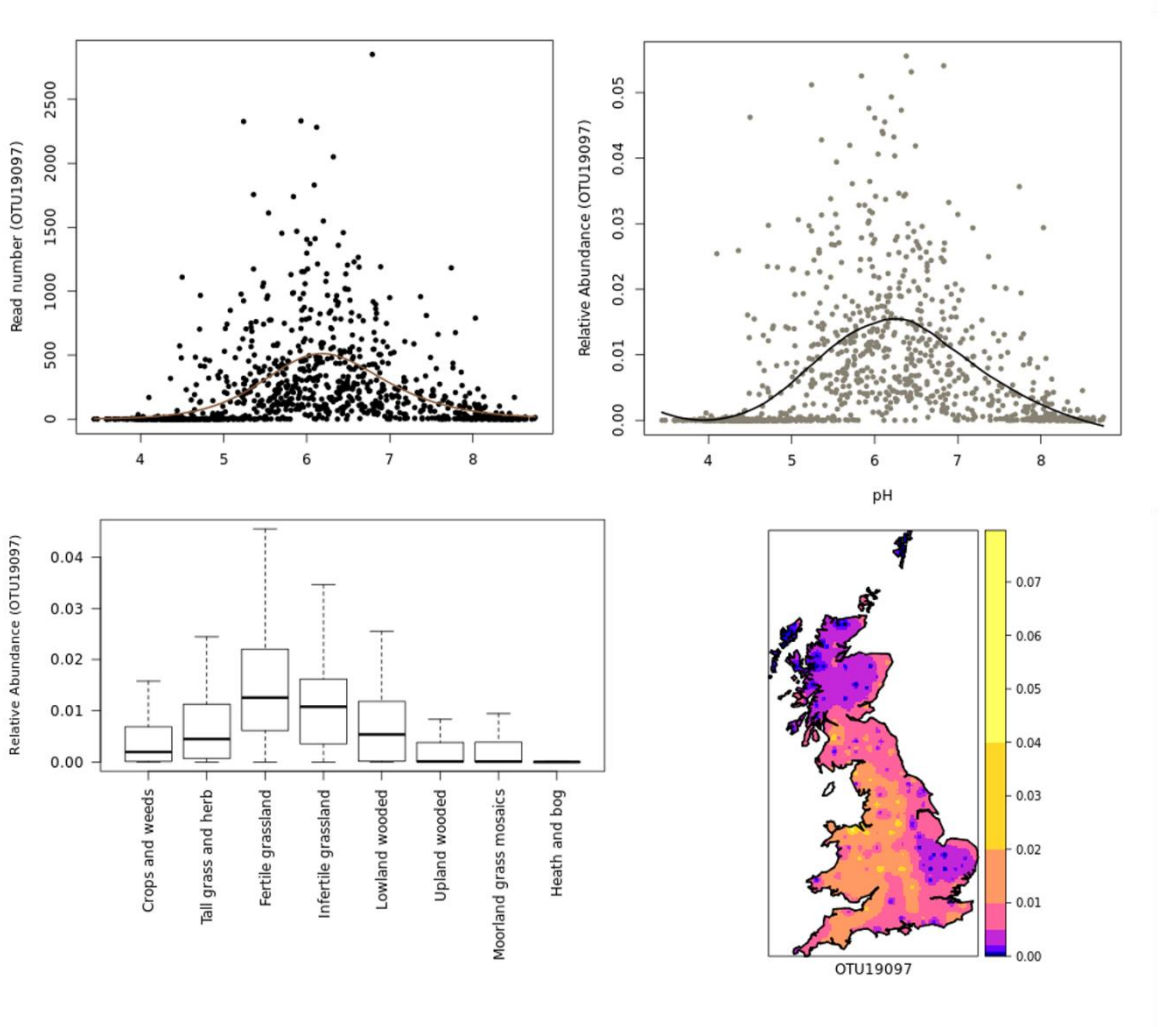


Fig.2.6. Example outputs from the ID-TaxER online portal. Using the *DA101 /Ca. U. copiosus* (Brewer, et al., 2016) 16S sequence (GenBank: Y07576.1) as a query, we found 98.3% identity to CS OTU19097 taxonomy=*k_Bacteria; p_Verrucomicrobia; c_Spartobacteria; o_Chthoniobacterales; f_Chthoniobacteraceae; g_DA101*): a) HOF model output showing the number of reads of CS OTU19097 per sample plotted against soil pH; with the line representing the model fit (Model V, unimodal response to pH with an optima at pH 6.18) b) the relative abundance of OTU19097 against sample pH, with the line representing a LOESS fit; c) boxplot showing the median and ranges of the relative abundance of OTU19097 per CS habitat class; d) inverse distance weighted interpolation map of the relative abundance of OTU19097 across Britain.

2.3.5 Utility in predicting pH preferences and community structure using a query dataset

To demonstrate both the utility of the reference sequence database, and the HOF modelling approach to identify environmental responses of soil bacterial taxa, we used a query dataset of >400 samples collected across Britain (dataset 1, **Table.2.1**). Since this survey focussed on productive habitats (grassland and arable land uses), with only a few acidic samples, it was not appropriate to generate independent HOF models. Instead we classified the samples according to the same pH cut-off levels identified above (pH 5.2 and 7) and then determined pH responsive taxa using Indicator species analyses (Dufrene & Legendre, 1997). As can be seen in **Fig.2.7a**, the pH groupings were clearly evident in the sample based ordination. Representative sequences from this dataset were then blasted against the CS database, and optimum pH and pH classification metrics retrieved from the top hit for subsequent comparison. In total 477 indicators for the three pH groupings were retrieved, of which 454 had a match greater than 97% similarity to the CS database. Of the 155 acidic indicator taxa identified in the query dataset, 129 (83%) were reliably classified as acidic OTUs based on matches to the CS database (**Fig.2.7b**), with 20 OTUs "incorrectly" classified as having a mid-pH optima. However, the predicted optima of these OTUs were mainly below pH 6 and most lie very close to pH 5.2. Similarly, for the 226 query taxa identified as indicating neutral soils, 203 (90%) had a neutral pH classification in the CS database, with 15 being incorrectly classed as mid, though the optima for these taxa were between pH 6.5 and 7. Sixty-seven indicators of the query mid pH soils were obtained of which 64 (96%) had a mid pH classification based on match to the CS database. Overall, this analysis shows that information on soil pH preferences from independent datasets can be reliably obtained using our approach.

We then sought to test whether we could reliably predict community structure using the CS HOF model outputs to predict query OTU abundances. We identified the most abundant OTUs in the query dataset and blasted these taxa against the CS database. CS HOF models were then used to predict the abundances of the 100 matched dominant OTUs within the 424 query samples. This predicted community matrix was then subject to NMDS ordination with the first axis scores plotted against the actual observed ordination scores generated

from 24260 OTUs. The results in **Fig.2.7c** show that the observed and predicted first axis ordination scores were highly related (r^2 = 0.88) demonstrating that it is possible to predict broad scale community change from individual OTU relative abundance pH models. These findings add to a growing body of literature on the predictability of soil bacterial communities (Bickel, Chen, Papritz, & Or, 2019; Fierer et al., 2013; Griffiths et al., 2016); but furthermore demonstrate the utility of our overall approach in deriving meaningful ecological information from matches to a 16S rRNA sequence database incorporating ecological responses.

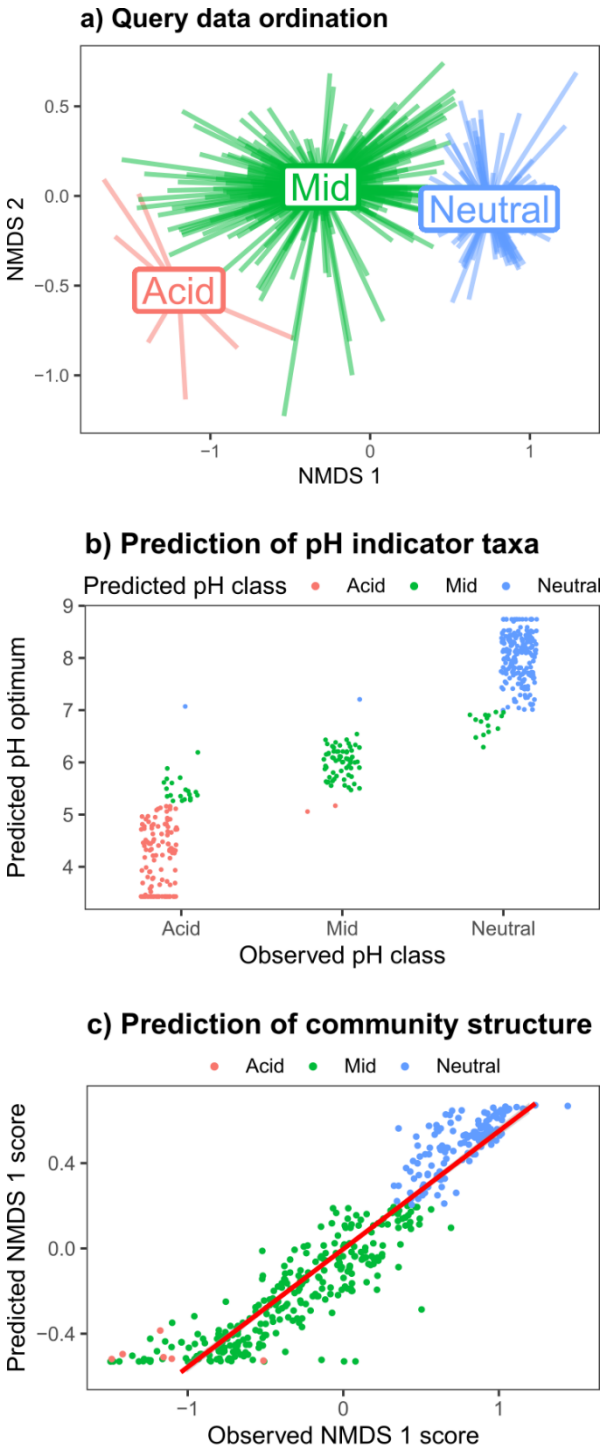


Fig. 2.7. Validating the pH models using a query dataset. Taxa strongly responsive to soil pH were identified from Query dataset 1 (**Table.2.1**) and then matched to the CS database to evaluate utility of the approach. **a)** NMDS ordination plot of the query dataset, with pH groupings denoted by colour (red =pH<5.2; green=pH>5.2<7; and blue=ph>7). **b)** Indicator species analyses on the query dataset revealed 477 OTUs strongly associated with the three pH classes ("Observed pH class"). The y axis *values,* and point colour denote the predicted pH optimum, and predicted pH class following matching to CS database. **c)** The relative abundances of the 100 most abundant taxa in the query dataset were predicted using the CS HOF models of matched *taxa and* subjected to NMDS ordination. The plot shows that the predicted abundances of these taxa reliably predicted the observed data first axis NMDS scores.

2.4 Conclusions

This work demonstrates how large scale soil molecular survey data can be used to build robust predictive models of bacterial abundance responses across environmental gradients. The models were applied to the single soil variable of pH which is known globally to be the strongest predictor of soil bacterial community structure in surveys spanning wide environmental gradients. We have produced an informatics tool incorporating extensive sequence data from a wide range of soils, linked to taxonomic and ecological response information. This currently includes data on the modelled pH optima, and the predictive utility in this regard was demonstrated using an independent dataset. Other ecological information is also made available via an online portal including habitat association, spatial distribution, and metrics relating to abundance and occurrence. We are currently working on incorporating other information on the sensitivities of discrete OTUs to land use change; and there is the wider potential for users to update the trait matrix with other observations (more information provided at https://github.com/brijon/ID-TaxER-flat-files). Such information could include sensitivities to perturbations such as climate change, as well as rRNA derived links to wider genome data to inform on function.

We anticipate this simple database and tool will be of use to the soil molecular community, but also hope it prompts further global efforts to better capture relevant ecological information on newly discovered microbial taxa. We acknowledge some limitations of the current tool, and identify some possibilities to develop further: Firstly being a 16S rRNA amplicon dataset, the database inventory will be affected by known biases relating to PCR primers and amplification conditions (Thijs et al., 2017); and obviously, user datasets built on a different regions of the 16S rRNA gene will not produce any matches. Additionally the length of sequences means only limited taxonomic resolution is currently provided, and ecological inferences based on BLAST matches must consider the strength of match, and variance within the matched region with respect to taxonomic discrimination (Fox, Wisotzkey, & Jurtshuk, 1992). Emerging long read sequencing technologies applied to survey nucleic acid archives in the future may improve these current constraints (Singer et al., 2016). With respect to the pH models, many other factors can of course influence bacterial abundances (Fierer, 2017; Thomson et al., 2010) and we note the large degree of

variance in relative abundance for a taxon even within its apparent pH niche optima (**Fig.2.3**). Such variance could be caused by nutrient availability, stress etc. and more complex models, albeit constrained by pH, need to be formulated to advance predictive accuracy. More generally, we assert that observed taxon relative abundance only inform on relative taxon success at a given soil pH, and does not identify any explicit underpinning ecological mechanism (e.g. pH stress tolerance versus competitive fitness) (Austin, 1999). However, linking emerging genomic data to detailed environmentally relevant sequence databases such as detailed here, will likely improve future understanding in relation to elucidating specific functional response traits and determining mechanisms underpinning bacterial community assembly along soil gradients. Finally and importantly, the CS database is spatially constrained to a temperate island in Northern Europe and would benefit from a more global extent to capture other soil biomes such as drylands. Improvements here could be made from integrating data from global sequencing initiatives or leveraging data from sequence repositories provided consistent environmental metadata can also be retrieved in order to reliably predict response trait characteristics.

2.5 Bibliography

Austin, M. P. (1999). The potential contribution of vegetation ecology to biodiversity research. *Ecography, 22*(5), 465-484. doi:10.1111/j.1600-0587.1999.tb01276.x

Bickel, S., Chen, X., Papritz, A., & Or, D. (2019). A hierarchy of environmental covariates control the global biogeography of soil bacterial richness. *Scientific Reports, 9*(1), 12129. doi:10.1038/s41598-019-48571-w

Brewer, T. E., Handley, K. M., Carini, P., Gilbert, J. A., & Fierer, N. (2016). Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. *Nat Microbiol, 2*(October 2016), 16198. doi:10.1038/nmicrobiol.2016.198

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., . . . Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences, 108*(Supplement 1), 4516-4522. doi:10.1073/pnas.1000080107

Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., . . . Howe, A. (2017). Strategies to improve reference databases for soil microbiomes. *The ISME Journal, 11*(4), 829-834. doi:10.1038/ismej.2016.168

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., . . . Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science, 359*(6373), 320-325. doi:10.1126/science.aap9516

Diaz, S., Purvis, A., Cornelissen, J. H., Mace, G. M., Donoghue, M. J., Ewers, R. M., . . . Pearse, W. D. (2013). Functional traits, the phylogeny of function, and ecosystem service vulnerability. *Ecol Evol, 3*(9), 2958-2975. doi:10.1002/ece3.601

Dufrene, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs, 67*(3), 345-366. doi:10.1890/0012-9615(1997)067[0345:Saaist]2.0.Co;2

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ, 5*(6226), e3889. doi:10.7717/peerj.3889

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol, 15*(10), 579-590. doi:10.1038/nrmicro.2017.87

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A, 103*(3), 626-631. doi:10.1073/pnas.0507535103

Fierer, N., Ladau, J., Clemente, J. C., Leff, J. W., Owens, S. M., Pollard, K. S., . . . McCulley, R. L. (2013). Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science, 342*(6158), 621-624. doi:10.1126/science.1243768

Fox, G. E., Wisotzkey, J. D., & Jurtshuk, P. (1992). How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *International Journal of Systematic and Evolutionary Microbiology, 42*(1), 166-170. doi:https://doi.org/10.1099/00207713-42-1-166

Gans, J., Wolinsky, M., & Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science, 309*(5739), 1387-1390. doi:10.1126/science.1112665

Gao, J. et al. (2017) 'The impact of land-use change on water-related ecosystem services: A study of the Guishui River Basin, Beijing, China', Journal of Cleaner Production. Elsevier Ltd, 163, pp. S148–S155. doi: 10.1016/j.jclepro.2016.01.049.

George, P. B. L., Lallias, D., Creer, S., Seaton, F. M., Kenny, J. G., Eccles, R. M., . . . Jones, D. L. (2019). Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems. *Nature Communications, 10*(1), 1107. doi:10.1038/s41467-019-09031-1

Griffiths, B. S. and Philippot, L. (2013) 'Insights into the resistance and resilience of the soil microbial community', FEMS Microbiology Reviews. Oxford Academic, pp. 112–129. doi: 10.1111/j.1574-6976.2012.00343.x.

Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., & Whiteley, A. S. (2011). The bacterial biogeography of British soils. *Environ Microbiol, 13*(6), 1642-1654. doi:10.1111/j.1462-2920.2011.02480.x

Griffiths, R. I., Thomson, B. C., Plassart, P., Gweon, H. S., Stone, D., Creamer, R. E., . . . Bailey, M. J. (2016). Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets. *Applied Soil Ecology, 97*, 61-68. doi:https://doi.org/10.1016/j.apsoil.2015.06.018

Jansen, F., & Oksanen, J. (2013). How to model species responses along ecological gradients – Huisman–Olff–Fresco models revisited. *Journal of Vegetation Science, 24*(6), 1108-1117. doi:10.1111/jvs.12050

Jia, Y. and Whalen, J. K. (2020) 'A new perspective on functional redundancy and phylogenetic niche conservatism in soil microbial communities', Pedosphere. Soil Science Society of China, 30(1), pp. 18–24. doi: 10.1016/S1002-0160(19)60826-X.

Jones, D. L., Cooledge, E. C., Hoyle, F. C., Griffiths, R. I., & Murphy, D. V. (2019). pH and exchangeable aluminum are major regulators of microbial energy flow and carbon use efficiency in soil microbial communities. *Soil Biology and Biochemistry, 138*, 107584. doi:https://doi.org/10.1016/j.soilbio.2019.107584

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing

Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology, 79*(17), 5112-5120. doi:10.1128/aem.01043-13

Lavorel, S., & Garnier, E. (2002). Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology, 16*(5), 545-556. doi:DOI 10.1046/j.1365-2435.2002.00664.x

Malik, A. A., Puissant, J., Buckeridge, K. M., Goodall, T., Jehmlich, N., Chowdhury, S., . . . Griffiths, R. I. (2018). Land use driven change in soil pH affects microbial carbon cycling processes. *Nature Communications, 9*(1), 3591. doi:10.1038/s41467-018-05980-1

Martiny, A. C., Treseder, K., & Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal, 7*(4), 830-838. doi:10.1038/ismej.2012.160

Martiny, J. B., Bohannan, B. J., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., . . . Staley, J. T. (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol, 4*(2), 102-112. doi:10.1038/nrmicro1341

Martiny, J. B. H., Jones, S. E., Lennon, J. T., & Martiny, A. C. (2015). Microbiomes in light of traits: A phylogenetic perspective. *Science, 350*(6261), aac9323. doi:10.1126/science.aac9323

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., . . . Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J, 6*(3), 610-618. doi:10.1038/ismej.2011.139

Muyzer, G., de Waal, E. C., & 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), 695-700.

Oksanen, J. et al. (2018) 'vegan: Community Ecology Package'. Available at: https://cran.r-project.org/package=vegan.

Parks, D. H. et al. (2018) 'A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life', Nature Biotechnology. Nature Publishing Group, 36(10), p. 996. doi: 10.1038/nbt.4229.

Parks, D. H. et al. (2020) 'A complete domain-to-species taxonomy for Bacteria and Archaea', Nature Biotechnology. Nature Research, 38(9), pp. 1079–1086. doi:

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One, 5*(3), e9490. doi:10.1371/journal.pone.0009490

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res, 41*(Database issue), D590-596. doi:10.1093/nar/gks1219

Reynolds, B., Chamberlain, P. M., Poskitt, J., Woods, C., Scott, W. A., Rowe, E. C., . . . Emmett, B. A. (2013). Countryside Survey: National "Soil Change" 1978–2007 for Topsoils in Great Britain—Acidity, Carbon, and Total Nitrogen Status. *Vadose Zone Journal, 12*. doi:10.2136/vzj2012.0114

Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Km, A., Kent, A. D., . . . Triplett, E. W. (2010). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal, 1*(4), 283--290. doi:10.1038/ismej.2007.53.Pyrosequencing

Rognes, T. et al. (2016) 'VSEARCH: A versatile open source tool for metagenomics', PeerJ. doi: 10.7717/peerj.2584

Sinclair, L., Ijaz, U. Z., Jensen, L. J., Coolen, M. J. L., Gubry-Rangin, C., Chronakova, A., . . . Pafilis, E. (2016). Seqenv: linking sequences to environments through text mining. *PeerJ, 4*(e2690), e2690. doi:10.7717/peerj.2690

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., . . . Woyke, T. (2016). High-resolution phylogenetic microbial community profiling. *ISME J, 10*(8), 2020-2032. doi:10.1038/ismej.2015.249

Slessarev, E. W., Lin, Y., Bingham, N. L., Johnson, J. E., Dai, Y., Schimel, J. P., & Chadwick, O. A. (2016). Water balance creates a threshold in soil pH at the global scale. *Nature, 540*, 567. doi:10.1038/nature20139

Suding, K. N., Lavorel, S., Chapin, F. S., Cornelissen, J. H. C., Diaz, S., Garnier, E., . . . Navas, M. L. (2008). Scaling environmental change through the community-level: a trait-based

response-and-effect framework for plants. *Global Change Biology, 14*(5), 1125-1140. doi:10.1111/j.1365-2486.2008.01557.x

Takahashi, S. et al. (2014) 'Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing', PLoS ONE, 9(8). doi: 10.1371/journal.pone.0105592.

Thijs, S., Op De Beeck, M., Beckers, B., Truyens, S., Stevens, V., Van Hamme, J. D., . . . Vangronsveld, J. (2017). Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. *Frontiers in microbiology, 8*, 494-494. doi:10.3389/fmicb.2017.00494

Thomson, B. C., Ostle, N., McNamara, N., Bailey, M. J., Whiteley, A. S., & Griffiths, R. I. (2010). Vegetation affects the relative abundances of dominant soil bacterial taxa and soil respiration rates in an upland grassland soil. *Microb Ecol, 59*(2), 335-343. doi:10.1007/s00248-009-9575-z

Wamelink, G. W. W., Walvoort, D. J. J., Sanders, M. E., Meeuwsen, H. A. M., Wegman, R. M. A., Pouwels, R., & Knotters, M. (2019). Prediction of soil pH patterns in nature areas on a national scale. *Applied Vegetation Science, 22*(2), 189-199. doi:10.1111/avsc.12423

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol, 73*(16), 5261-5267. doi:10.1128/AEM.00062-07

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Available at: http://ggplot2.org.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2013). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30*(5), 614-620. doi:10.1093/bioinformatics/btt593

Chapter 3

Land use intensification effects on soil functions assessed through metagenomic profiling

Abstract

Land use intensification to meet current human food needs can lead to soil degradation and is known to affect the microbial communities responsible for orchestrating soil processes. However, to date there is little information available on how land use intensification affects the abundances of microbial functional genes responsible for a variety of soil processes, nor any appreciation of whether consistent responses are observed across different soils and land use systems. Here I analyse 96 metagenomes from distributed paired land management intensity contrasts, applying statistical and machine learning approaches to identify the functional genetic differences between high and low intensity land use. Ordination and clustering methods demonstrated that, whilst pH was a major influence on functional profiles, there was also separation based on organic matter contents, which was influenced by land use change. Pairwise differences in functional profiles mirrored patterns observed for bacterial taxonomic biodiversity, assessed using 16S rRNA gene amplicon sequencing. Using a random forest (RF) approach, I was able to identify functional genes and categories that responded consistently to high and low management factors, irrespective of specific management or other abiotic factors. Components of the Nitrate reductase (Nar) complex were found to be consistently higher within high intensity sites, likely due to the application of nitrogen rich fertilizers. More generally I used this large geographically distributed metagenomic study encompassing various land uses and soil types to better understand the environmental preferences of functional genes in terms of pH and organic matter and wider soil attributes.

3.1 Introduction

In order to meet current human demands, soils are dramatically transformed through both largescale land use change (human transformation of the landscape) and more subtle modifications of land management (the soil regimen applied e.g. liming, fertilizer, or tillage). Whilst this is valuable in the short term to meet societal demands for food and fibre production, it can come at the cost of environmental degradation and long term loss of important soil functions (Foley et al., 2005; Guo & Gifford, 2002). Indeed increased agriculture can lead to increased greenhouse gas emissions (Boetius, 2019) of carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O), as well as depletion of soil organic carbon (SOC). However, some land use practices are both simultaneously beneficial and detrimental to ecosystem services, for example minimal tillage is associated with increased carbon sequestration (Paustian et al., 2016) but also with increased levels of N_2O emissions (Badagliacca et al., 2018; Bayer et al., 2015). As such improved understanding of soil trade-offs prior to modifications in land use is desirable in order to design and plan management regimes to both produce biomass for human consumption and limit negative environmental consequences (Maskell et al., 2013); and this is likely to require a deeper understanding of the underlying mechanisms controlling various soil functions. More generally, given uncertainties about the benefits of altered management practices for soil carbon sequestration alone, there is increased interest in how increases in topsoil organic matter content from less intensive management, provides additional benefit for soil functionality and the ecosystem services it provides.

Microbial communities make up the majority of biodiversity in soils and are known to be large contributors to soil functionality (Daniel, 2005). We know microbes play an important role in carbon dynamics, where they carry out both mineralisation of soil organic carbon (SOC) (resulting in subsequent CO_2 loss) alongside carbon stabilisation into microbial biomass (Jansson & Hofmockel, 2020; Trivedi et al., 2016). They are also involved in numerous other biogeochemical cycles such as nitrogen, hydrogen and phosphorus cycling and are generally important to soil fertility by recycling essential plant nutrients (Turner, 2010). Since microbial communities in soil are highly diverse with most taxa being unculturable, we lack deep insight as to how the functioning of these communities will be

altered by land use transitions. This is largely due to uncertainties with regards to the extent of functional redundancy across microbial lineages (Griffiths & Philippot, 2013), some studies have shown that generic processes such as respiration are insensitive to diversity change, but we know little about the redundancy of rarer functions for example nitrogen fixation, methane production or pesticide degradation etc. (Jia and Whalen, 2020 ; Grządziel, 2017). Improved understanding of the effects of land use on a variety of processes of relevance to soil functions, could help inform future land use policy as well as provide novel understanding of the functional ecology of previously understudied soil microbial taxa.

Previously, exploring microbial function in soils has been hampered by the fact that the majority of bacteria are unable to be cultured in lab conditions. The advent of high-throughput sequencing has enabled us to gain new insights in spite of this, furthering understanding of community biodiversity change using amplicon analyses (assessment of defined taxonomic marker genes), and functional change in the "collective genome" of communities through metagenomics (Handelsman, 2004). New discoveries through using these techniques, could help us better understand how resilient soil systems are to future changes in the environment and anthropogenic pressures; this is particularly pertinent today with impending future climate and land use change (Boetius, 2019). As an example of the power of these technologies, the recent discovery of microbes conducting complete oxidation of ammonia to nitrate without releasing N_2O (commamox) (Van Kessel et al., 2015; Daims et al., 2015), has revolutionised our understanding of the Nitrogen cycle and its abundance in soils (Xia et al., 2018) shows how looking at soil microbial function can help us to understand mechanisms that have the potential to reduce the impact of agricultural practices and help mitigate climate change (Jansson & Hofmockel, 2020).

Whilst many studies have assessed land use change effects using amplicon approaches; functional metagenomics studies applied in this context are comparatively in their infancy. Large scale surveys employing amplicon analyses across many different soils have revealed pH in particular to be a strong correlate of bacterial taxonomic biodiversity (Fierer & Jackson, 2006; Griffiths et al., 2011); consistent with the known effects of pH observed more generally on broad soil functions (Jones, et al., 2019; Malik et al., 2018). Furthermore there

are numerous localised studies demonstrating the impact of land use change on taxonomic composition of microbial communities (Banerjee et al., 2019; Hartmann, Frey, Mayer, Mäder, & Widmer, 2015; Pershina et al., 2015; Schöps et al., 2018).

Previously, metagenomics has been applied to N and P addition experiments on globally distributed grasslands and found clear differences in carbohydrate metabolism between N and P treatments, alongside shifts in bacterial life strategies (Leff et al., 2015). Other more localised N addition experiments have shown consistent increases in respiration related genes at higher nitrogen inputs (Barberán, Bates, Casamayor, & Fierer, 2012). Another molecular approach Geochip (microarray) has been used to look at the long term impact of inorganic fertilizer on microbial function and plant interactions at the park grass experiment treated with inorganic fertilizers for over 100 years, and showed that long term fertilization lead to less connectivity between plant and microbial communities (Huang et al., 2019). There have also been local metagenome studies looking at land use change between tropical forest, grasslands and arable (Goss-Souza et al., 2017), pasture compared to pristine rainforest (Kroeger et al., 2018) and assessing the impact of mono-cropping vs crop rotation on microbial function on dryland soils (Li et al., 2019). There is now a need to perform similar work across UK soils, specifically at geographically distributed sites looking the same land use contrasts to identify whether there are functional consistencies in genomic responses induced by real world management practices.

3.1.1 Chapter Aims

Here, I addressed this need by examining metagenomes from UK distributed land use contrasts. To do this I used a dataset with a unique study design whereby soils were obtained from ten geographically distributed sites where there were existing paired land use contrasts. Each low intensity grassland was positioned adjacent to a high intensity grassland or arable soil and an additional bare fallow site was included as a functionally depauperate contrast (no inputs > 50yrs). Nintey six HiSeq shotgun metagenomes were sequenced and a range of statistical techniques were used to assess supervised and unsupervised groupings within the data. I chose to employ a machine learning approach to identify global genomic variables for classifying high and low land use intensities to alleviate

the issues associated with working with high dimensional data (Touw et al., 2013). Since I expected that soil metagenomes from geographically distributed sites would be influenced by a range of abiotic factors, I wished to specifically explore the degree of control exerted by land use management.

The specific aims are:

(i) **To determine overall land use associated functional variables:** To pinpoint genes and wider SEED subsystems that are globally important in discriminating between high and low intensity sites.

(ii) **To identity whether land use associated functional variables are consistent across sites:** To determine if specific genes or SEED subsystems consistently increase or decrease in response to land use intensity across sites. Despite the subtle differences within site specific land use intensity treatments and various abiotic factors that are also likely to also exert control on microbial populations.

(iii) **To contribute more widely to the understanding of functional gene responses in soils**: Through identifying genomic variables associated with wider abiotic factors such as organic matter and pH.

3.2 Materials and Methods

3.2.1 Soil sampling

Samples were collected between April and August 2015 as part of the Soil Security programme's UGRASS project. Paired sample sites were chosen where pristine fields were adjacent to intense agricultural sites. A 100 m transect was used to take 5 pairs of cores (15cm depth, 5cm diameter) at the boundary of the two intensities every 25 m. An additional 5 unpaired cores were taken at sites with intermediate land uses. Surface litter was removed from soil cores. Soil cores were homogenised wet without sieving prior to subsampling for DNA extraction.

3.2.2 16S Sequencing

To analyse bacterial communities, DNA was extracted using 0.25 g of soil and the PowerSoil-htp 96 Well DNA Isolation kit (Qiagen) according to manufacturer's protocols. The dual indexing protocol of Kozich et al (Kozich, Westcott, Baxter, Highlander, & Schloss, 2013) was used for Illumina MiSeq sequencing of the V3–V4 hypervariable regions of the bacterial 16S rRNA gene using primers 341F (Muyzer, De Waal, & Uitterlinden, 1993) and 806R (Yu, Lee, Kim, & Hwang, 2005). Amplicon concentrations were normalized using SequalPrep Normalization Plate Kit (Thermo Fisher Scientific), sequencing was then conducted on the Illumina MiSeq using V3 chemistry. Sequenced paired end reads were joined with PEAR (sco.h-its.org/exelixis/web/software/pear) and then quality filtered using FASTX tools (hannonlab.cshl.edu). Chimeras were then removed with VSEARCH_UCHIME_REF and clustering was conducted with VSEARCH_CLUSTER into 97% OTUs (github.com/torognes/vsearch). There were an average of ~23770 reads per sample, samples with < 1000 reads were removed. Relative abundance of OTU's were calculated by dividing raw OTU counts by total read counts per sample using Vegan Decostand function (method "total") (Oksanen et al., 2018).

3.2.3 Metagenome Sequencing

DNA was extracted from 2g of soil using the power max soil DNA isolation soil kit, and subsequently purified using a millipore amplicon ultra buffer exchange. 96 Illumina libraries were constructed using the Illumina TruSeq library preparation kit (insert size < 500- 600 bp). Paired-end sequencing (2 x 150 bp) was conducted using the Illumnia Hiseq 4000 platform, 96 indexed libraries were multiplexed across 8 lanes and generated in excess 280M clusters per lane.

Reads then underwent bioinformatic pre-processing, Illumina adaptor sequences were detected and removed using Cutadapt 1.2.1, reads were subsequently trimmed with Sickle 1.200 with a minimum window quality score of 20. Reads shorter than 20bp after trimming were discarded. Resulting in ~15000000 to ~38000000 trimmed reads per sample and an average trimmed read length of ~148 bp.

All trimmed reads were functionally annotated to SEED subsystems (a hierarchical classification system, based on biological groupings related by process or structure) (Overbeek et al., 2005) using kmers (k=9) to detect similarity using standalone RAST server (Aziz et al., 2012; Overbeek et al., 2014). On average 11.6% of reads were annotated to SEED per sample. Relative abundance of genes was calculated by dividing raw counts by total gene number per sample using Vegan Decostand function (method "total") (Oksanen et al., 2018).

3.2.4 Statistical analysis

To assess which genomic variables were globally important in discriminating land uses, two random forest models were generated on all data both on the subsystem (SEED subsystem level 1) and gene level. Random forest is an ensemble learning approach which uses many decision trees, with each tree using a sub sample of data and variables (Breiman, 2001), this approach is becoming more widely used in genomic analyses, due to the methods ability to identify significant variables in high dimensional data. For each random forest model, genes and subsystems that were present in less than 30% of samples were discarded and data was z score transformed to make relative abundance of genomic variables more comparable across samples. For both model's data was first split into a test and training dataset (70%/30% split of samples), model parameters were then tuned using the training dataset with a cross validation K fold of 10. Genomic variables identified as important discriminators of land use were further analysed in order to assess the significance of differences in total relative abundance between high and low land intensities. This was conducted using linear modelling (using relative abundance of gene/ subsystem as dependent variable and management intensity and site as covariates with interaction) and a two way anova. The significance of within site differences were assessed using TukeyHSD.

To examine interactions between genomic variables and further understand the influence of abiotic factors I produced a network based on correlations. Spearmans rank correlation was used to compare all genomic variables (SEED subsystem level 3) with each other as well as a number of soil characteristics. Correlations weaker than 0.7 and with a Benjamini-Hochberg corrected p value of more than 0.01 were excluded. The network was visualised with igraph

and visNetwork in R. Subgraphs (clusters of variables with common connections (Pons & Latapy, 2006)) were computed using igraph's cluster_walktrap function.

3.3 Results

3.3.1 Supervised/ Unsupervised clustering of annotated gene relative abundance

NMDS was used to assess overall similarity of both 16S (**Fig.3.1a**) and functional gene profiles across land uses (**Fig.3.1b**). The high and low intensity 16S profiles show some separation, most markedly on the first axis, which appears to be related to pH (green contours). The functional profiles also showed separation between high and low intensity sites (**Fig.3.1b**), although the difference in functional profiles appeared more striking than that in the 16S communities, with clearer separation on both axes. However consistent with the 16S communities the first axis appeared to be heavily related to pH (green contours **Fig.3.1**).

The bare fallow 16S (**Fig.3.1a**) community appear particularly distinct on both NMDS axis, which was also seen in the functional profiles (**Fig.3.1b**), given that these soils have not been cropped in 50 years and therefore are low in organic matter and are heavily degraded, their distinct taxonomic and functional profile is unsurprising. The greater distinction in the functional profiles in comparison to the 16S communities could indicate a redundancy in taxa able to deliver soil functions. Alternatively, it could be indicative of functional differences not driven by bacterial taxa that are potentially orchestrated by archaea or fungi, and thus not picked up in these 16S analyses, although some 16S primers detect archaea well (for example EMP) the primers used here (covering the V3-V4 region) are known to be comparatively poor at archaea detection.

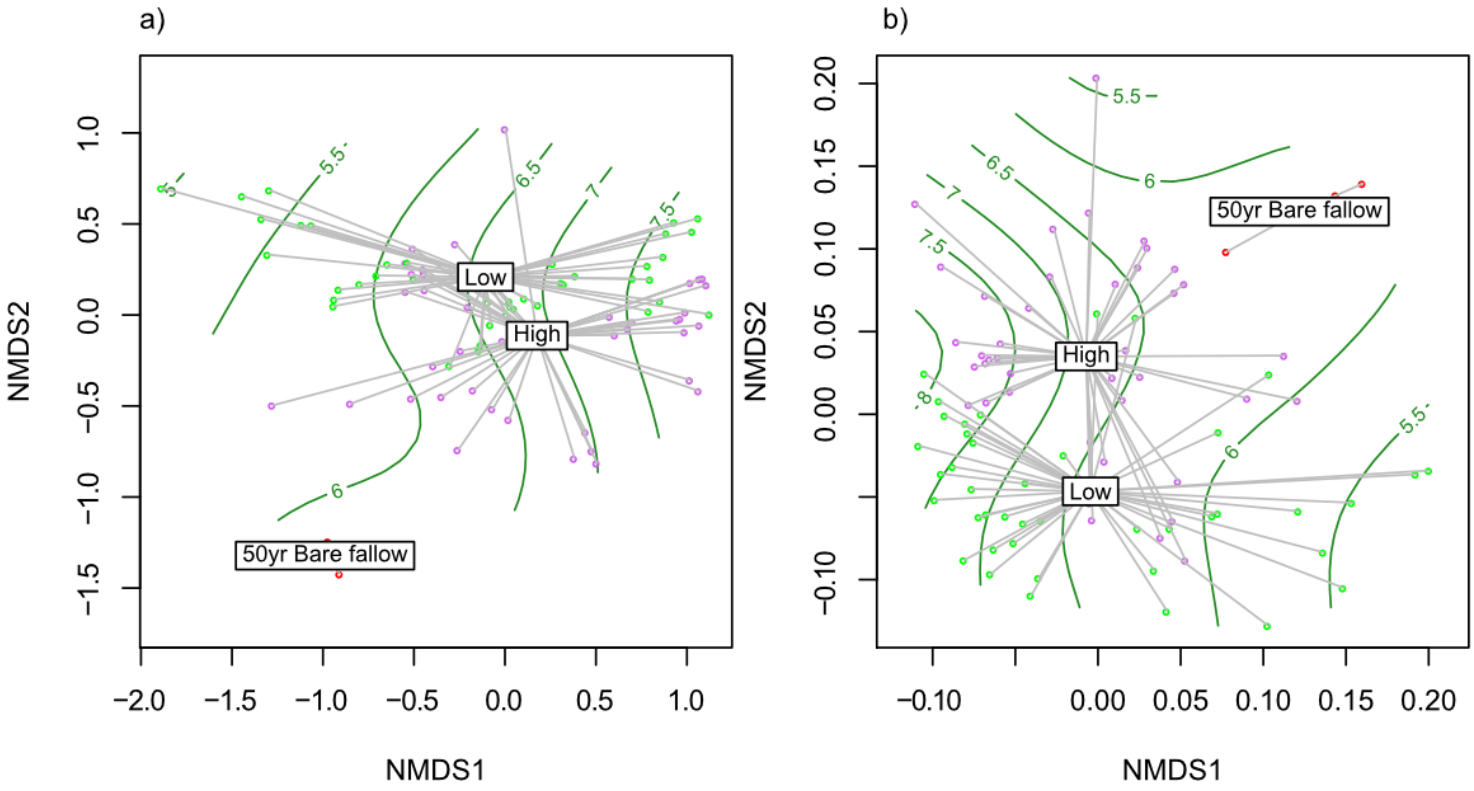


Fig. 3.1. Ordinations of 16S **(a)** and functional gene **(b)** relative abundances across high and low land use intensities and bare fallow soils. Labels and point colour indicate land use intensity, green contours represent soil pH gradient.

To assess unsupervised groupings within the data I used k-means clustering, groups of 2 and 3 were shown to describe the data most accurately based on their BIC scores. Separating the data into two groups shows clear clustering with pH (**Fig.2a**), with cluster one containing sites with pH's of 6.24 > pH > 8.12 and cluster 2 with sites between 4.83> pH > 6.9. Whilst separating by three groups (**Fig.2b**), shows the third group is separated on the NMDS2 axis which appears to have a relationship to organic matter with sites within that grouping having a 3.99 > loi >14.5 (loss of ignition). This highlights the challenge of pinpointing the functional differences at varying land use when there are clear gradients within the landscape such as pH and organic matter appearing to be highly influential on functional profiles. The groupings of sites (k=3) and average pH and loi per site and intensity can be seen in **Table.3.1**.

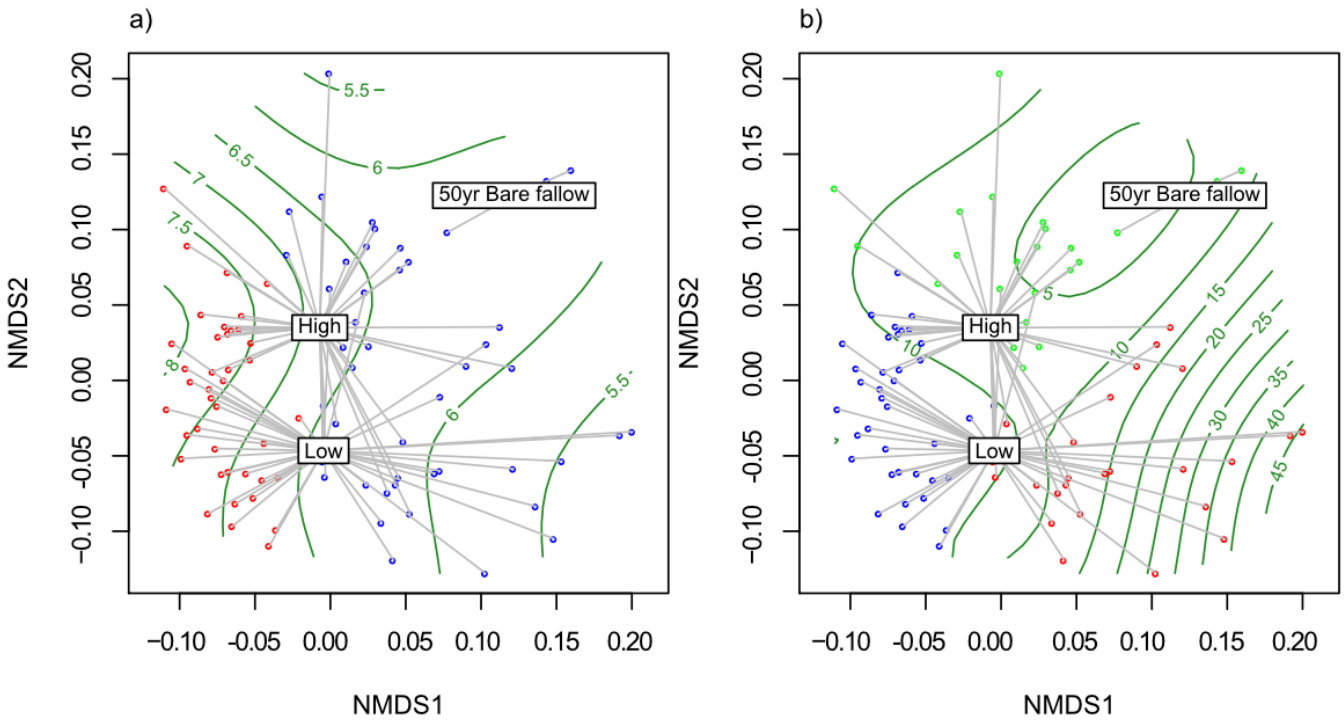


Fig. 3.2. NMDS plots based upon relative abundance of functional genes, labelled by land use intensity. Points coloured by k means clustering groupings with groups of 2 **(a)** and 3 **(b)** respectively. Green contours show pH **(a)** and organic matter gradients **(b)**.

									Proportion of samples in k-means cluster		
Site Name	**Contrast type**	**County**	**Intensity**	**Management**	**Latitude**	**Longitude**	**pH**	**OM (%)**	**Cluster 1 (Red)**	**Cluster 2 (Green)**	**Cluster 3 (Blue)**
Affeton Moor	Working farm	Devon	High	Intensive grass (tillage and N)	50.92193	-3.75794688	6.35	20.48	1	0	0
			Low	Unimproved grass	50.92064	-3.75743940	5.73	28.05	1	0	0
Foster	Experiment	Hertfordshire	High	Arable (tillage+N)	51.81292	-0.37775159	6.87	4.69	0	1	0
			Low	Unimproved grass	51.81304	-0.37775613	6.48	7.52	0	0.5	0.5
Highfield	Experiment	Hertfordshire	High	bare fallow (tillage-N)	51.80423	-0.36145143	6.26	4.44	0	1	0
			High	Arable (tillage+N)	51.80429	-0.36247759	6.37	5.57	0	1	0
			Low	Unimproved grass	51.80424	-0.36217155	6.39	8.7	0	0	1
North Wyke	Working farm	Devon	High	Intensive grass (tillage and N)	50.77085	-3.90788125	6.6	9.47	1	0	0
			Low	Unimproved grass	50.78145	-3.91716375	6.15	14.25	1	0	0
Park Grass pH 5	Experiment	Hertfordshire	High	Intensive grass (N)	51.80387	-0.37354547	5.76	11.34	0	1	0
			Low	Unimproved grass	51.80366	-0.37439303	5.48	7.48	1	0	0
Park Grass pH 7	Experiment	Hertfordshire	High	Intensive grass (N + lime)	51.80356	-0.37330848	7.5	10.27	0	0.25	0.75
			Low	Unimproved grass (+Lime)	51.80338	-0.37417658	7.38	9.78	0	0	1
Parsonage Down	Working farm	Wiltshire	High	Arable (tillage + N)	51.17486	-1.91215707	8.02	7.91	0	0	1
			Low	Unimproved grass	51.17483	-1.91891150	7.65	20.4	0	0	1
RSPB_Hope Farm	Working farm	Cambridgeshire	High	Arable (tillage + N)	52.24454	-0.05109275	7.92	8.62	0	0	1
			Low	Unimproved grass	52.24419	-0.05014525	7.57	15.6	0	0	1
SRUC Kirkton	Experiment	Perthshire	High	Intensive grass (tillage and N)	56.41664	-4.66064800	6.37	8.89	1	0	0
			Low	Unimproved grass	56.42050	-4.66829500	5.18	48.57	1	0	0
Strawberry Farm	Working farm	Leicestershire	High	Arable (tillage+N)	52.75973	-0.76247125	6.22	7.31	0	0.75	0.25
			Low	Unimproved grass	52.75972	-0.76430975	6.8	13.12	0	0	1
Top clumps	Working farm	Oxfordshire	High	Arable (tillage+N)	51.62806	-1.18471980	7.82	4.96	0	0.67	0.33
		Oxfordshire	Low	Unimproved grass	51.62871	-1.18438537	7.5	13.17	0	0	1

Table.3.1. Description of treatments at each site, site coordinates (latitude, longitude), average pH, organic matter (% loss on ignition). Proportion of samples assigned to k-means clusters (based upon metagenomics functional profiles) are also reported per treatment/site, with colours cross referencing with **Fig 3.2b**.

3.3.2 Genomic correlates of soil characteristics

As clustering analyses appeared to suggest clear influences of pH and organic matter (OM) on functional gene content, I next sought to look at the specific genomic variables correlating with abiotic factors through generating a correlation network. Nodes within the network represent SEED subsystems (level 3) alongside abiotic variables, whilst edges represent correlations with an R value of > 0.7 and Benjamini-Hochberg corrected p value of < 0.01. Subgraphs were identified using igraphs cluster_walktrap function, represented by node colour in **Fig.3.3**. This network and subsequent clustering analyses showed a clear subgraph of genes related to organic matter (OM) and moisture. Whilst pH was found in a separate subgraph (**Fig.3.3**), providing further evidence of pH independent effects driving genomic profiles. Bulk density (BD) was also found within a distinct subgraph to organic matter and moisture. This makes intuitive sense as higher bulk density results in less pore space and consequently less moisture and OM, so one would therefore expect OM and moisture to have few genomic connections in common with BD.

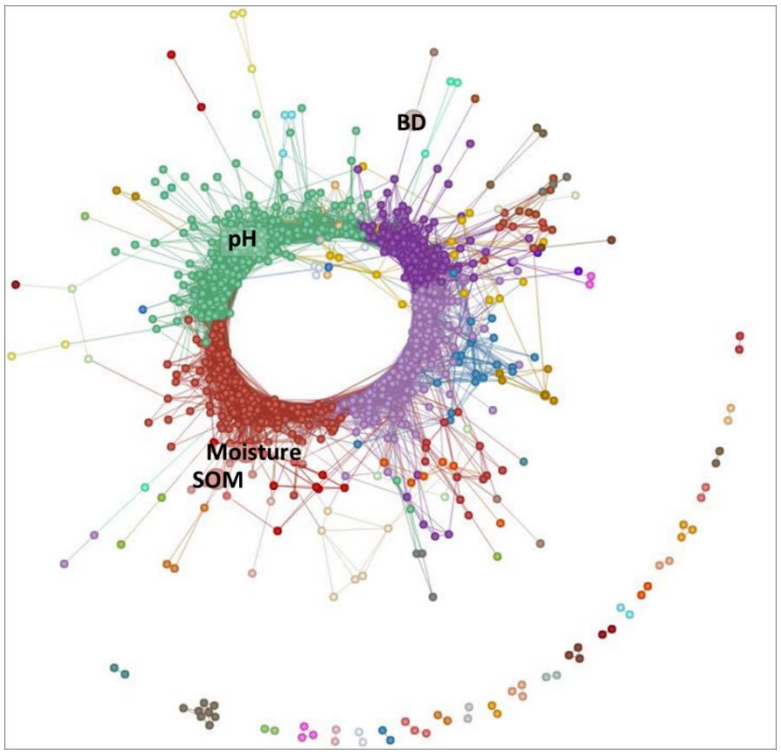


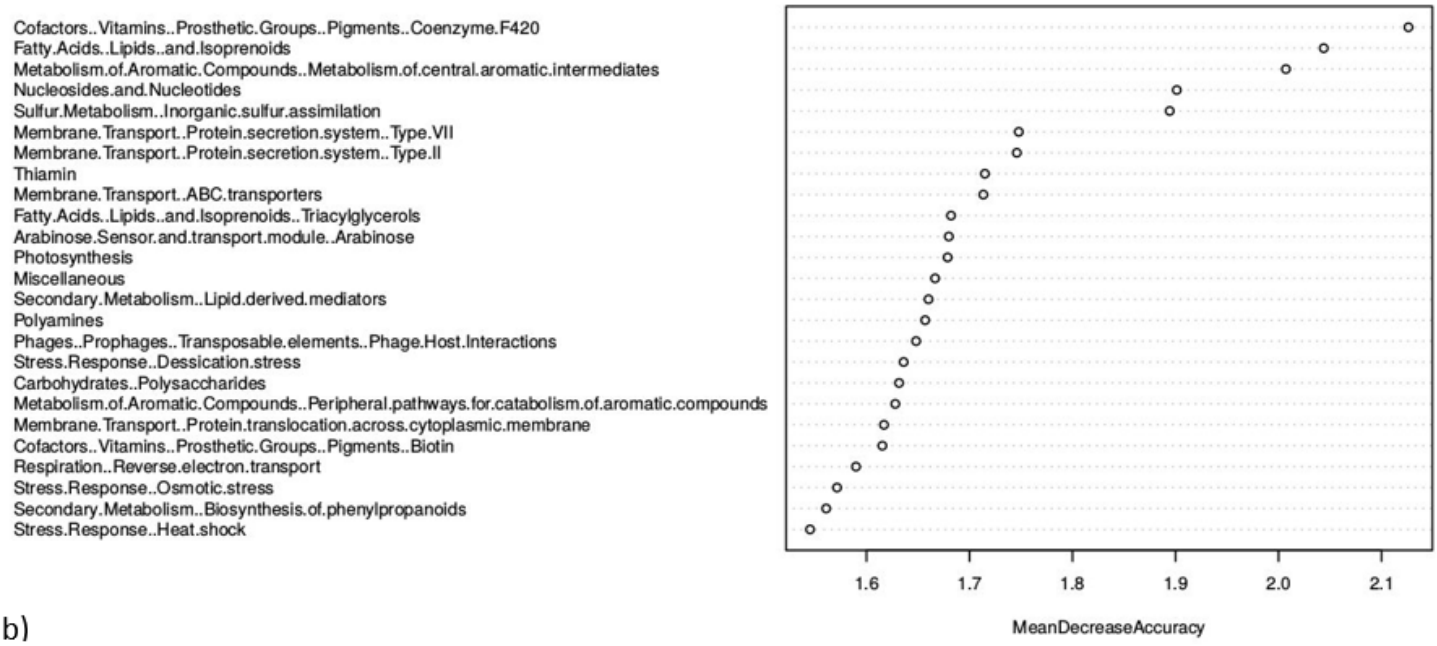
Fig.3.3. Correlation Network (based upon Spearman's rank correlations) of all genomic variables (level 3 SEED subsystems) and soil characteristics. Connections indicate an R value of > 0.7 and Benjamini-Hochberg corrected p value of < 0.01. Nodes coloured by subgraphs (computed using random walk clustering).

3.3.3 Are there consistent functional indicators of land use change?

In order to gain a global picture of genomic variables that were strong classifiers of different land uses I used two separate random forest models using gene level and broader SEED subsystem (level 1) classification abundances respectively. I explored classifiers at different levels as its possible a functional group may be a weak determinant of land use, but a specific gene may be a stronger classifier, or vice versa.

The model generated at the broad seed subsystem level had an 86% level of accuracy according to cross validation and had 73% accuracy on the training set. Important determinants of land use change highlighted by the model included Fatty acids lipids and isoprenoids (mean decrease of accuracy 2.04), and various membrane transport related subsystems including Protein secretion system Type VII (mean decrease accuracy 1.75) and Type II (mean decrease accuracy 1.75), and ABC transporters (mean decrease accuracy 1.71) (**Fig.3.4a**). The random forest on functional gene abundance data had an 86.3% level of accuracy according to cross validation and had 76.9% accuracy on the training set. On the gene level (**Fig.3.4b**) the most important classifier appeared to be respiratory nitrate alpha chain (NarG) which led to a mean decrease accuracy of 3.83, the respiratory nitrate beta chain (NarH) also appeared important with a mean decrease accuracy of 2.57. Other important classifiers included flagellar FlbT (motility) (decrease 2.76), DipZ protein (decrease 2.71) and protein co-occurring with transport systems (decrease 2.9).

a)



b)

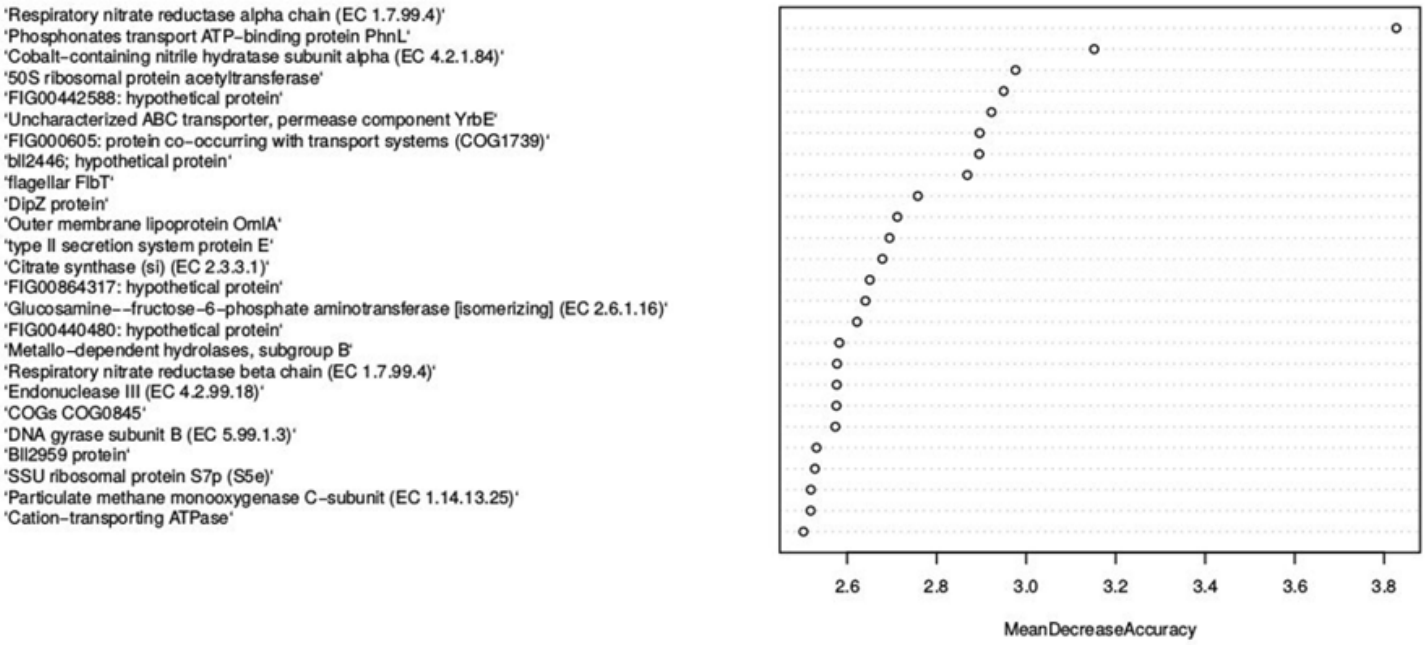


Fig.3.4. Random forest variable importance plots identifying genomic variables which contribute to land use intensity classification model accuracy. Random forest models are based on **a)** subsystem (level 1) and **b)** gene relative abundance.

To further explore the classifiers identified through random forest and to investigate how consistent these classifiers are at discriminating land use, I next plotted their relative abundances per site. The significance of differences of genes/ subsystems between land uses (total relative abundance) was assessed using linear modelling (with relative abundance of the gene/ subsystem as the dependent variable and management intensity and site as covariates with interaction) and a two way anova. Within site differences were calculated using the post hoc test TukeyHSD.

On the subsystem level the total relative abundance of Membrane Transport system Type II (T2SS) was significantly increased within high Intensity soils compared to low intensity soils (pval 0.012). Increased levels of T2SS in high intensity soils was also visible within some individual sites, although these differences were not significant (**Fig.3.5a**). Interestingly Membrane Transport system Type VII (T7SS) was also an important variable in distinguishing site land use, but showed the inverse trend of T2SS, with a significantly higher total relative abundance within low Intensity soils in comparison to high intensity soils (pval 5.32e-08) (**Fig.3.5b**). There were also significantly elevated levels of T7SS within Foster and Highfield low intensity sites when compared to their corresponding high intensity sites. Other sites (Park Grass pH7, Parsonage Down, SRUC Kirkton, Strawberry Farm and Top Clumps) also showed increased levels of T7SS within low intensity soils, though these differences were not significant.

On the gene level, Respiratory nitrate reductase alpha chain (NarG), the catalytic unit of transmembrane nitrate reductase (which reduces nitrate to nitrite) appeared to be in increased relative abundance in high intensity soils in comparison low intensity soils within almost all sites (**Fig.3.6a**). Correspondingly respiratory nitrate reductase beta chain (NarH) the electron transport unit was found in in higher relative abundance within high intensity land use within all sites (**Fig.3.6b**). Further, the total relative abundances of both NarG and NarH were significantly increased within high intensity soils in comparison to low intensity soils (both p val 2e-16). NarG and NarH also demonstrated common statistical significance levels within multiple sites (Affeton Moor, North Wyke, Park Grass pH7, Park Grass pH5 and SRUC Kirkton) as expected given they are components of the same complex (**Fig.3.6a**, **Fig.3.5b**). Whilst other components of Nar were not identified as key discriminators of land

use within the random forest (and thus were not explored in greater detail), the relative abundance of NarG, NarH, NarL and NarJ were all highly correlated (Spearmans rank correlation of R> 0.8 and Benjamini-Hochberg corrected pval < 0.01).

The total relative abundance of Flagellar FlbT (flagellum biogenesis repressor) was significantly increased in low intensity soils (**Fig.3.6c**) (pval 2.79e-07). FlbT also appeared to be found in increased levels in low intensity soils within all sites, although no within site differences were statistically significant. Type II secretion system protein E (GspE, secretion ATPase) demonstrated significantly higher total relative abundance in high intensity soils in comparison to low (pval 5.66e-14) **(Fig.3.6d**), consistent with what was observed at the broader SEED subsystem level. Further GsPE showed increased levels within high intensity soils within most sites, with significant differences observed within Affeton Moor, Park Grass pH 7 and Parsonage Down. The reasonable classification rate of land use intensities within the RF models and consistent differences in relative abundance for specific genomic variables (particularly NarG and NarH) illustrate that despite various natural gradients, there are consistent genomic variables that enable discrimination between land use intensities.

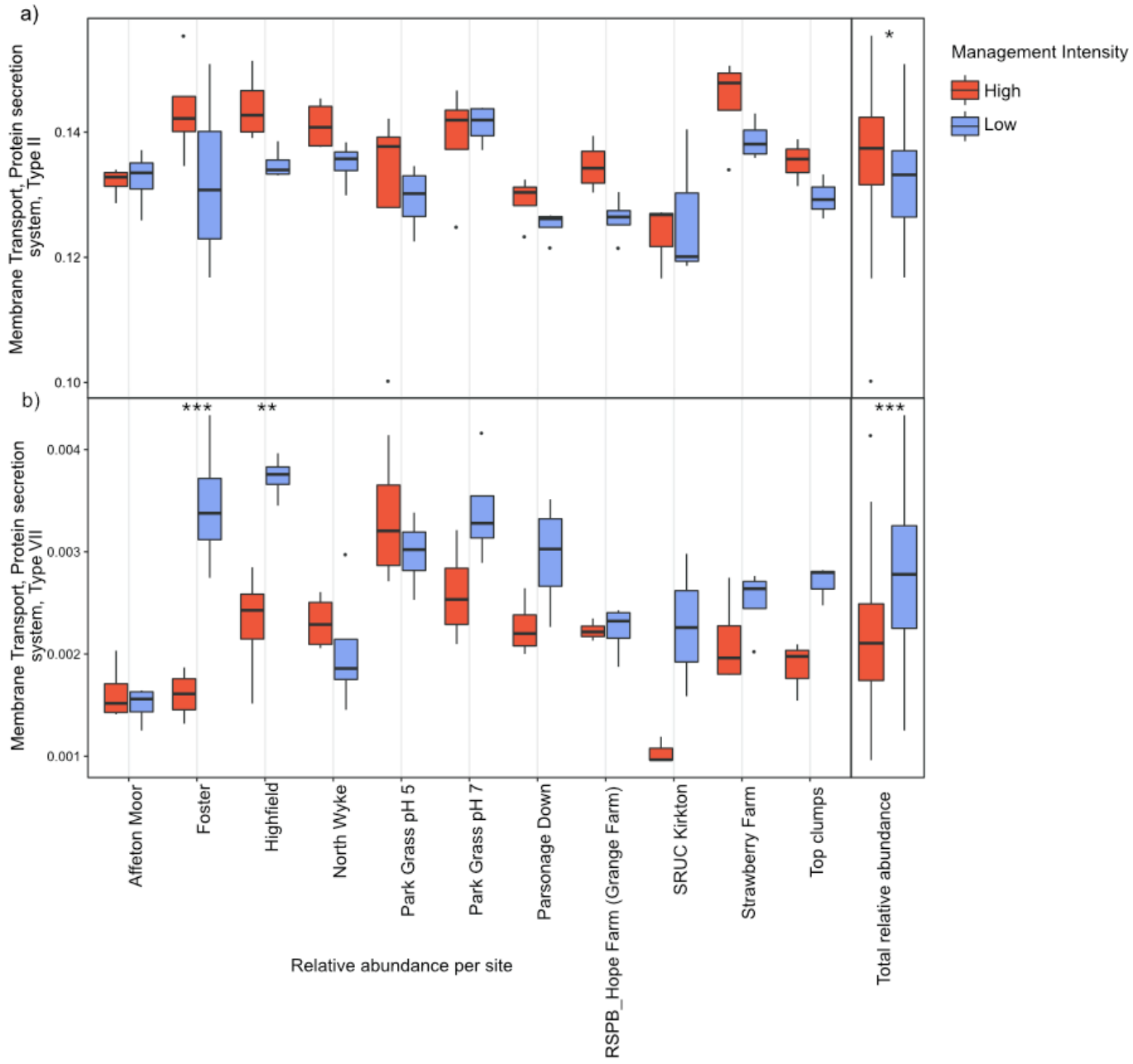


Fig.3.5. Boxplots of subsystems abundance across sample sites. These plots are representative of a sub selection of genomic variables shown to be important in SEED subsystem (level 1) based random forest model, in respect to model accuracy. Statistical differences between high and low management intensities (total relative abundance) was determined using linear modelling (relative subsystem abundance as dependent variable and management intensity and site as covariates with interaction) and a two way anova. Statistical differences between sites was determined using a post hoc test (TukeyHSD). * denotes pval < 0.05, ** pval < 0.01, *** pval < 0.001, blank denotes pval > 0.05.

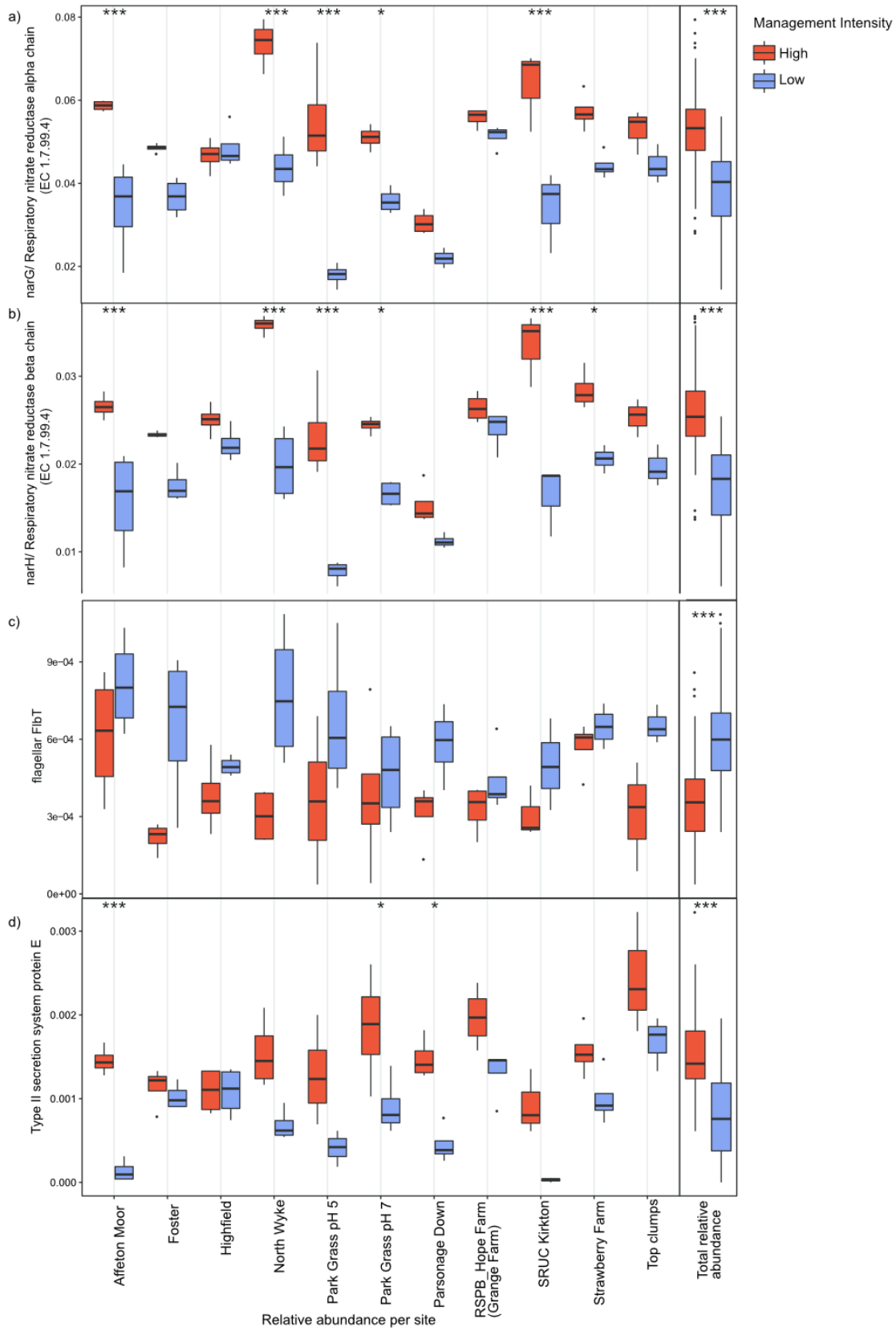


Fig.3.6. Boxplots of gene abundance across sample sites. These plots are representative of a sub selection of genomic variables shown to be important in the gene based random forest model, in respect to model accuracy. Statistical differences between high and low management intensities (total relative abundance) was determined using linear modelling (relative gene abundance as dependent variable and management intensity and site as covariates with interaction) and a two way anova. Statistical differences between sites was determined using a post hoc test (TukeyHSD). * denotes pval < 0.05, ** pval < 0.01, *** pval < 0.001, blank denotes pval > 0.05.

3.4 Discussion

This work found that whilst pH was a major driver of functional profiles (as previously reported (Malik et al., 2018)), organic matter and land use also appear influential. Similarly, although a subset of genomic variables correlated with pH, there were other variables correlating with moisture and organic matter. Random forest (RF) models, both on the gene and subsystem level had a reasonably good rate of classifying the samples land use intensity in an unseen subset of the data (76.9% and 73% respectively). Suggesting specific genomic variables are effective at discriminating between varying land use intensity.

The respiratory nitrate reductase complex (Nar) was important to the gene level RF's accuracy. Nar converts nitrate to nitrite, which is the first step of denitrification, which can result in the further reduction to nitic and nitrous oxide (catalysed by enzyme complexes NirS/NirK and NorB) which are both potent greenhouse gases, alternatively nitrite can be exported out of the cell by the nitrite extrusion protein NarK. I found that the respiratory nitrate alpha subunit (NarG, the complexes catalytic subunit), and the respiratory nitrate beta subunit (NarH, electron transfer subunit (Bertero et al., 2003; Moreno-Vivián, Cabello, Martínez-Luque, Blasco, & Castillo, 1999)) were particularly important to the models accuracy. These two subunits form a complex together with Narl, which links the complex to the inner membrane. Both NarG and NarH were found in significantly higher total relative abundance in high intensity soils compared to low intensity soils and consistently showed increased levels within high intensity samples across sites. These results are likely to be related to the nitrogen based fertilisers used within the arable fields studied, given that the application of nitrogen fertiliser has been associated with large losses of nitrogen to the environment via increased denitrification rates (Philippot, Hallin, & Schloter, 2007; Zumft, 1997).

Through my random forest analyses on the subsystem level I found that both T2SS and T7SS were important in distinguishing between high and low intensity land uses, generally finding increased levels of T2SS in high intensity and increased levels of T7SS within low intensity soils. T2SS are found in Gram negative bacterium whilst T7SS are found in Gram positive, given Gram positive and negative bacteria differ extensively in respect to cell structure (e.g Gram positive bacterium often contain a dense lipid layer termed a mycomembrane),

different cellular machinery (and thus secretory systems) are needed in order to expel substrates (Green & Mecsas, 2016). Further, within the gene level random forest, a component of T2SS, Type 2 secretion protein E (GsPE) was also found to both be important to model accuracy and found in increased levels within high intensity land use (consistent with the trend seen in broad T2SS subsystem relative abundance). GsPE is an ATPase that sits in the cytosol close to the inner membrane and helps form the pseudopillis, a piston like structure which expel proteins out of the cell and is therefore key to T2SS functioning (McLaughlin, et al., 2012; Patrick, et al., 2014). Together these findings indicate increased Gram positive bacterial secretion in low intensity soils and increased Gram negative bacterial secretion in high intensity soils. These results are most likely indicative of land use related shifts in microbial community composition, which were apparent within the 16S analyses. Other cellular machinery genes were also identified as important variables within the random forest models including the flagellum FlbT gene which was consistently found in higher relative abundance in low intensity soils (with significantly increased total relative abundance in low intensity soils overall). Flagellum play a key role in motility , whilst also being involved in wider cellular functioning such as mechanosensing, indeed it has previously been reported that flagellum are able to sense wetness in the environment to control their own biogenesis (Wang, et al., 2005). The specific gene highlighted in the random forest analyses, FlbT is a post-transcriptional inhibitor, inhibiting the translation of the FliK transcript, a flagellum protein responsible for controlling the length of the flagellum hook (Anderson, Smith, & Hoover, 2010; Mangan et al., 1999). Whilst land use induced shifts in soil conditions could explicitly be selecting for motility / sensing traits, equally these results may just be indicative of land use induced taxonomic changes leading to the subsequent changes in types of cellular structures found.

When genomic variables that were identified as important to random forest model accuracy were plotted, less variables acted consistently on the broad SEED subsystem level in comparison to those on the gene level (including genomic variables not shown here). This may be due to the fact that whilst these variables are important to model accuracy they may be more context dependent, relying on interactions with other subsystems, i.e. they may not act consistently in the same land use across sites and instead depend upon other subsystem responses. SEED subsystems may also be too broad to be good determinants of

land use; indeed there will likely be variation in abundance of genes categorised under a subsystem. It is also worth considering when conducting analyses on the subsystem level that there are likely to be biases in the systems used to catalogue genes into these subsystems in the first instance. However regardless of which gene ontology classification system is used (SEED (Overbeek et al., 2005), KEGG (Ogata et al., 1999), MetaCyc (Caspi et al., 2016), COG (Kristensen et al., 2010), GO (Gene Ontology Consortium 2000) etc.) there will invariably be biases introduced whether these systems are manually curated or classified by algorithms. Despite this gene ontology systems remain a useful reference point and are valuable for understanding the big picture of what is occurring within genomic data. This work does however highlight that there is a need for a better understanding of what many genes are doing within the environmental context, and what their documented function means within soil systems. Many of the land use associated genomic variables identified in this work are difficult to contextualise within soils. Whilst nitrogen cycling genes are very interpretable in soils, other genes including those related to molecular or cellular machinery, are harder to interpret and less directly linked to soil functional processes. There is a need therefore for more environmentally specific gene and ontology databases for improved interpretation of these large metagenomic datasets.

It worth noting that the normal limitations of relative abundance methods apply to this work; it's very hard to get full coverage of a sample when sequencing thus it's not possible to say the relative abundance of reads in a sample corresponds directly to the actual relative abundance of genes in the soil. Further, the approach taken here i.e. annotating and analysing metagenomes based on short reads only provides potential links between functions and land use intensity and does not provide insights into the likely taxa conducting these functions. It would be beneficial therefore for future work to assemble soil metagenome data to gain more contiguous information in order to map taxa to function. Understanding a phylotype's functional traits alongside their responses to environmental drivers is valuable as it can help us begin to assess whether bacteria that respond similarly to environmental drivers are also similar functionally thus enabling insights into how resilient soil systems are to land use and wider environmental change.

3.5 Conclusions

This work suggests land use is influential on microbial functional profiles and that there are consistent functional genomic characteristics of high and low management intensity soils, irrespective of pH and other subtleties in land use management regimens. It also suggests the importance of nitrate reductases in distinguishing land use; since I observed a higher relative abundance of nitrate reductase genes within high intensity land use possibly due to the application of nitrogen based fertilizers. This work also highlights the usefulness of large-scale studies employing metagenomics approaches to identify potentially relevant soil functions affected by land use. However, it also demonstrates a clear need for better functional categorical systems that groups genes into soil relevant functions rather than broad, context neutral categories that are hard to interpret. Further, there is now a need to focus on applying assembly techniques to map taxa to function, in order to gain insights into not only how microbes respond to environmental change but also which functions they are likely delivering, in order to better assess the resilience of soil microbial communities under different land use conditions, to better inform farmers and policy makers decisions.
This dataset is therefore explored in greater detail within the next two chapters (**Chapters 4 and 5**) using assembly based approaches to link specific taxa with functional capacities. **Chapter 4** assembles organic matter decomposition enzyme sequences from metagenomes to gain insights into taxonomically associated shifts in carbon cycling in response to pH, whilst **Chapter 5** uses a novel metagenomic assembly and binning approach to link land use induced shifts in functional genes (of relevance to various soil processes) to specific taxonomic groupings.

3.6 Bibliography

Anderson, J. K., Smith, T. G., & Hoover, T. R. (2010). Sense and sensibility: flagellum-mediated gene regulation. *Trends in Microbiology*, *18*(1), 30–37. https://doi.org/10.1016/j.tim.2009.11.001

Aziz, R. K., Devoid, S., Disz, T., Edwards, R. A., Henry, C. S., Olsen, G. J., … Xia, F. (2012). SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. *PLoS ONE*, *7*(10), 1–10. https://doi.org/10.1371/journal.pone.0048053

Badagliacca, G., Benítez, E., Amato, G., Badalucco, L., Giambalvo, D., Laudicina, V. A., & Ruisi, P. (2018). Long-term effects of contrasting tillage on soil organic carbon, nitrous oxide and ammonia emissions in a Mediterranean Vertisol under different crop sequences. *Science of the Total Environment*, *619–620*, 18–27. https://doi.org/10.1016/j.scitotenv.2017.11.116

Banerjee, S., Walder, F., Büchi, L., Meyer, M., Held, A. Y., Gattinger, A., … van der Heijden, M. G. A. (2019). Agricultural intensification reduces microbial network complexity and the abundance of keystone taxa in roots. *ISME Journal*, *13*(7), 1722–1736. https://doi.org/10.1038/s41396-019-0383-2

Barberán, A., Bates, S. T., Casamayor, E. O., & Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME Journal*, *6*(2), 343–351. https://doi.org/10.1038/ismej.2011.119

Bayer, C., Gomes, J., Zanatta, J. A., Vieira, F. C. B., Piccolo, M. de C., Dieckow, J., & Six, J. (2015). Soil nitrous oxide emissions as affected by long-term tillage, cropping systems and nitrogen fertilization in Southern Brazil. *Soil and Tillage Research*, *146*(PB), 213–222. https://doi.org/10.1016/j.still.2014.10.011

Bertero, M. G., Rothery, R. A., Palak, M., Hou, C., Lim, D., Blasco, F., … Strynadka, N. C. J. (2003). Insights into the respiratory electron transfer pathway from the structure of nitrate reductase A. *Nature Structural Biology*, *10*(9), 681–687. https://doi.org/10.1038/nsb969

Boetius, A. (2019). Global change microbiology — big questions about small life for our future. *Nature Reviews Microbiology*, *17*(6), 331–332. https://doi.org/10.1038/s41579-019-0197-2

Breiman, L. (2001). No Title, 1–33.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., … Karp, P. D. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, *44*(D1), D471–D480. https://doi.org/10.1093/nar/gkv1164

Consortium, T. G. O. (2000). Gene ontology: Tool for the identification of biology. *Natural Genetics*, *25*(may), 25–29.

Daniel, R. (2005). The metagenomics of soil. *Nature Reviews. Microbiology*, *3*(6), 470–478. https://doi.org/10.1038/nrmicro1160

Daims, H. et al. (2015) 'Complete nitrification by Nitrospira bacteria', Nature. Nature Publishing Group, 528(7583), pp. 504–509. doi: 10.1038/nature16461.

Fierer, N., & 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), 626–631. https://doi.org/10.1073/pnas.0507535103

Foley, J. a, Defries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., … Snyder, P. K. (2005). Global consequences of land use. *Science (New York, N.Y.)*, *309*(5734), 570–574. https://doi.org/10.1126/science.1111772

Goss-Souza, D., Mendes, L. W., Borges, C. D., Baretta, D., Tsai, S. M., & Rodrigues, J. L. M. (2017). Soil microbial community dynamics and assembly under long-term land use change. *FEMS Microbiology Ecology*, *93*(10). https://doi.org/10.1093/femsec/fix109

Green, E. R., & Mecsas, J. (2016). Bacterial Secretion Systems: An Overview. *Virulence Mechanisms of Bacterial Pathogens, Fifth Edition*, *4*(1), 215–239. https://doi.org/10.1128/microbiolspec.vmbf-0012-2015

Griffiths, B. S., & Philippot, L. (2013, March 1). Insights into the resistance and resilience of the soil microbial community. *FEMS Microbiology Reviews*. Oxford Academic. https://doi.org/10.1111/j.1574-6976.2012.00343.x

Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., & Whiteley, A. S. (2011). The bacterial biogeography of British soils. *Environmental Microbiology*, *13*(6), 1642–1654. https://doi.org/10.1111/j.1462-2920.2011.02480.x

Grządziel, J. (2017). Functional redundancy of soil microbiota – Does more always mean better? *Polish Journal of Soil Science*, *50*(1), 75–81. https://doi.org/10.17951/pjss.2017.50.1.75

Guo, L. B., & Gifford, R. M. (2002). Soil carbon stocks and land use change: A meta analysis. *Global Change Biology*, *8*(4), 345–360. https://doi.org/10.1046/j.1354-1013.2002.00486.x

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews : MMBR*, *68*(4), 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hartmann, M., Frey, B., Mayer, J., Mäder, P., & Widmer, F. (2015). Distinct soil microbial diversity under long-term organic and conventional farming. *ISME Journal*, *9*(5), 1177–1194. https://doi.org/10.1038/ismej.2014.210

Huang, R., McGrath, S. P., Hirsch, P. R., Clark, I. M., Storkey, J., Wu, L., … Liang, Y. (2019). Plant–microbe networks in soil are weakened by century-long use of inorganic fertilizers. *Microbial Biotechnology*, *12*(6), 1464–1475. https://doi.org/10.1111/1751-7915.13487

Jansson, J. K., & Hofmockel, K. S. (2020). Soil microbiomes and climate change. *Nature Reviews Microbiology*, *18*(1), 35–46. https://doi.org/10.1038/s41579-019-0265-7

Jia, Y., & Whalen, J. K. (2020). A new perspective on functional redundancy and phylogenetic niche conservatism in soil microbial communities. *Pedosphere*, *30*(1), 18–24. https://doi.org/10.1016/S1002-0160(19)60826-X

Jones, D. L., Cooledge, E. C., Hoyle, F. C., Griffiths, R. I., & Murphy, D. V. (2019). pH and exchangeable aluminum are major regulators of microbial energy flow and carbon use efficiency in soil microbial communities. *Soil Biology and Biochemistry*, *138*(August), 0–4. https://doi.org/10.1016/j.soilbio.2019.107584

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.01043-13

Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, *26*(12), 1481–1487. https://doi.org/10.1093/bioinformatics/btq229

Kroeger, M. E., Delmont, T. O., Eren, A. M., Meyer, K. M., Guo, J., Khan, K., … Nüsslein, K. (2018). New biological insights into how deforestation in amazonia affects soil microbial

communities using metagenomics and metagenome-assembled genomes. *Frontiers in Microbiology*, *9*(JUL), 1–13. https://doi.org/10.3389/fmicb.2018.01635

Leff, J. W., Jones, S. E., Prober, S. M., Barberán, A., Borer, E. T., Firn, J. L., … Fierer, N. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), 10967–10972. https://doi.org/10.1073/pnas.1508382112

Li, X., Jousset, A., de Boer, W., Carrión, V. J., Zhang, T., Wang, X., & Kuramae, E. E. (2019). Legacy of land use history determines reprogramming of plant physiology by soil microbiome. *ISME Journal*, *13*(3), 738–751. https://doi.org/10.1038/s41396-018-0300-0

Malik, A. A., Puissant, J., Buckeridge, K. M., Goodall, T., Jehmlich, N., Chowdhury, S., … Griffiths, R. I. (2018). Land use driven change in soil pH affects microbial carbon cycling processes. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-05980-1

Mangan, E. K., Malakooti, J., Caballero, A., Anderson, P., Ely, B., & Gober, J. W. (1999). FlbT couples flagellum assembly to gene expression in Caulobacter crescentus. *Journal of Bacteriology*, *181*(19), 6160–6170. https://doi.org/10.1128/jb.181.19.6160-6170.1999

Maskell, L. C., Crowe, A., Dunbar, M. J., Emmett, B., Henrys, P., Keith, A. M., … Smart, S. M. (2013). Exploring the ecological constraints to multiple ecosystem service delivery and biodiversity. *Journal of Applied Ecology*, *50*(3), 561–571. https://doi.org/10.1111/1365-2664.12085

McLaughlin, L. S., Haft, R. J. F., & Forest, K. T. (2012). Structural insights into the Type II secretion nanomachine. *Current Opinion in Structural Biology*, *22*(2), 208–216. https://doi.org/10.1016/j.sbi.2012.02.005

Moreno-Vivián, C., Cabello, P., Martínez-Luque, M., Blasco, R., & Castillo, F. (1999). Prokaryotic nitrate reduction: Molecular properties and functional distinction among bacterial nitrate reductases. *Journal of Bacteriology*, *181*(21), 6573–6584.

Muyzer, G., De Waal, E. C., & 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), 695–700. https://doi.org/10.1128/aem.59.3.695-700.1993

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *27*(1), 29–34. https://doi.org/10.1093/nar/27.1.29

Oksanen, J. et al. (2018) 'vegan: Community Ecology Package'. Available at: https://cran.r-project.org/package=vegan.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., … Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, *33*(17), 5691–5702. https://doi.org/10.1093/nar/gki866

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., … Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(D1), 206–214. https://doi.org/10.1093/nar/gkt1226

Patrick, M., Shah, R., Sandkvist, M., Bush, M. F., & Hol, W. G. J. (2014). Hexamers of the Type II Secretion ATPase GspE from Vibrio cholerae with Increased ATPase Activity, *21*(9), 1707–1717. https://doi.org/10.1016/j.str.2013.06.027.Hexamers

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G. P., & Smith, P. (2016). Climate-smart soils. *Nature*, *532*(7597), 49–57. https://doi.org/10.1038/nature17174

Pershina, E., Valkonen, J., Kurki, P., Ivanova, E., Chirak, E., Korvigo, I., … Andronov, E. (2015). Comparative analysis of prokaryotic communities associated with organic and conventional farming systems. *PLoS ONE*, *10*(12), 1–16. https://doi.org/10.1371/journal.pone.0145072

Philippot, L., Hallin, S., & Schloter, M. (2007, January 1). Ecology of Denitrifying Prokaryotes in Agricultural Soil. *Advances in Agronomy*. Academic Press. https://doi.org/10.1016/S0065-2113(07)96003-4

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191–218. https://doi.org/10.7155/jgaa.00124

Schöps, R., Goldmann, K., Herz, K., Lentendu, G., Schöning, I., Bruelheide, H., … Buscot, F. (2018). Land-use intensity rather than plant functional identity shapes bacterial and fungal rhizosphere communities. *Frontiers in Microbiology*, *9*(NOV), 2711. https://doi.org/10.3389/fmicb.2018.02711

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & Sacha van Hijum, A. F. T. (2013). Data mining in the life science swith random forest: A walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326. https://doi.org/10.1093/bib/bbs034

Trivedi, P., Delgado-baquerizo, M., Trivedi, C., Hu, H., Anderson, I. C., Jeffries, T. C., … Singh, B. K. (2016). Microbial regulation of the soil carbon cycle : evidence from gene – enzyme relationships. *The ISME Journal*, *10*(11), 1–12. https://doi.org/10.1038/ismej.2016.65

Turner, B. L. (2010). Variation in ph optima of hydrolytic enzyme activities in tropical rain forest soils. *Applied and Environmental Microbiology*, *76*(19), 6485–6493. https://doi.org/10.1128/AEM.00560-10

Van Kessel, M. A. H. J., Speth, D. R., Albertsen, M., Nielsen, P. H., Op Den Camp, H. J. M., Kartal, B., … Lücker, S. (2015). Complete nitrification by a single microorganism. *Nature*, *528*(7583), 555–559. https://doi.org/10.1038/nature16459

Wang, Q., Suzuki, A., Mariconda, S., Porwollik, S., & Harshey, R. M. (2005). Sensing wetness: A new role for the bacterial flagellum. *EMBO Journal*, *24*(11), 2034–2042. https://doi.org/10.1038/sj.emboj.7600668

Xia, F., Wang, J. G., Zhu, T., Zou, B., Rhee, S. K., & Quan, Z. X. (2018). Ubiquity and diversity of complete ammonia oxidizers (comammox). *Applied and Environmental Microbiology*, *84*(24), 1–14. https://doi.org/10.1128/AEM.01390-18

Yu, Y., Lee, C., Kim, J., & Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnology and Bioengineering*. https://doi.org/10.1002/bit.20347

Zumft, W. G. (1997). Cell biology and molecular basis of denitrification. *Microbiology and Molecular Biology Reviews : MMBR*, *61*(4), 533–616. https://doi.org/10.1128/.61.4.533-616.1997

Chapter 4

Functional relevance of soil pH change for carbon cycling enzyme taxonomic producers and genetic diversity

Abstract

Soil pH is known to affect the biodiversity of soil microbes, though how pH related taxonomic change translates to altered soil processes remains uncertain. Extracellular enzyme (EE) production is an important functional trait of microbes for degrading complex carbohydrates to simple compounds for growth and metabolism, and is thought to be a key modulator of soil organic matter decomposition. Recently Puissant et al (2019) demonstrated using physiological assays that the pH activity optima of β-glucosidases (responsible for hydrolysing glucosides to glucose) is altered by long term changes in soil pH. Here I seek to further explore the genomic basis of the pH related differences in enzyme activity by interrogating metagenomic assemblies for β-glucosidase sequences from the same soils. Firstly, a much larger proportion of *Acidobacterial* β-glucosidase sequences were found at pH5 compared to pH7, which infers that different microbial taxa may produce specific pH adapted enzymes. Relatedly, there were also large pH related differences in specific glycoside hydrolase family abundances, with pH5 soils exhibiting a greater proportion of GH2 annotated sequences (mainly *Acidobacteria* produced), and pH7 soils showing elevated GH1 annotated sequences (mainly *Actinobacteria*, *Proteobacteria* and *Verrucomicrobia* produced). In attempts to assess how other genetic features of the enzymes were effected by soil pH, a phylogenetic analyses of the most abundant and ubiquitous glycoside hydrolase family (GH3) was performed. Here, it was found that phylum was highly influential on enzyme sequence and to a lesser extent the presence of signal peptides. In addition, analyses of the presence of signal peptides revealed that *Acidobacterial* enzymes associated with β-glucosidase activity may be more likely to be secreted than enzymes from other dominant soil phyla. Together this work identifies *Acidobacteria* as playing a key role in extracellular enzyme secretion in acid soils, and further identifies new avenues for research into the functionality of novel enzymes discovered through metagenomic sequencing.

4.1 Introduction

Microbial communities are key players in the carbon cycle whereby they both sequester carbon into more stable forms of organic matter and mineralize carbon into soil organic matter, resulting in the release of CO_2 (Jansson and Hofmockel, 2018). A large amount of organic matter (OM) decomposition is conducted by extracellular enzymes (EE) secreted by bacteria and fungi which degrade complex carbohydrates derived from plants and microbes into simple compounds which are in turn used in microbial growth and metabolism (Allison, 2005; Burns et al., 2013; Cantarel et al., 2009; German et al., 2011). Given the importance of EE's to the carbon cycle, there is increasing interest in quantifying enzyme activity in different soil types in order to build more process-informed carbon models (Allison, 2014, 2012; Wang et al., 2013). Typically, measurement of enzymatic processes relies on quantifying the rate of the total enzyme pool in an environment through the use of assays measuring changes within a chosen substrate. As such, typical quantification of enzyme activity can be considered a "black box" approach, whereby we don't know the underlying mechanisms of variation in activity nor the microbial contributors to the enzymes we are measuring. Improving understanding of the influence of specific microbial producers may allow for increased understanding of process mechanisms, and better prediction of rates within specific environments where there is already understanding of the microbial community.

The widespread use of high-throughput sequencing has enabled us to develop an improved understanding of soil microbial distributions and accurately predict relative abundances of important phylotypes based on co-located soil environmental data (**Chapter 1**) (Fierer and Jackson, 2006; Griffiths et al., 2011). Through using this knowledge, we can begin to ask more advanced questions as to how taxonomic biodiversity affects soil processes, and whether it is possible to predict functional processes based upon taxonomic distributions within that environment. Given the widespread acknowledgement of the importance of extracellular enzymes to soil processes (Dick and Kandeler, 2005; Duly and Nannipieri, 1998), it is crucial that we gain a better understanding of how microbial diversity and community dynamics affect EE process rates. A key question is whether microbes are functionally redundant in respect to EE production and OM decomposition; i.e. do all

microbes possess the capability to produce EE's with equal activity rates under different environmental conditions, or do certain microbes possess EE production traits, which give rise to altered soil nutrient cycling processes? Suggesting the latter, microcosm experiments and reciprocal transplant studies in the field have found that different microbial inoculations altered litter decomposition rates (Allison et al., 2013; Cleveland et al., 2014; Strickland et al., 2009). These studies did not however conduct enzyme assays and therefore the mechanism by which microbes could be affecting OM decomposition was not explicitly addressed. It is also of note that other work using microcosms has shown that microbial inoculation did not impact on organic matter decomposition and concluded there may be functional redundancy amongst microbes with respect to decomposition processes (Banerjee et al., 2016). Improved understanding in this area will likely not only enhance understanding of direct soil decomposition processes but may be informative with respect to our wider understanding of the biological mechanisms that affect process rates. We know for example enzyme production is costly and enzyme products can be opportunistically used by taxa that do not secrete these enzymes and so there are fundamental questions over the wider ecological and evolutionary interactions within communities which may also impact upon process rates (Allison, 2005; Allison et al., 2014). Further many industrial processes make use of microbial extracellular enzymes and so a better understanding of the enzymatic production capacity of microbes in natural environments may permit targeted discovery of novel enzymes for a number of biotechnological purposes (Ahmed et al., 2017; Gangoiti et al., 2018; Srivastava et al., 2019).

Metagenomics studies either targeting whole genomes or specific enzyme genes are emerging as a powerful approach to better understand the ecology of EE producers in the natural environment. For example metagenomics has been used to show that cellulase gene content can be predicted from microbial community composition (genus level) in semi-arid grasslands (Berlemont et al., 2014). This finding is consistent with results from large metagenome meta-analyses across soil, marine, human and animal microbiomes which also found genus level association with OM degrading enzyme gene content, with most carbohydrate degrading genes found in just 77 genera (Berlemont and Martiny, 2016). Other work has looked at the distribution of important OM degradation enzymes across sequenced bacterial genomes and found that though β-glucosidases are widespread across

bacteria (in 80% percent of ~5000 sequenced taxa), taxa within the same genera had more similar β-glucosidase gene content in terms of the glycoside hydrolase families present. Additionally the majority of taxonomic groups did not contain all enzyme encoding genes necessary to conduct complete cellulose degradation and instead are likely to rely upon scavenging disaccharides produced by other organisms to obtain a glucose supply (Berlemont and Martiny, 2013). Another question in this area is whether phylogenetically related clades demonstrate high sequence similarity within EE genes. Whilst relatively little work has been done in this area, qPCR of enzymes involved in the early cellulose degradation has shown little agreement between species and enzyme phylogeny (Merlin et al., 2014), suggesting the possibility that OM enzyme sequences are commonly exchanged between taxa.

Understanding how EE producers are affected under different abiotic conditions is critical, particularly in a time of rapid land use and climate change. The adaption of extracellular enzymes to soil temperature is widely studied, and its known some EEs are particularly efficient at cold temperatures, possibly due to increased flexibility within the active site or protein surface (Lonhienne et al., 2000; Zanphorlin et al., 2016). It's also been suggested that shifts in temperature can result in community level transitions from cold to warm adapted species/genotypes (Bradford, 2013; Wei et al., 2014) as well as physiological changes within individual taxa, resulting in different isoenzymes being expressed (Bradford, 2013). EE responses to pH have been less well studied, which is surprising given the relationship between pH change and land use transition, whereby the application of ammonium based fertilizers are known to contribute to soil acidification (Goulding, 2016; Tian and Niu, 2015) and liming is used to intentionally neutralize acidity. The way we manage our soils clearly affects soil pH (Malik et al., 2018) and understanding how this affects core soil functions including organic matter decomposition is of great value to farmers and policy makers. In terms of enzyme kinetics, it is well established that general enzyme activity is sensitive to pH and that enzymes have a specific pH where they operate most efficiently, termed a pH optimum (German et al., 2011). However, whilst there have been studies speculating some soil EE pH optima may vary according to soil pH (Niemi and Vepsäläinen, 2005; Turner, 2010) this has never been critically evaluated. Furthermore, whist we know taxonomic biodiversity of microbes is strongly influenced by pH (Fierer and

Jackson, 2006; Griffiths et al., 2011) along with specific carbon cycling processes (Jones et al., 2019); we have little understanding as to the relationships between pH and EE producers and the subsequent impact on EE activity rate.

4.1.1 Chapter Aims

Recent work at UKCEH as part of the UGRASS project on long term pH manipulated plots at Rothamsted demonstrated that for a number of EE, the pH optima was shifted in the direction of source soil pH (Puissant et al., 2019). Within this article, I contributed bioinformatics analyses of metagenomics assemblies, to show that these changes in pH optima were also accompanied by changes in the communities of EE producing bacteria. These bioinformatics analyses focussed on β-glucosidase which is involved in late stage cellulose degradation through hydrolysing the glycosidic bonds of glucosides (e.g. cellobiose) to produce glucose. β-glucosidase was chosen specifically since it is thought to be a key enzyme responsible for soil OM processing and also exhibited large changes in efficiency at different pH sites within the UGRASS experiments (**Fig.4.1**) (Puissant et al., 2019), suggesting soil pH is heavily influential on β-glucosidase activity. β–glucosidases are also a commonly used soil health indicator (Bandick and Dick, 1999), are well characterized enzymes (in terms of sequence and structure) (Henrissat et al., 1995); and relevant HMMER profiles are available based upon CAZY (Carbohydrate-Active enZymes) database classifications (Cantarel et al., 2009; Zhang et al., 2018) .

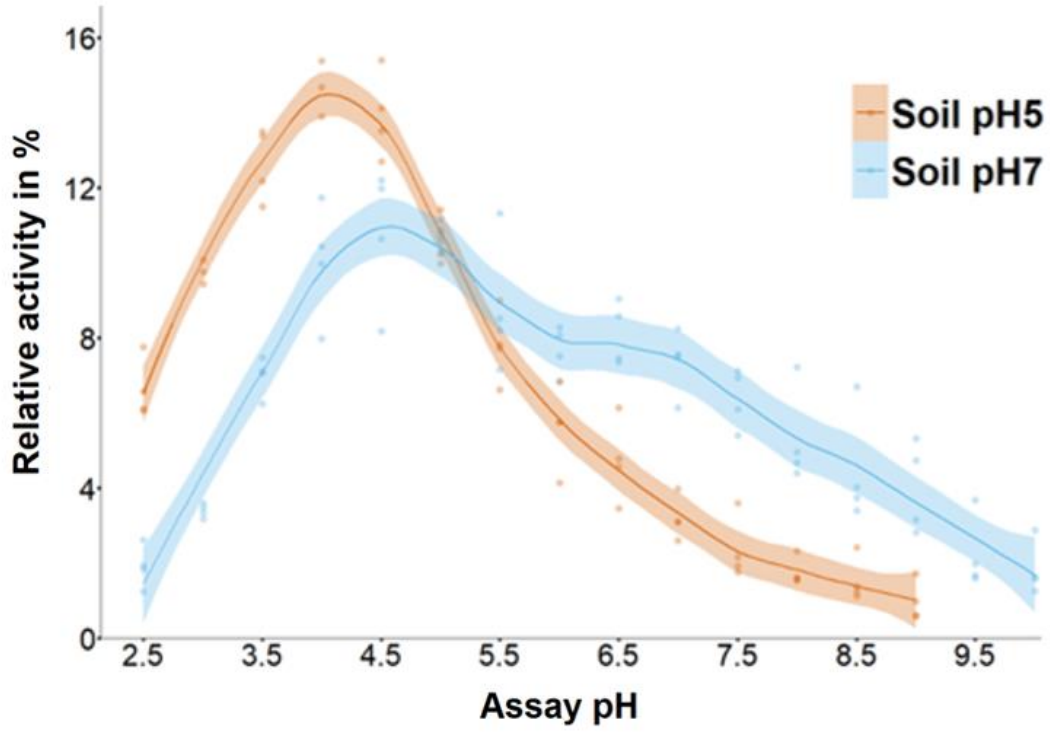


Fig.4.1. β-glucosidase activity from grassland soils maintained at either pH5 or 7 assayed at different pH levels (from Puissant et al., 2019). Activity is expressed as a percentage of the total activity measured across the pH2.5 -10 range assayed. Orange and blue lines correspond to pH5 and pH7 soils respectively. Shaded area represents 95% confidence intervals around the trend line generated using LOESS smoothing.

The following chapter describes in more detail the metagenomics analyses that was included in the paper (attached in **appendix 1**) and also includes some follow up work which did not appear in the article, exploring how pH affects soil C cycling enzyme genes beyond the genes taxonomic assignment.

The key aims were to:

(i) **Determine how specific CAZY families vary with pH**: CAZY families are classified by sequence and structure opposed to function, and therefore it is possible for families to have multiple activities and for multiple families to perform the same activity. For this reason I examined shifts in specific families linked to β–glucosidase, to gain further understanding of the exact enzymatic activities occurring at each pH.

(ii) **Examine the amino acid composition of sequences and their relationship with pH**: Enzyme sequences are known to exist in alternate forms termed isoenzymes where they vary in terms of sequence but not enzymatic activity. I therefore analysed amino acid composition across pHs to determine whether there was evidence of pH induced isoenzymes that could explain differences in enzyme activity observed within UGRASS assays.

(iii) **Investigate how phylogeny corresponds to taxonomy**: To investigate the redundancy of β–glucosidase sequences within microbial communities, through examining whether phylogeny of β–glucosidases matched that of taxonomic phylogeny. Specifically, I aimed to determine whether there was evidence of taxonomic related isoenzymes, or whether β–glucosidase phylogeny suggests high rates of horizontal gene transfer and sequence redundancy.

(iv) **Establish whether β–glucosidases annotated are secreted:** Given β–glucosidases exist both extracellularly and intracellularly, I wished to distinguish between these two groups using signal peptide annotation to identify which sequences are most relevant to the UGRASS extracellular assays conducted. Further to this I aimed to determine whether the sequences of extracellular and intracellular β–glucosidases made them inherently distinguishable, with this having the potential to be informative to future annotation methods.

(v) **Which factors are most determinant over β–glucosidase sequence diversity soil pH, cellular location (secreted or not) or taxonomy of producer:** and the implications this could have on organic matter decomposition given different abiotic and biotic scenarios.

4.2 Methods

4.2.1 Study site

Soils were taken from the long term Park grass experiment based at Rothamsted research, whereby soils have been maintained at pH5 and 7 for over 100 years. The experiment was initially setup in 1856 to understand how the application of different fertilizers affect yield from hay meadows. The original plots were later divided in 1903 and subjected to different

pH treatments (Silvertown et al., 2006). For this study soil cores (15cm depth, 4cm diameter) were sampled in November 2015, from 'Nil plot 12' which has never received fertilizer treatment. Soil pH has been controlled by liming whereby 'subplot a' has been kept at ~ pH7 since 1903 (limed every four years until 1976 and every three years since), and 'subplot C' has been kept at ~pH5 since 1965 (limed every three years). As the natural soil pH is already between 5.4-5.6, liming of 'Subplot C' has been minimal and primarily used to mitigate natural soil acidification. Surface litter was removed from soil cores. Soil cores were homogenised wet without sieving prior to subsampling for DNA extraction.

4.2.2 Metagenome Sequencing

DNA was extracted from 2g of soil from 4 field replicates for the two pH treatments using the PowerMax Soil DNA Isolation kit (Qiagen) and subsequently concentrated and purified using Amicon® Ultra filters. Illumina libraries were constructed using the Illumina TruSeq library preparation kit (insert size < 500- 600 bp) and paired-end sequencing (2 x 150 bp) was conducted using the Illumina HiSeq 4000 platform. Prior to annotation, Illumina adapters were removed from raw fastq files using Cutadapt 1.2.1 (Martin, 2011), reads were trimmed using Sickle (Joshi et al., 2011) with a minimum window quality score of 20 and short reads were removed (<20bp). On average there were ~22700000 trimmed reads per sample with an average read length of 148.2 bp.

Preliminary analyses were conducted using MG-RAST to functionally annotate with SEED subsystems and taxonomically annotate with refseq. For more detailed analyses of β-glucosidase sequences, all reads from the 8 samples were co-assembled using MEGAHIT (Li et al., 2015) with a minimum contig length of 1000 bp resulting in 576612 contigs and an average length of 1925 bp. Sequences were translated and open reading frames were predicted using FragGeneScan (Rho et al., 2010). Contigs were assigned CAZY (Carbohydrate-Active enZYmes) families (Lombard et al., 2014) using a HMMER search (Finn et al., 2011) against dbCAN2 profiles with an eval of 1e-15 (Zhang et al., 2018).
Of the 576612 contigs, there were a total of 23238 annotations to CAZY domains, 1314 of these were annotations to β-glucosidase associated CAZY families (GH1, GH2, GH3, GH5, GH9, GH30, GH39 and GH116).

Contigs were taxonomically annotated against the NCBI Blast non-redundant protein database using Kaiju, a fast translated method, which identifies protein-level maximum exact matches (MEM's) (Menzel et al., 2016). The taxonomic names assigned by the NCBI database will be used throughout, though we acknowledge these names may vary to those used within the Genome Taxonomy Database (GTDB) (Parks et al., 2018, 2020). Regions of contigs annotated as relevant β-glucosidase CAZY families were extracted.

To identify pH associations of these sequences, DNA reads were mapped back to assembled domain amino acid sequences using BlastX, mappings with an identity percentage of < 97% and/or an evalue of > 0.001 were discarded. Mapping outputs were used to identify the relative abundance of assembled domain sequences across pH5 and pH7 samples, multinomial species classification method (CLAM) (Chazdon et al., 2011) was used to classify pH generalists and specialists and to discount sequences that were too rare to meaningfully categorise.

4.2.3 Annotation of secretory motifs

To determine whether the enzyme genes studied were extracellular opposed to intracellular, the whole contig containing a GH3 annotated region was further mined for secretory motifs using SignalP. SignalP identifies secretory motifs using neural networks and a training data set of 20,758 proteins from UniProtKB/SwissProt with experimental evidence of a cleavage site. It further classifies these secretory motifs by the type of secretory motif /signal peptide present and the signal peptidase (SPase) used to cleave the signal peptide after membrane translocation (Sec/SP1, Sec/SP11 and Tat/SP1) (Almagro Armenteros et al., 2019). SignalP was run with default parameters using Gram positive and Gram negative databases (given 99.7% of taxonomic annotations of sequences were bacterial).
As 94.67% of secretory motif annotations were the same regardless of which database was used, annotations that were consistent for both databases were kept, 2.06% of sequences that were labelled as having different secretory motifs depending on the database used were labelled "secretory mismatch" and those that had an secretory motif annotation in one database and no secretory motif annotation in the other were discarded (3.27%).

Permutational testing was used to test the significance of differences between intracellular and extracellular sequences for each phyla and each pH class using the permute package with 10000 permutations.

4.2.4 CAZY family specific alignments and phylogenetic analyses

β-glucosidase sequences related to the most abundant GH family (GH3) were selected for further phylogenetic analyses. GH3 sequences that had a length of more than 200bp were aligned with ClustalO. A distance matrix was generated from the sequence alignment, which was subsequently analysed using K-means clustering to establish groupings of sequences and Adonis to determine relationships between contig attributes (pH specialism, phyla, and secretory motif annotations) and sequence variance. The alignment was also used to build a phylogenetic tree with FastTree 2.1.7 using neighbour-joining (NJ).

4.3 Results

4.3.1 Taxonomic classification of β-glucosidase sequences

To gain a broad understanding of the phyla contributing to the β-glucosidase sequences at each pH, all metagenomics reads from the soils were subjected to SEED subsystems functional annotation, and taxonomic annotation (Refseq) using the online MG-RAST pipeline (Wilke et al., 2016). By then examining the taxonomy of the β-glucosidase annotated reads, a higher relative abundance of *Acidobacterial* annotated sequences was observed in pH5 soils, and an increased relative abundance of *Actinobacteria* within pH7 (**Fig.4.2a**). When normalized using a housekeeping gene (DNA gyrase subunit B), *Acidobacterial* β-glucosidase were twice as abundant in pH5 soils compared to pH7 (**Fig.4.2b**). This variation in *Acidobacterial* β-glucosidase sequences could therefore play a role in the differences in enzyme efficiency observed in the physiological assays (Puissant et al., 2019 and **Fig.4.1**), but equally its possible subclades at a finer phylogenetic scale could also be influential over enzyme activity.

To address this I sought to examine taxonomic variance at finer scales, using a more resolved enzymatic database (CAZY database using the dbCAN2 pipeline). This was

conducted through assembling contigs from all reads (from pH5 and pH7 soils); extracting sequences annotated to β-glucosidase related CAZY families; and then mapping individual reads back to these sequences. I then classified these contig sequences as pH specialists, generalists or "too rare" to categorise using multinomial species classification method (CLAM). As seen in **Fig.4.3**, the majority of *Acidobacteria* sequences were classed as pH5 specialists, suggesting not only is there a higher relative abundance of *Acidobacteria* β-glucosidase sequences at pH5 but that the majority of these sequences are also more unique to pH5 soils. Sequences annotated as other dominant phyla such as *Actinobacteria* and *Proteobacteria* appeared to have a higher proportion of pH7 specialist and generalist sequences, whilst *Verrucomicrobia* included a clear subclade of pH7 specialist sequences (**Fig.4.3**).

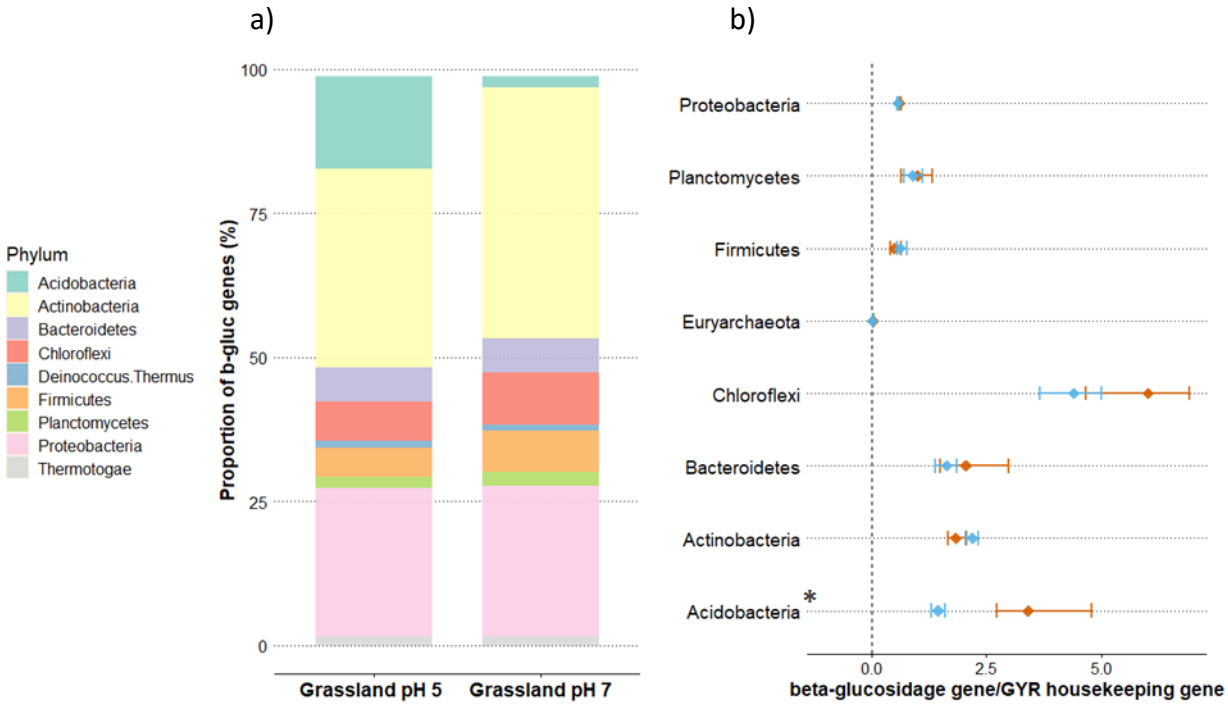


Fig.4.2. Abundances of β-glucosidase genes from different microbial taxa, from MG-RAST annotated metagenomes (SEED Subsystems) (figure from Puissant et al., 2019). **a)** Stacked plot representing the total proportion of β-glucosidase genes from dominant bacterial phylum. **b)** The proportional change of β-glucosidase gene abundance compared to the abundance of the DNA gyrase subunit B gene. Orange and blue colours correspond to pH 5 and pH 7 soil respectively. The x-axis shows the relative fold change on log2 scale. Error bars indicate +/- standard deviation and the means are indicated by filled diamond shape. Asterisks indicate significance difference between pH5 and pH7 soil (ANOVA $p<0.05$).

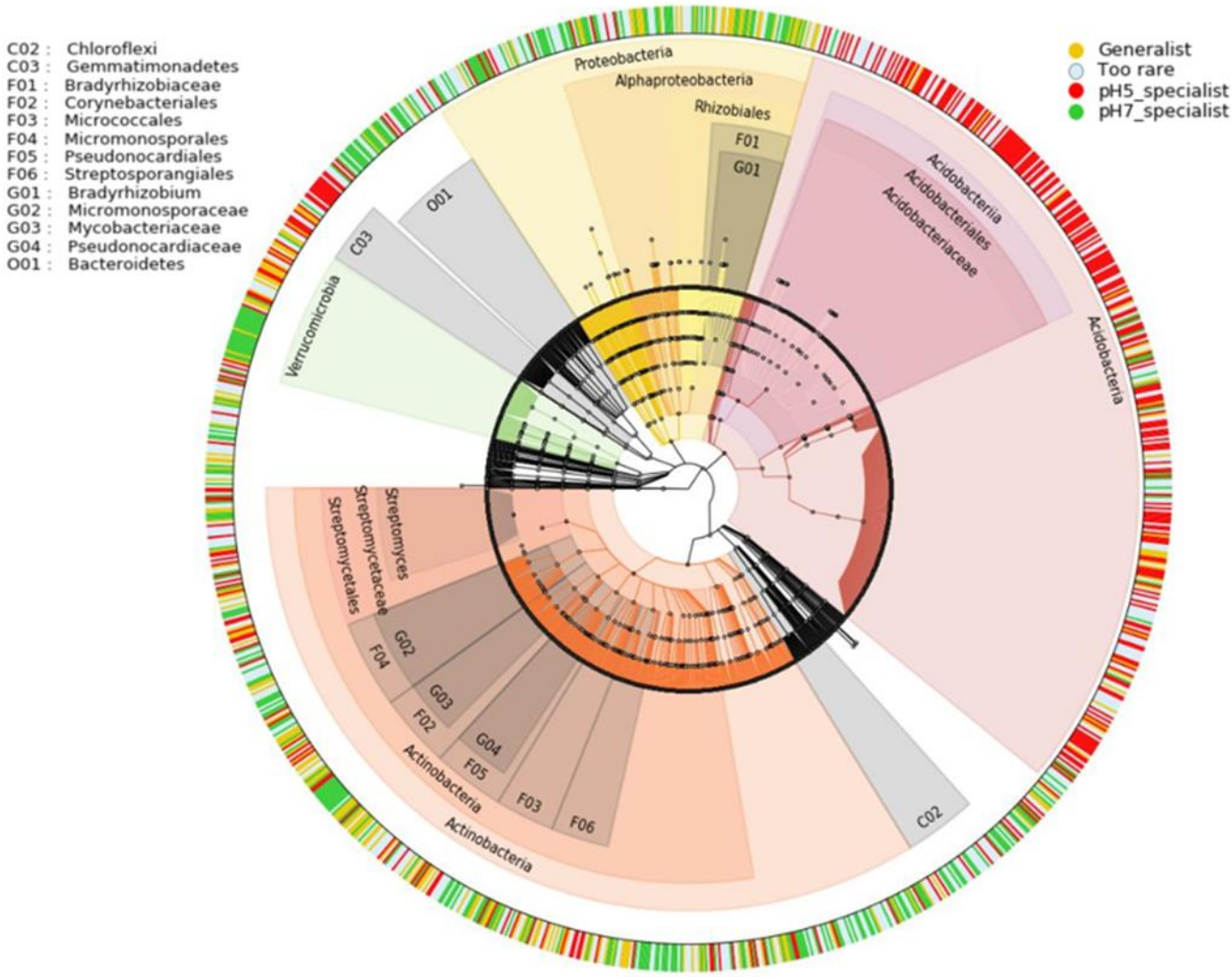


Fig.4.3. Taxonomy and pH associations of β-glucosidase related sequences (annotated to CAZY families GH1, GH2, GH3, GH5, GH9, GH30, GH39 and GH116) assembled from metagenomes. Inner tree and labels depict the taxonomy of β-glucosidase associated gene assemblies constructed from pooled metagenomes from the pH 5 and pH 7 soils (n=4). Outer ring shows putative pH associations of each assembled gene, following tabulation of reads mapped to the contigs from each of the eight soil metagenomes, and statistical classification using a multinomial model based on relative abundance across the two soils (CLAM).

4.3.2 Domain classification of β-glucosidases

As β-glucosidases are associated with multiple CAZY families, and each family varies in terms of the other enzymatic activities/ substrate specificities they are associated with, I next sought to examine the relative abundance of specific CAZY families at each pH treatment.

As seen in **Fig.4.4a** soils at pH7 demonstrated a higher relative abundance of sequences annotated as GH1 in comparison to pH5, whilst pH5 soils showed a larger proportion of GH2. Given GH1 is primarily involved in β-glucosidase activity and GH2 is more commonly involved in other activities such as β -galactosidase, β -mannosidase and β –glucuronidase, this could imply greater β-glucosidase activity at pH7. In contrast to GH1, GH3, another family with common β-glucosidase activity, showed little variation between the two soil pHs.

I next looked at the taxonomy of sequences within dominant families (GH1, GH2 and GH3) and their variation in relative abundance between the two pHs. As **Fig.4.4b** demonstates, there is a drastic shift in the relative abundance of *Acidobacteria* GH1 sequences between the two soils, with a substantially larger amount of *Acidobacteria* GH1 sequences at pH5 compared to pH7, and much of the pH7 GH1 pool arising from *Veruccomicrobia*, *Actinobacteria* and *Proteobacteria*. A higher proportion of *Acidobacterial* GH2 and GH3 sequences were also observed at pH5 in comparison to pH7 (**Fig.4.4c, Fig.4.4d**).
In summary the broad differences in β-glucosidase taxonomy observed in MG-RAST annotations, were also consistent within individual β-glucosidase associated GH families, with more *Acidobacterial* sequences observed within pH5, in addition these results also revealed an elevated GH2/GH1 ratio in pH5 soils.

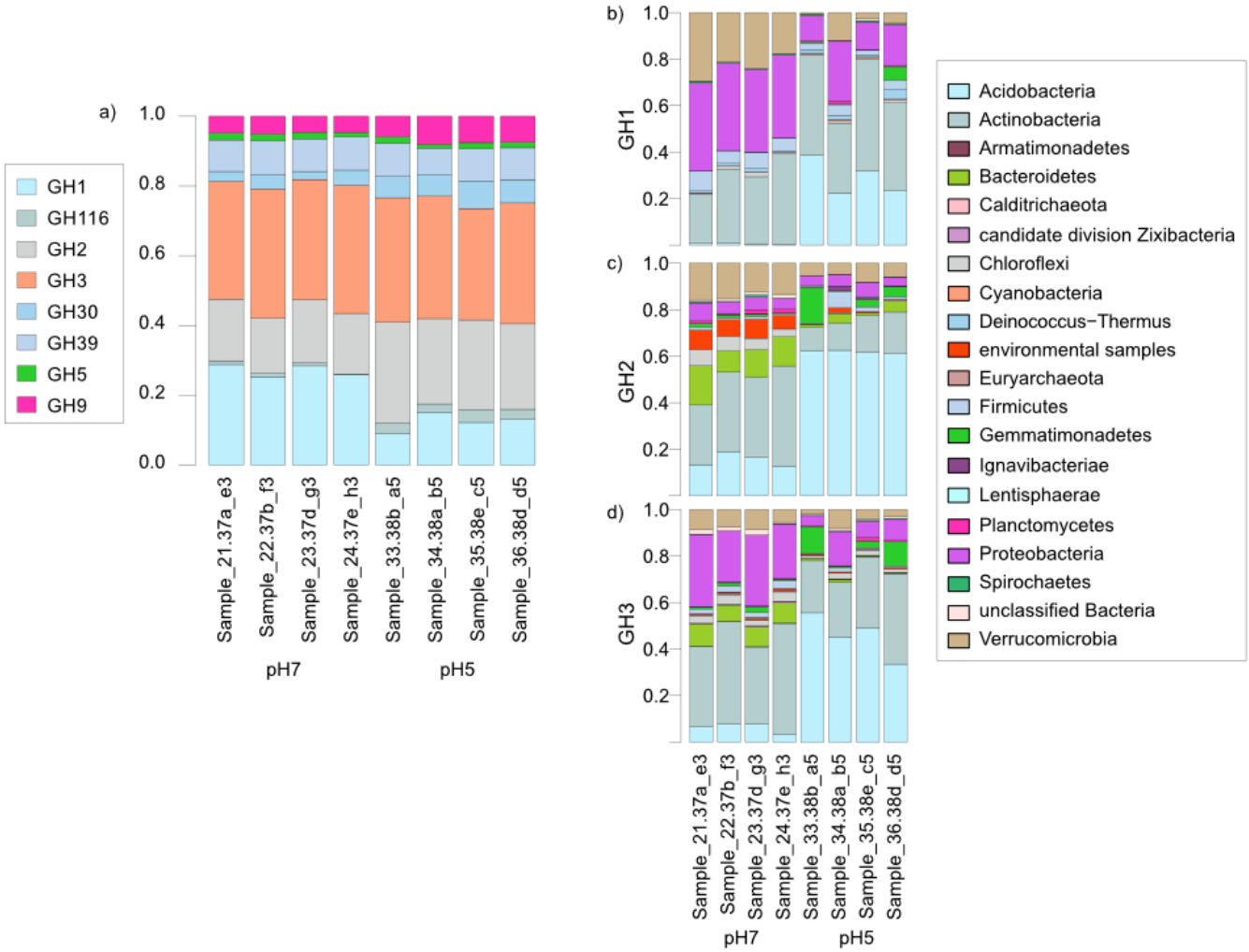


Fig.4.4. a) Stacked bar plot showing the total proportion of β-glucosidase related genes associated with differing CAZY Glycoside hydrolase (GH) families (annotated using dbCAN2) in pH7 and pH5 soils. GH family specific plots show the proportion of different phyla for all **b)** GH1, **c)** GH2 and **d)** GH3 annotated sequences within pH7 and pH5 soils.

4.3.3 Secretory motif annotations

As β-glucosidases are active both intracellularly and extracellularly I next sought to ascertain the likely cellular location of the enzyme sequences. Here, I took contigs containing β-glucosidase related sequences (identified through using dbCAN2 to annotate relevant CAZY families) and determined the presence or absence of secretory motifs using SignalP. Of the 1313 contigs containing β-glucosidase associated families, 469 (35.72%) were found to contain a secretory motif, and were classified as “extracellular”, 801 (61%) did not contain secretory motifs and were classed as “intracellular”, while 43 (3.23%) had conflicting

annotations depending on whether Gram negative or positive databases were used, and therefore were not classified as either category.Overall, a larger proportion of "extracellular" β-glucosidases demonstrated pH5 specialism (36.36%) opposed to pH7 specialism (14.94%) (**Fig.4.5a**), whilst a larger proportion of "intracellular" β-glucosidases were pH7 specialists (26.93%) opposed to pH5 (17.83%) (**Fig.4.5b**).

Across both intra and extracellular classifications *Acidobacteria* had clear groupings of pH5 specialists (**Fig.4.5**), however extracellular enzymes were represented by a greater proportion of pH5 specialists within *Acidobacteria* (58.54%) compared to "intracellular" (37.23%). "Intracellular" sequences showed a much higher proportion of pH7 specialism within *Verrucomicrobia* (43.18 %) compared to "extracellular" sequences (18.18%) (**Table.4.1**). A substantial subset of *Actinobacterial* β-glucosidases were pH7 specialists, although these were represented reasonably equally within extracellular (26.15%) and intracellular (29.21%) sequences (**Fig.4.5**, **Table.4.1**).

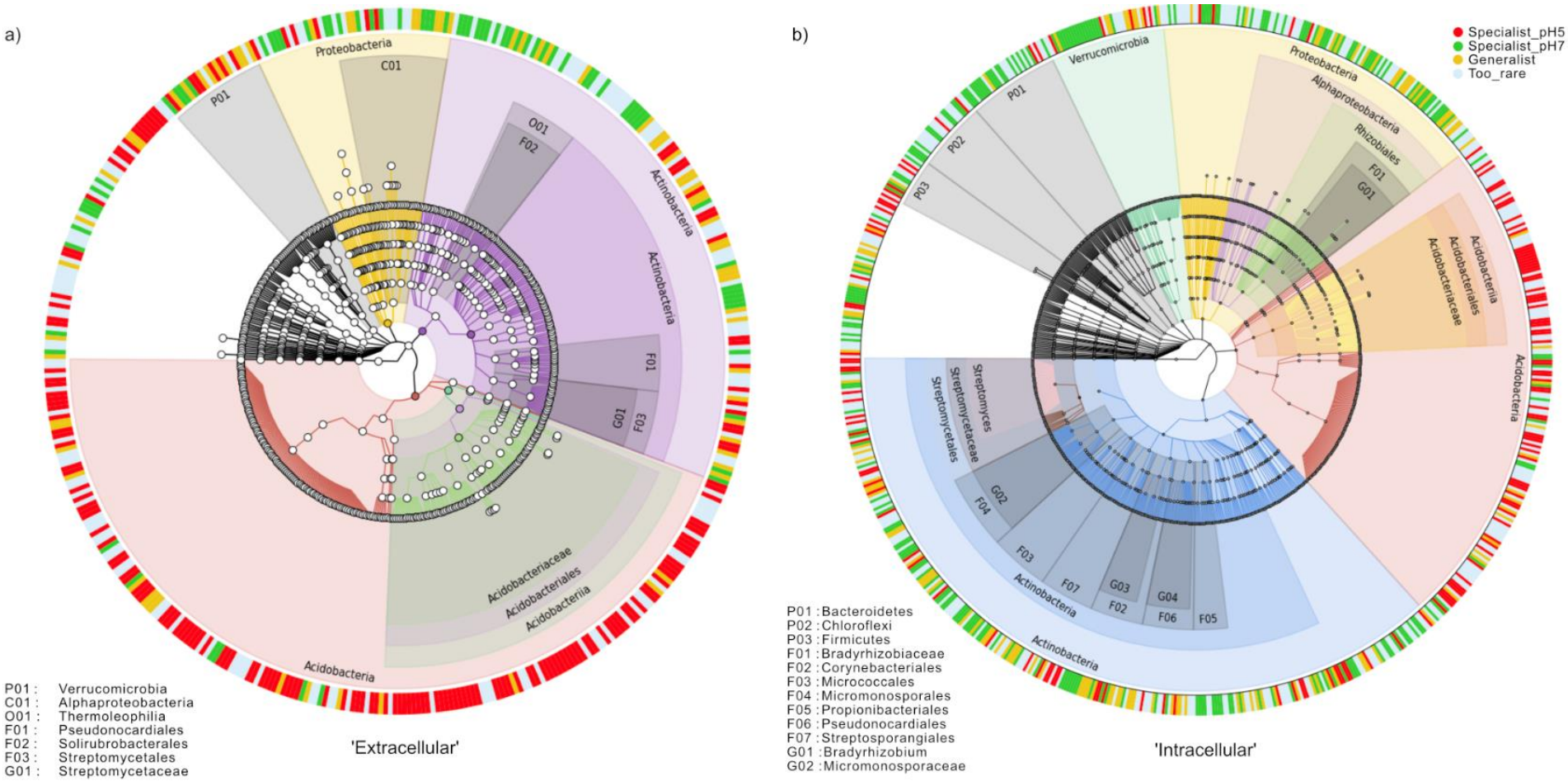


Fig.4.5. Taxonomy and pH associations of β-glucosidase related sequences (annotated to CAZY families GH1, GH2, GH3, GH5, GH9, GH30, GH39 and GH116) assembled from metagenomes for contigs with **a)** secretory motifs present ("Extracellular") and **b)** without secretory motifs present ("intracellular"). Inner tree and labels depict the taxonomy of β-glucosidase related gene assemblies constructed from pooled metagenomes from the pH5 and pH7 soils (n=4). Outer ring shows putative pH associations of each assembled gene, following tabulation of reads mapped to the contigs from each of the 8 soil metagenomes, and statistical classification using a multinomial model based on relative abundance across the two soils (CLAM).

Phylum	No of reads	Subset	Generalist		Specialist pH5		Specialist pH7		Too rare	
Acidobacteria	393	All	6.9		48.28		4.68		40.15	
		'Extracellular'	7.32	p 0.88	58.54	p 1e-04 ***	2.93	p 0.073	31.22	p 3e-04 ***
		'Intracellular'	6.91		37.23		6.91		48.94	
Actinobacteria	421	All	20.27		14.58		28.47		36.67	
		'Extracellular'	22.31	p 0.56	13.85	p 0.79	26.15	p 0.5	37.69	p 0.74
		'Intracellular'	19.93		14.78		29.21		36.08	
Bacteroidetes	50	All	5.88		7.84		47.06		39.22	
		'Extracellular'	6.67	p 0.61	13.33	p 0.35	33.33	p 0.17	46.67	p 0.54
		'Intracellular'	2.86		5.71		54.29		37.14	
Chloroflexi	36	All	2.78		11.11		30.56		55.56	
		'Extracellular'	11.11	p 0.022 *	22.22	p 0.24	22.22	p 0.53	44.44	p 0.4
		'Intracellular'	0		7.41		33.33		59.26	
Gemmatimonadetes	30	All	9.68		58.06		3.23		29.03	
		'Extracellular'	13.33	p 0.59	66.67	p 0.31	0	p 0.38	20	p 0.28
		'Intracellular'	6.67		46.67		6.67		40	
Proteobacteria	171	All	19.21		5.65		36.16		38.98	
		'Extracellular'	18.18	p 1	13.64	p 0.0027 **	36.36	p 0.93	31.82	p 0.2
		'Intracellular'	18.11		3.15		37.01		41.73	
Verrucomicrobia	66	All	27.27		16.67		34.85		21.21	
		'Extracellular'	40.91	p 0.084	27.27	p 0.097	18.18	p 0.043*	13.64	p 0.29
		'Intracellular'	20.45		11.36		43.18		25	

Table.4.1. Percentage of β-glucosidase related gene sequences (annotated to CAZY families GH1, GH2, GH3, GH5, GH9, GH30, GH39 and GH116) per bacterial phylum (with no of contigs >=30) of each pH class for all sequences, 'extracellular' (with secretory motif annotation) and 'intracellular' (without secretory motif annotation). Permutational p values (10,000 perm) test significance of difference between 'intracellular' and 'extracellular' sequences per phyla and pH class. * denotes pval < 0.05, ** pval < 0.01, *** pval < 0.001, blank denotes pval > 0.05.

4.3.4 Sequence variation and contributing factors

After looking at variation in the β-glucosidases in terms of taxonomic and CAZY family, I next sought to look at the variation at the sequence level (amino acids). Essentially, I wished to evaluate whether there are distinct amino acid signatures in the enzymes which correspond with either taxonomic origin, soil pH classification or whether the enzyme is likely to be secreted or not. These analyses were focused on sequences annotated to GH3 as it was the most dominant family within the data studied, and it is known to commonly be associated with β-glucosidase activity (Nijikken et al., 2007). I aligned all GH3 regions of contigs, where the GH3 annotation spanned more than 200 aa (annotated using 216 aa GH3 HMMER profile). I initially used this alignment to create a distance matrix and used K-means clustering to examine natural groupings within the data, which revealed two clear clusters of sequences, as visible in **Fig.4.6**.

However neither taxonomic annotation (at varying levels) nor soil pH appeared to be related to these clusters upon visualisation (not shown). There did however appear to be a relationship between clustering and secretory motif annotation (**Fig.4.6**) with cluster 1 having a much larger proportion of Sec/SP1 secretory motif annotation (39.84%) than cluster 2 (9.21%), and a larger proportion of sequences within cluster 2 (73.68%) having no secretory motif present compared to cluster 1 (49%).Though, through further statistical analyses using Adonis (permutational multivariate analysis test) I found that phyla was significantly related to matrix variance (R^2 0.11467, p value 0.001). Secretory motif annotation was also significantly related to matrix variance, although more weakly than phyla (R2 0.05291, p value 0.001). No significant relationship was seen between matrix variance and pH (R^2 of 0.00830 and a p value of 0.112).

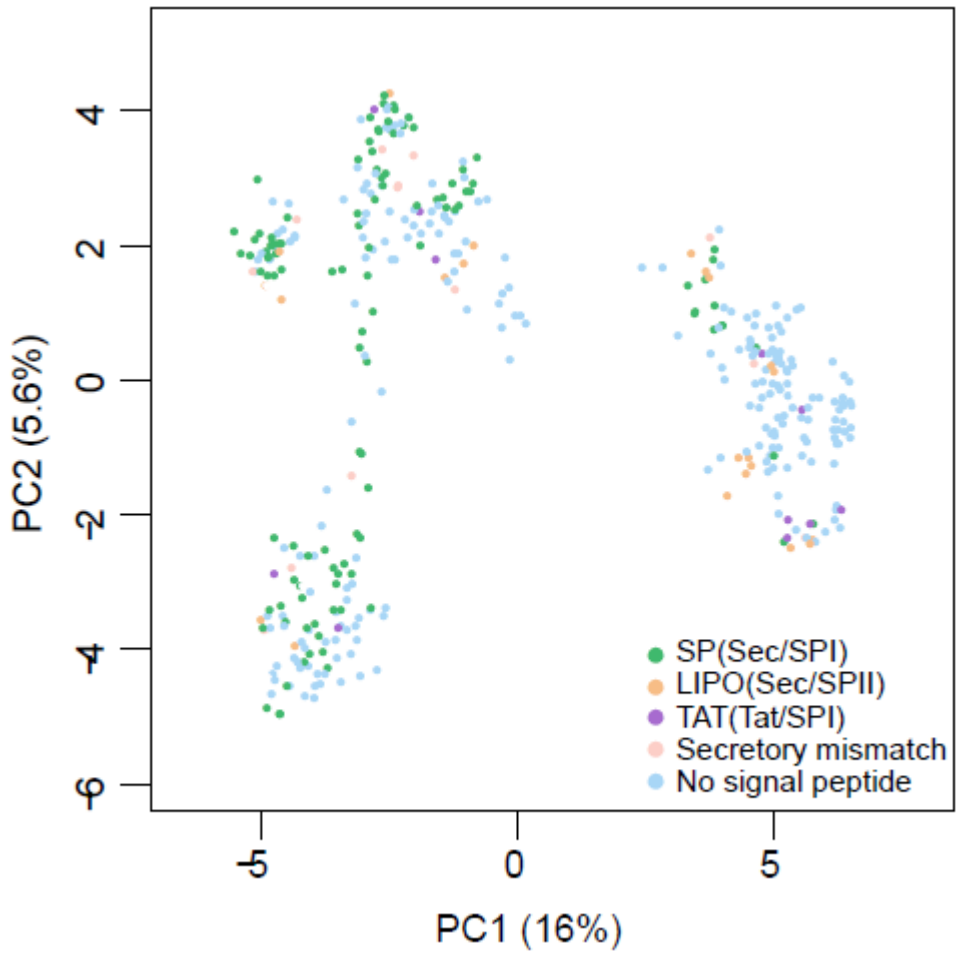


Fig.4.6. Ordination of GH3 sequences (annotated using dbCAN2) with a length >200 aa based upon a distance matrix generated on the sequence level (protein), colour depicts secretory motif annotation (annotated using SignalP).

To further examine differences in GH3 sequence composition I constructed a phylogenetic tree and examined the distribution of pH specialism, taxonomic annotation and secretory traits across the tree. From the circle plot (**Fig.4.7**) it is evident there are clear clusters of taxonomic phylogeny (shown as node colour), although these are fragmented.

Annotation depicting secretory motifs present (outer ring) shows relatively large clusters of 'intracellular' sequences, with smaller clusters of 'extracellular' β-glucosidases with Sec/SP1 secretory motifs present. There are also several acidic and neutral clusters throughout the tree (inner ring), although in general these appear smaller in comparison to clusters related to secretory motif annotation and phyla.

As both phyla and secretory motif annotation appear to be determinants of sequence similarity (from Adonis results) one might expect to see two clusters of phylogenetic

annotation per phyla (one intracellular and one extracellular) but instead there are numerous groups of the same phyla spread throughout the phylogenetic tree, for example 5 or 6 clusters of nodes of *Actinobacteria*, which could suggest lateral gene transfer of β-glucosidases between phylogenetic groups.

When comparing the relative abundances of GH3 sequences subset by pH specialism, phyla and cellular location (**Fig 4.8**), it is apparent there are significantly more "Intracellular" pH7 specialists in comparison to "extracellular" pH7 specialists (pval 0.0016). This directly contrasts with what is seen within pH5 specialists where there are more "extracellular" sequences compared to "intracellular". Although the difference between "extracellular" and "intracellular" pH5 specialist sequences was not statistically significant, pH5 specialists made up a significantly larger proportion of the total pool of "extracellular" sequences in comparison to "intracellular" (pval 1e-05). Importantly both 'Intracellular' and 'Extracellular" pH5 specialists exhibited a larger amount of *Acidobacteria* sequences than was observed pH7 specialists or generalists. These results are similar to what was seem in all β-glucosidase related sequences (annotated to GH1, GH2, GH3, GH5, GH9, GH30, GH39, GH116), where there were also a larger amount of *Acidobacteria* sequences within pH5 specialists, compared to pH7 specialists or generalists (**Fig.A4.1, Tabel.4.1**).

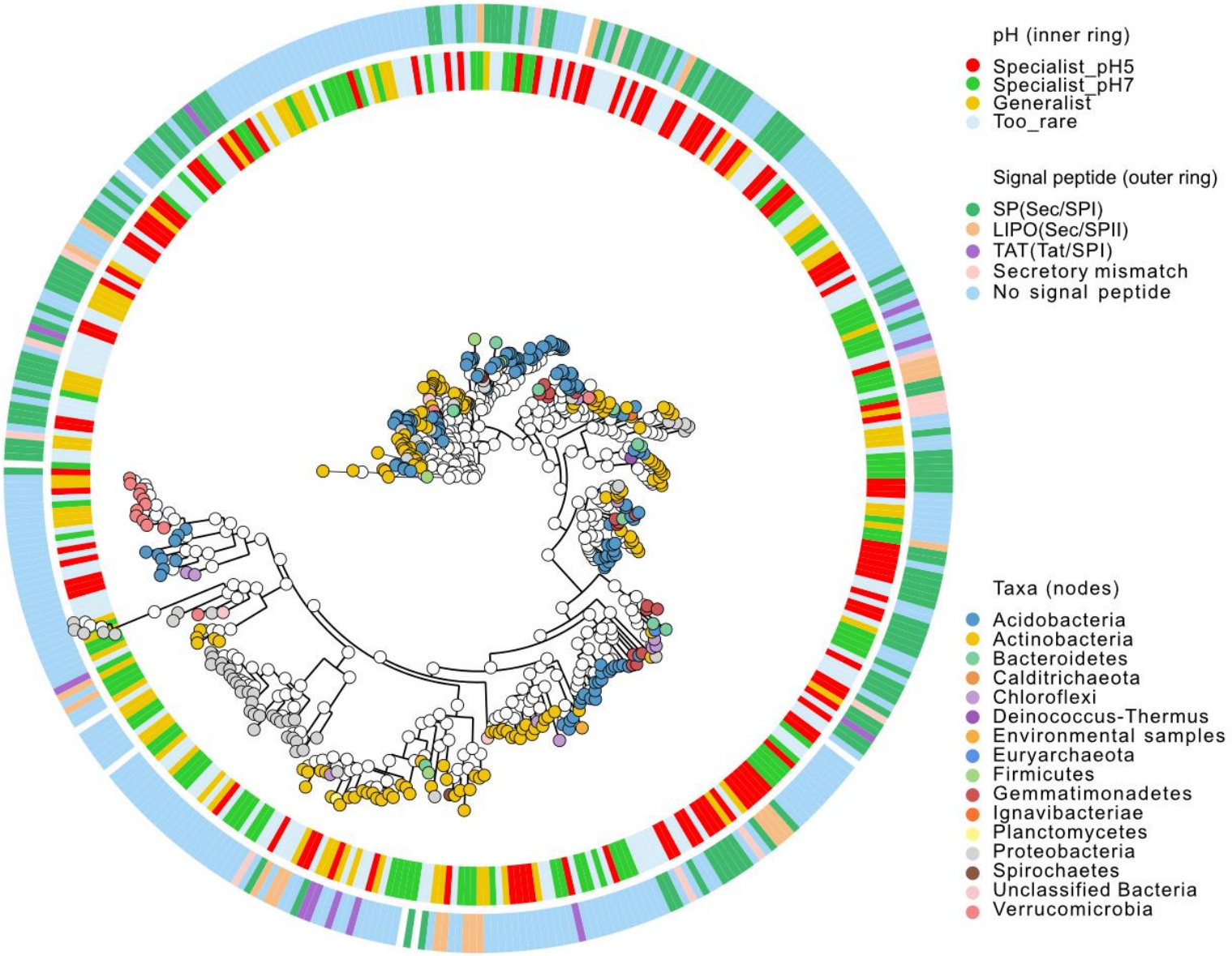


Fig.4.7. Phylogenetic tree of GH3 sequences (annotated with dbCAN2) with a length >200 aa. Inner ring shows pH preference, outer ring describes secretory motif annotation (annotated using SignalP), whilst node colour depicts phyla (annotated with Kaiju).

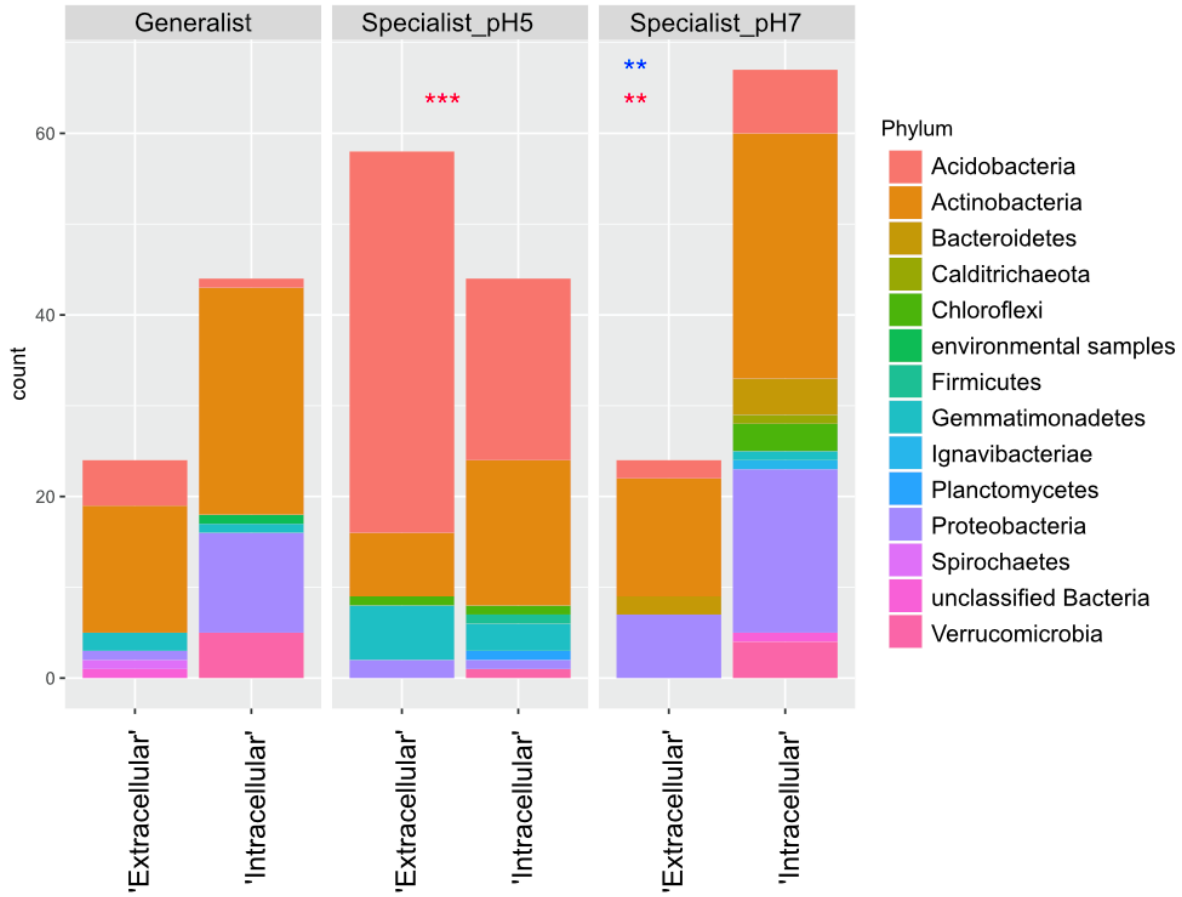


Fig.4.8. Count of GH3 annotated sequences (annotated using dbCAN2), subset by pH specialism, cellular location (inferred from secretory motif annotations conducted using SignalP) and phyla (annotated with Kaiju). Permutational p values (10,000 perm) test significance of difference between 'intracellular' and 'extracellular' within each pH class (blue *), and significance of difference in proportions of pH classes within total pool of 'intracellular' and 'extracellular' sequences (red *). * denotes pval < 0.05, ** pval < 0.01, *** pval < 0.001, blank denotes pval > 0.05.

4.4 Discussion

4.4.1 Taxonomic shifts and soil pH

This work applied metagenomics to the long term Park Grass experiment and found that *Acidobacteria* are a larger contributor to the β-glucosidase gene pool in pH5 than pH7, and that a large proportion of *Acidobacteria* β-glucosidase associated sequences are exclusive to pH5. Previous work looking at the microbial communities within the Park Grass experiment demonstrated shifts in microbial composition related to pH (Zhalnina et al., 2014), with higher relative abundances of *Acidobacteria* found at pH5 compared to pH7 (Puissant et al., 2019). Whilst it's not possible to definitively say whether the finding of more *Acidobacteria*

β-glucosidases at pH5 is due to there being more *Acidobacteria* within pH5 soils or whether this is because there are specifically more *Acidobacteria* β-glucosidases present, the marked difference in β-glucosidase abundance between the two sites even when normalised by housekeeping genes (gyrB) suggests the latter. Previously work which assembled genomes from soil bacteria isolates also emphasised the importance of *Acidobacteria* to the carbohydrate degradation processes, with *Acidobacteria* assembled genomes containing a significantly higher number of genes encoding carbohydrate degrading enzymes in comparison to *Proteobacteria* genomes (Lladó et al., 2019). Supplementary to this work , extracellular β-glucosidase enzyme assays were conducted on Park Grass pH5 and pH7 soils and revealed clear shifts in β-glucosidase pH optima between pH5 and pH7 (**appendix 1**) (Puissant et al., 2019). The differences in relative abundance of *Acidobacteria* β-glucosidase genes at pH5 and pH7 found in the present study provides a potential genomic mechanism behind the variation in pH optima seen in β-glucosidase enzyme assays. However, it is worth noting that as the genomic information used within the present study was extracted from metagenomes it's impossible to definitively say these sequences encoded the specific enzymes active within the assays. I did however try to increase the relevance of the sequences extracted by also mining contigs for secretory motifs. An alternate approach to obtain a more causal link between genomic data and enzyme activity would be to directly sequence and assay enzymes from isolates (Lladó et al., 2019, 2016) .

Indeed future work could also delve deeper into the sequence signatures and characteristics of pH5 and pH7 specialist β-glucosidase sequences. In this current study I did additionally examine proportions of amino acid subtypes (tiny, small, aliphatic, aromatic, non-polar, polar, charged, basic and acidic) at each pH, but found very little differences between pH 5 and pH7 specialists (results not shown). Work relating enzyme sequences to pH preference is a growing area with methods being developed to predict pH optima from cellulase sequences using amino acid position and distribution probabilities in neural networks (Yan and Wu, 2012) and alternate machine learning methods being used to discriminate between "acidic" and "alkaline" enzymes using g-gap dipeptide compositions (correlation of amino acids separated by g number of residues) (Lin et al., 2013). Such tools could be

employed for further understanding of pH related differences in sequence composition within β-glucosidases.

A better grasp of the sequence and structure of the active site region may also aid understanding of pH5 and pH7 specialists, with other work finding that mutations within the active site of carboxylesterases are influential over pH preference, whereby acidophilic carboxylesterases demonstrate extended hydrogen bond networks within the active site which are not present in their alkaliphilic equivalents (Ohara et al., 2014). In the current study, I conducted some preliminary analyses in this area whereby I extracted active site regions from the metagenomic derived β-glucosidase sequences, but I found little evidence of active site sequence variation being related to pH in ordinations (results not shown).

4.4.2 CAZY subfamilies

I also identified shifts in β-glucosidase associated CAZY families in pH5 and pH7 soils. Annotating β-glucosidases using CAZY families is not without its challenges as most Glycoside Hydrolase families are 'polyspecific' meaning they possess multiple enzymatic activities opposed to 'monospecific' where one family maps perfectly to a single enzyme activity (Aspeborg et al., 2012; Nguyen et al., 2018). When comparing differences in the relative abundance of β-glucosidase associated families at each site, I found a larger proportion of GH1 at pH7 compared to pH5 and a larger proportion of GH2 sequences at pH5 compared to pH7. As GH1 has common β-glucosidase activity, whilst GH2 has rarer β-glucosidase activity and more common β-galactosidase and β-mannosidase activities this suggests a larger proportion of β-glucosidase encoding genes at pH7. GH3 another family with common β-glucosidase activity showed greater consistency across pH5 and pH7 sites. Understanding the likely locations of these families provides further context to interpret these findings, with β-glucosidases in GH3 being thought to be more likely to be extracellular or cell bound, whilst β-glucosidases in GH1 are thought to be predominantly intracellular (Nijikken et al., 2007; Zhou et al., 2012). This suggests the shifts in GH1 maybe not necessarily be directly relevant to the extracellular enzymes assayed in previous work (Puissant et al., 2019).GH3 did however show a large shift in *Acidobacteria* sequences in pH5 compared to pH7, although similar results were also observed within GH1 and GH2.

4.4.3 Sequence phylogeny and secretory motif annotation

Through drilling down to just GH3 sequences and conducting sequence alignments and subsequently NMDS analyses (on an alignment derived sequence distance matrix), I identified two clear clusters of sequences, which loosely demonstrated relation to annotation of secretory motifs. Accurately distinguishing between intracellular and extracellular enzyme sequences has been a longstanding challenge in molecular biology (Duly and Nannipieri, 1998), however the advent of high-throughput sequencing has led to a number of tools being developed to detect secretory motifs (Almagro Armenteros et al., 2019; Yu et al., 2010). Work employing such tools have provided us with previously unknown insights into the relationship between microbial community and function, with a recent study integrating 16S data with assembled genomes showing that microbes living in more structured habitats had a greater amount of genes encoding extracellular proteins (Barberán et al., 2012).

The differences seen in the present study in proportions of intracellular and extracellular sequences, particularly in the context of *Acidobacteria* pH5 specialists, suggests further relevance of the *Acidobacteria* pH5 specialist sequences to the differences seen previously in enzyme assays (Puissant et al., 2019), as it is apparent that a significantly larger proportion of these sequences have secretory motifs and thus are likely to be extracellular opposed to intracellular. Interestingly as mentioned most β-glucosidases in GH3 are thought to be extracellular (Ahmed et al., 2017; Nijikken et al., 2007; Zhou et al., 2012) however in the current study I found a large amount of GH3 sequences did not appear to have secretory motifs. This could suggest these GH3 sequences are not secreted, or that these GH3 sequences do not possess β-glucosidase activity (and instead perhaps are involved in alternate GH3 associated enzymatic activity) it may also be due to limitations in detecting certain taxon's secretory motifs within the SignalP database.

Nevertheless, both in the case of GH3 and when looking at combined counts of all β-glucosidase associated families (annotated to GH1, GH2, GH3, GH5, GH9, GH30, GH39 and GH116) I found a larger proportion of pH5 specialists within 'extracellular' sequences

(containing secretory motifs) sequences compared to 'intracellular' sequences (no secretory motif). This may be related to the fact that acidic soils are typically nutrient poor; indeed economic theories in microbiology hypothesise that enzyme production will increase when the environment lacks simple nutrients but possesses a wealth of complex nutrients. Although equally resource constraints could also reduce enzyme production (Allison and Vitousek, 2005). Studies on tundra soils (acidic soils with low nutrient concentrations) , have indicated the later may be true, as while both pH and nutrient availability appear to exert control on cellulose decomposition (Koyama et al., 2013; Stark et al., 2014), they demonstrated opposing affects. With nutrient limitation being linked to lower enzyme activity (Koyama et al., 2013; Stark et al., 2014) and low pH being related to increased enzyme activity (Stark et al., 2014).

4.4.4 Sequence phylogeny and taxonomic assignment

Whilst I found that GH3 sequence variance was significantly related to phyla (using Adonis)**,** the clusters of GH3 sequences sharing the same phyla were fragmented within the phylogenetic tree. This likely suggests there is some degree of exchange of β-glucosidases between taxa for example by horizontal gene transfer. This is consistent with what has previously been seen in GH6 (another glycoside hydrolase family, involved in earlier stages of cellulose degradation) where qPCR has demonstrated ambiguity between taxonomic and GH6 phylogeny with sequences from different taxonomic domains and phyla appearing to be closely related within the phylogenetic tree (Merlin et al., 2014). Other work has found that while GH gene content is extremely variable at the phyla level, GH1 and GH3 shows greater conservation at the genus level (Berlemont and Martiny, 2015; 2013), this could provide a possible explanation for the smaller clusters of GH3 phylogenetic similarity found here.

4.5 Conclusions

These results show that soil pH demonstrates influence over the taxonomic annotation of β-glucosidases with pH5 showing a much larger proportion of *Acidobacteria* sequences and these sequences appearing to be more unique to pH5 soils. Further to this, pH5 specialist β-

glucosidase associated sequences had a significantly larger amount of *Acidobacteria* 'extracellular' sequences opposed to 'intracellular' based upon annotations of secretory motifs. Shifts in *Acidobacteria* sequences were seen in GH1, GH2 and GH3 CAZY families which vary in terms of how commonly they are involved β-glucosidase activity as well as their associated location (secreted or not). Phylogeny of GH3 sequences (common β-glucosidase activity and reported extracellular association) appears to be largely dependent on secretory motifs opposed to pH preference. Further, the clusters of phylogenetically related GH3 sequences sharing taxonomic annotations were reasonably small, suggesting an exchange of GH3 sequences potentially by horizontal gene transfer. This work highlights the use of assembling and annotating enzyme genes from metagenomes by contributing new knowledge in terms of the microbial contributors of β-glucosidase sequences across soils of different pH, and new insights into contributors to GH3 phylogeny, which could be potentially useful to β-glucosidase applications in biotechnology.

4.6 Bibliography

Ahmed, A., Nasim, F., Batool, K., Bibi, A., 2017. Microbial β -Glucosidase : Sources , Production and Applications. doi:10.12691/jaem-5-1-4

Allison, S.D., 2014. Modeling adaptation of carbon use efficiency in microbial communities. Frontiers in Microbiology 5, 1–9. doi:10.3389/fmicb.2014.00571

Allison, S.D., 2012. A trait-based approach for modelling microbial litter decomposition. Ecology Letters. doi:10.1111/j.1461-0248.2012.01807.x

Allison, S.D., 2005. Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. Ecology Letters 8, 626–635. doi:10.1111/j.1461-0248.2005.00756.x

Allison, S.D., Lu, L., Kent, A.G., Martiny, A.C., 2014. Extracellular enzyme production and cheating in Pseudomonas fluorescens depend on diffusion rates. Frontiers in Microbiology 5, 1–8. doi:10.3389/fmicb.2014.00169

Allison, S.D., Lu, Y., Weihe, C., Goulden, M.L., Martiny, A.C., Treseder, K.K., Martiny, J.B.H., 2013. Microbial abundance and composition influence litter decomposition response to environmental change. Ecology 94, 714–725. doi:10.1890/12-1243.1

Allison, S.D., Vitousek, P.M., 2005. Responses of extracellular enzymes to simple and complex nutrient inputs. Soil Biology and Biochemistry 37, 937–944. doi:10.1016/j.soilbio.2004.09.014

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology 37, 420–423. doi:10.1038/s41587-019-0036-z

Aspeborg, H., Coutinho, P.M., Wang, Y., Brumer, H., Henrissat, B., 2012. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). BMC Evolutionary Biology 12, 1. doi:10.1186/1471-2148-12-186

Bandick, A.K., Dick, R.P., 1999. Field management effects on soil enzyme activities. Soil Biology and Biochemistry 31, 1471–1479. doi:10.1016/S0038-0717(99)00051-6

Banerjee, S., Kirkby, C.A., Schmutter, D., Bissett, A., Kirkegaard, J.A., Richardson, A.E., 2016. Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. Soil Biology and Biochemistry 97, 188–198. doi:10.1016/j.soilbio.2016.03.017

Barberán, A., Bates, S.T., Casamayor, E.O., Fierer, N., 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME Journal 6, 343–351. doi:10.1038/ismej.2011.119

Berlemont, R., Allison, S.D., Weihe, C., Lu, Y., Brodie, E.L., Martiny, J.B.H., Martiny, A.C., 2014. Cellulolytic potential under environmental changes in microbial communities from grassland litter. Frontiers in Microbiology 5, 1–10. doi:10.3389/fmicb.2014.00639

Berlemont, R., Martiny, A.C., 2016. Glycoside Hydrolases across Environmental Microbial Communities. PLoS Computational Biology 12, 1–16. doi:10.1371/journal.pcbi.1005300

Berlemont, R., Martiny, A.C., 2015. Genomic potential for polysaccharide deconstruction in bacteria. Applied and Environmental Microbiology 81, 1513–1519. doi:10.1128/AEM.03718-14

Berlemont, R., Martiny, A.C., 2013. Phylogenetic distribution of potential cellulases in bacteria. Applied and Environmental Microbiology 79, 1545–1554. doi:10.1128/AEM.03305-12

Bradford, M.A., 2013. Thermal adaptation of decomposer communities in warming soils. Frontiers in Microbiology 4, 1–16. doi:10.3389/fmicb.2013.00333

Burns, R.G., DeForest, J.L., Marxsen, J., Sinsabaugh, R.L., Stromberger, M.E., Wallenstein, M.D., Weintraub, M.N., Zoppini, A., 2013. Soil enzymes in a changing environment: Current knowledge and future directions. Soil Biology and Biochemistry. doi:10.1016/j.soilbio.2012.11.009

Cantarel, B.I., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B., 2009. The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. Nucleic Acids Research 37, 233–238. doi:10.1093/nar/gkn663

Chazdon, R.L., Chao, A., Colwell, R.K., Lin, S.Y., Norden, N., Letcher, S.G., Clark, D.B., Finegan, B., Arroyo, J.P., 2011. A novel statistical method for classifying habitat generalists and specialists. Ecology 92, 1332–1343. doi:10.1890/10-1345.1

Cleveland, C.C., Reed, S.C., Keller, A.B., Nemergut, D.R., O'Neill, S.P., Ostertag, R., Vitousek, P.M., 2014. Litter quality versus soil microbial community controls over decomposition: A quantitative analysis. Oecologia 174, 283–294. doi:10.1007/s00442-013-2758-9

Dick, R.P., Kandeler, E., 2005. ENZYMES IN SOILS, in: Hillel, D.B.T.-E. of S. in the E. (Ed.), . Elsevier, Oxford, pp. 448–456. doi:https://doi.org/10.1016/B0-12-348530-4/00146-6

Duly, O., Nannipieri, P., 1998. Intracellular and extracellular enzyme activity in soil with reference to elemental cycling. Zeitschrift Für Pflanzenernährung Und Bodenkunde 161, 243–248. doi:10.1002/jpln.1998.3581610310

Fierer, N., 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, 626–631. doi:10.1073/pnas.0507535103

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Research 39, 29–37. doi:10.1093/nar/gkr367

Gangoiti, J., Pijning, T., Dijkhuizen, L., 2018. Biotechnological potential of novel glycoside hydrolase family 70 enzymes synthesizing α-glucans from starch and sucrose. Biotechnology Advances 36, 196–207. doi:10.1016/j.biotechadv.2017.11.001

German, D.P., Weintraub, M.N., Grandy, A.S., Lauber, C.L., Rinkes, Z.L., Allison, S.D., 2011. Optimization of hydrolytic and oxidative enzyme methods for ecosystem studies. Soil Biology and Biochemistry. doi:10.1016/j.soilbio.2011.03.017

Goulding, K.W.T., 2016. Soil acidification and the importance of liming agricultural soils with particular reference to the United Kingdom. Soil Use and Management 32, 390–399. doi:10.1111/sum.12270

Griffiths, R.I., Thomson, B.C., James, P., Bell, T., Bailey, M., Whiteley, A.S., 2011. The bacterial biogeography of British soils. Environmental Microbiology 13, 1642–1654. doi:10.1111/j.1462-2920.2011.02480.x

Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.P., Davies, G., 1995. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proceedings of the National Academy of Sciences of the United States of America 92, 7090–7094. doi:10.1073/pnas.92.15.7090

Jansson, J.K., Hofmockel, K.S., 2018. The soil microbiome — from metagenomics to metaphenomics. Current Opinion in Microbiology 43, 162–168. doi:10.1016/j.mib.2018.01.013

Jones, B.A., Goodall, T., George, P., Gweon, H.S., Puissant, J., Read, D., Emmett, B.A., Robinson, D.A., Jones, D.L., Griffiths, R.I., 2019. Beyond taxonomic identification: integration of ecological responses to a soil bacterial 16S rRNA gene database. BioRxiv 843847. doi:10.1101/843847

Jones, D.L., Cooledge, E.C., Hoyle, F.C., Griffiths, R.I., Murphy, D. V., 2019. pH and exchangeable aluminum are major regulators of microbial energy flow and carbon use efficiency in soil microbial communities. Soil Biology and Biochemistry 138, 0–4. doi:10.1016/j.soilbio.2019.107584

Joshi, N.A., Fass, J.N., others, 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software].

Koyama, A., Wallenstein, M.D., Simpson, R.T., Moore, J.C., 2013. Carbon-Degrading Enzyme Activities Stimulated by Increased Nutrient Availability in Arctic Tundra Soils. PLoS ONE 8. doi:10.1371/journal.pone.0077212

Li, D., Liu, C.M., Luo, R., Sadakane, K., Lam, T.W., 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. doi:10.1093/bioinformatics/btv033

Lin, H., Chen, W., Ding, H., 2013. AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. PLoS ONE 8. doi:10.1371/journal.pone.0075726

Lladó, S., Větrovský, T., Baldrian, P., 2019. Tracking of the activity of individual bacteria in temperate forest soils shows guild-specific responses to seasonality. Soil Biology and Biochemistry 135, 275–282. doi:10.1016/j.soilbio.2019.05.010

Lladó, S., Žifčáková, L., Větrovský, T., Eichlerová, I., Baldrian, P., 2016. Functional screening of abundant bacteria from acidic forest soil indicates the metabolic potential of Acidobacteria subdivision 1 for polysaccharide decomposition. Biology and Fertility of Soils 52, 251–260. doi:10.1007/s00374-015-1072-6

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 42, 490–495. doi:10.1093/nar/gkt1178

Lonhienne, T., Gerday, C., Feller, G., 2000. Psychrophilic enzymes: Revisiting the thermodynamic parameters of activation may explain local flexibility. Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology 1543, 1–10. doi:10.1016/S0167-4838(00)00210-7

Malik, A.A., Puissant, J., Buckeridge, K.M., Goodall, T., Jehmlich, N., Chowdhury, S., Gweon, H.S., Peyton, J.M., Mason, K.E., van Agtmaal, M., Blaud, A., Clark, I.M., Whitaker, J., Pywell, R.F., Ostle, N., Gleixner, G., Griffiths, R.I., 2018. Land use driven change in soil pH affects microbial carbon cycling processes. Nature Communications 9, 1–10. doi:10.1038/s41467-018-05980-1

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17, 10–12. doi:10.14806/ej.17.1.200

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications 7, 1–9. doi:10.1038/ncomms11257

Merlin, C., Besaury, L., Niepceron, M., Mchergui, C., Riah, W., Bureau, F., Gattin, I., Bodilis, J., 2014. Real-time PCR for quantification in soil of glycoside hydrolase family 6 cellulase genes. Letters in Applied Microbiology 59, 284–291. doi:10.1111/lam.12273

Nguyen, S.T.C., Freund, H.L., Kasanjian, J., Berlemont, R., 2018. Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy. Applied Microbiology and Biotechnology 102, 1629–1637. doi:10.1007/s00253-018-8778-y

Niemi, R.M., Vepsäläinen, M., 2005. Stability of the fluorogenic enzyme substrates and pH optima of enzyme activities in different Finnish soils. Journal of Microbiological Methods 60, 195–205. doi:10.1016/j.mimet.2004.09.010

Nijikken, Y., Tsukada, T., Igarashi, K., Samejima, M., Wakagi, T., Shoun, H., Fushinobu, S., 2007. Crystal structure of intracellular family 1 β-glucosidase BGL1A from the basidiomycete Phanerochaete chrysosporium. FEBS Letters 581, 1514–1520. doi:10.1016/j.febslet.2007.03.009

Ohara, K., Unno, H., Oshima, Y., Hosoya, M., Fujino, N., Hirooka, K., Takahashi, S., Yamashita, S., Kusunoki, M., Nakayama, T., 2014. Structural insights into the low pH adaptation of a unique carboxylesterase from Ferroplasma: Altering the pH optima of two carboxylesterases. Journal of Biological Chemistry 289, 24499–24510. doi:10.1074/jbc.M113.521856

Parks, D. H. et al. (2018) ‘A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life’, Nature Biotechnology. Nature Publishing Group, 36(10), p. 996. doi: 10.1038/nbt.4229.

Parks, D. H. et al. (2020) ‘A complete domain-to-species taxonomy for Bacteria and Archaea’, Nature Biotechnology. Nature Research, 38(9), pp. 1079–1086. doi: 10.1038/s41587-020-0501-8.

Puissant, J., Jones, B., Goodall, T., Mang, D., Blaud, A., Gweon, H., Malik, A., Jones, D., Clark, I., Hirsch, P., Griffiths, R., 2019. The pH optimum of soil exoenzymes adapt to long term changes in soil pH. Soil Biology and Biochemistry 138, 107601. doi:10.1016/j.soilbio.2019.107601

Rho, M., Tang, H., Ye, Y., 2010. FragGeneScan: Predicting genes in short and error-prone reads. Nucleic Acids Research 38, 1–12. doi:10.1093/nar/gkq747

Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M., Biss, P.M., 2006. The Park Grass Experiment 1856-2006: Its contribution to ecology. Journal of Ecology 94, 801–814. doi:10.1111/j.1365-2745.2006.01145.x

Srivastava, N., Rathour, R., Jha, S., Pandey, K., Srivastava, M., Thakur, V.K., Sengar, R.S., Gupta, V.K., Mazumder, P.B., Khan, A.F., Mishra, P.K., 2019. Microbial beta glucosidase enzymes: Recent advances in biomass conversation for biofuels application. Biomolecules 9, 1–23. doi:10.3390/biom9060220

Stark, S., Männistö, M.K., Eskelinen, A., 2014. Nutrient availability and pH jointly constrain microbial extracellular enzyme activities in nutrient-poor tundra soils. Plant and Soil 383, 373–385. doi:10.1007/s11104-014-2181-y

Strickland, M.S., Lauber, C., Fierer, N., Bradford, M.A., 2009. Testing the functional significance of microbial community composition. Ecology 90, 441–451. doi:10.1890/08-0296.1

Tian, D., Niu, S., 2015. A global analysis of soil acidification caused by nitrogen addition. Environmental Research Letters 10. doi:10.1088/1748-9326/10/2/024019

Turner, B.L., 2010. Variation in ph optima of hydrolytic enzyme activities in tropical rain forest soils. Applied and Environmental Microbiology 76, 6485–6493. doi:10.1128/AEM.00560-10

Wang, G., Post, W.M., Mayes, M.A., 2013. Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses. Ecological Applications 23, 255–272. doi:10.1890/12-0681.1

Wei, H., Guenet, B., Vicca, S., Nunan, N., AbdElgawad, H., Pouteau, V., Shen, W., Janssens, I.A., 2014. Thermal acclimation of organic matter decomposition in an artificial forest soil is related to shifts in microbial community structure. Soil Biology and Biochemistry 71, 1–12. doi:10.1016/j.soilbio.2014.01.003

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K.P., Paczian, T., Trimble, W.L., Bagchi, S., Grama, A., Chaterji, S., Meyer, F., 2016. The MG-RAST metagenomics

database and portal in 2015. Nucleic Acids Research 44, D590–D594. doi:10.1093/nar/gkv1322

Yan, S.-M., Wu, G., 2012. Prediction of Optimal pH and Temperature of Cellulases Using Neural Network. Protein & Peptide Letters 19, 29–39. doi:10.2174/092986612798472794

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Cenk Sahinalp, S., Ester, M., Foster, L.J., Brinkman, F.S.L., 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26, 1608–1615. doi:10.1093/bioinformatics/btq249

Zanphorlin, L.M., De Giuseppe, P.O., Honorato, R.V., Tonoli, C.C.C., Fattori, J., Crespim, E., De Oliveira, P.S.L., Ruller, R., Murakami, M.T., 2016. Oligomerization as a strategy for cold adaptation: Structure and dynamics of the GH1 β-glucosidase from Exiguobacterium antarcticum B7. Scientific Reports 6, 1–14. doi:10.1038/srep23776

Zhalnina, K., Dias, R., de Quadros, P.D., Davis-Richardson, A., Camargo, F.A.O., Clark, I.M., McGrath, S.P., Hirsch, P.R., Triplett, E.W., 2014. Soil pH Determines Microbial Diversity and Composition in the Park Grass Experiment. Microbial Ecology 69, 395–406. doi:10.1007/s00248-014-0530-2

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., Yin, Y., 2018. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research 46, W95–W101. doi:10.1093/nar/gky418

Zhou, Q., Xu, J., Kou, Y., Lv, X., Zhang, X., Zhao, G., Zhang, W., Chen, G., Liu, W., 2012. Differential involvement of β-glucosidases from Hypocrea jecorina in rapid induction of cellulase genes by cellulose and cellobiose. Eukaryotic Cell 11, 1371–1381. doi:10.1128/EC.00170-12

Chapter 5

Functional and ecological properties of assembled genomes from soils under different land uses

Abstract

Advances in assembly and binning methods now enable functional characterisation of uncultivated microbes from metagenomic sequences. Such methods may also permit application of trait based approaches to soil microbial ecology, enabling prediction of soil functional change based on models of microbial taxon distributions coupled with information on specific functional genetic capabilities. Within this work I assembled 88 soil metagenomes from distributed grassland and arable soils to determine the functional attributes of specific dominant microbial taxa and examined their responses to soil pH and land use. Ordination of assembled genome "bin" abundances within sites, revealed arable soil communities to be highly distinct from grasslands, although intensive and unintensive grasslands were largely indistinguishable from one another and were strongly affected by natural gradients of soil pH. Using a random forest approach I found that the abundance of numerous *Thaurmarchaeota* bins were strong indicators of arable soils, and to a lesser extent genomes of *Actinobacteria* and *Chloroflex*i. Grassland indicators were found to be representatives of the *Alphaproteobacteria* and *Verrucomicrobia*. All genomes were then functionally annotated revealing broad phylogenetic clustering of functional gene content, which was a stronger predictor than ecological classifications based on modelled pH preference. *Thaurmarchaeota* bins in particular formed a highly distinct cluster, due largely to increases in genes relating to protein metabolism. Within short read analyses, I observed shifts in N, P, and S cycling genes in response to land use, which suggested varying nutrient acquisition strategies from organic and inorganic sources. I coupled these broad functional indicators of land use with indicators of taxonomic groupings based on bin functional gene content providing insight into the potential taxa mediating these functional changes.

5.1 Introduction

The concept of traits has been a longstanding theoretical framework, most commonly applied to plant and animal populations, placing emphasis on analyses of taxon attributes as opposed to taxonomy (Weiher and Keddy, 1995). Here a "response group" is defined as group of taxa that respond the same way to environmental stressors and an "effect group" is a group of taxa that impact one or more ecosystem functions in the same way (Suding *et al.*, 2008). Indeed applying such a framework to microbial communities could be of great value, given the significant role microbial communities play in biogeochemical cycles (Falkowski, Fenchel and Delong, 2008) and the impact that climate and land use change will likely have on them in terms of both community and function (Fichtner *et al.*, 2014; Paul *et al.*, 2019; Bardgett and Caruso, 2020; Jansson and Hofmockel, 2020). Understanding how much these response and effect groups are coupled is of particular value as it would enable insights into the likely functional consequences and the resilience of soil systems to environmental change. For example, if response and effect groups are highly related a change in an environmental factor could potentially wipe out a function entirely. Whilst if response and effect groups are not related an environmental change may have little impact on ecosystem functioning (Suding *et al.*, 2008). There is now an opportunity to apply a trait based framework to microbes given the wealth of phylogenetic and functional information we can rapidly obtain using amplicon and metagenome approaches respectively (Martiny *et al.*, 2015).

In relation to microbial responses to environmental gradients (response traits),it has traditionally been thought that microbes are globally distributed and that they are capable of proliferating anywhere with suitable conditions (Green, Bohannan and Whitaker, 2008), or as it was put by Lourens Gerhard Marinus Baas Becking "everything is everywhere but the environment selects" (Baas-Becking, 1934). Whilst this assertion has previously been found to be the case in some cultured organisms (Ramette and Tiedje, 2007), the advent of high-throughput sequencing has demonstrated little overlap between organisms that are cultured and organisms that dominate soils (Dunbar *et al.*, 1999; Amann and Ludwig, 2000), leading many to question whether this is actually the case within the uncultivated majority of soil organisms. Indeed applying molecular methods to large surveys has revealed new

insights into microbial distributions, particularly in relation to uncultivated microbial responses to environmental gradients. We now know for example that soil pH is broadly predictive of bacterial diversity as well as community composition (Fierer and Jackson, 2006; Griffiths *et al.*, 2011) and within **Chapter 2** I demonstrated how most soil bacterial taxa have predictable responses to soil pH at the OTU level. Molecular approaches have also been used to gain insights into land use effects on microbial communities, where numerous studies have reported changes in microbial diversity within intensively managed soils (Hartmann *et al.*, 2015; Mackelprang *et al.*, 2018; Sui *et al.*, 2019) as well as land use induced shifts in bacterial, archaeal and fungal community composition (Bissett *et al.*, 2011; Banerjee *et al.*, 2019). Key questions are now starting to be addressed as to how these reported changes (both in response to pH and land use) in taxa impact the functional gene pool and what consequences this may have on soil function and biogeochemical cycling.

In terms of taxon functional gene content (or microbial effect traits), it has long been postulated that microbes are largely functionally redundant, in part because bacteria can freely exchange genomic elements (Cohan and Koeppel, 2008; Martiny *et al.*, 2015). However, the amount of functional redundancy in soil microbial communities is difficult to assess, as the vast majority of microbes, including most dominant soil microbes remain uncultured. Its reported that just 1% of all microbes can be successfully cultivated using standardized procedures (Alneberg *et al.*, 2014) mostly due to a lack of understanding of their nutritional requirements and long generation times (Nannipieri *et al.*, 2020). As a result of this in 2018, half of the 60 known bacterial and archaeal phyla had solely been detected through 16S studies and had no actual cultured representatives (Marie E Kroeger *et al.*, 2018). Some bioinformatics approaches have tried to predict functional traits of microbial communities using marker gene data (Langille *et al.*, 2013; Nguyen *et al.*, 2016). For example Picrust uses 16S data to predict the functional profile of a bacterial community using available genomes for each taxon's closest common ancestor (Langille *et al.*, 2013). These methods, whilst useful for generating hypotheses are of course hampered by the lack of data available in genomic databases and one would imagine the closest common ancestor would be of little functional relevance if no genomes exist for the 16S sequence of interest. To an extent metagenomic approaches (whereby genomic data from across a community is sequenced) have provided new insights into microbial function in soils.

However traditional short read annotation methods typically only enable us to make connections between functional genes and the environment they are found to be present in and therefore provide limited insight into the taxonomic origin of the functional genes of interest. The development of novel bioinformatic approaches however has enabled the linking of specific potentially uncultivated taxa to function through generating metagenome assembled genomes (MAGs)(Tyson *et al.*, 2004; Bowers *et al.*, 2017).

Here MAGs are constructed by assembling short reads into longer contigs and then grouping these contigs into bins based on nucleotide composition and/or read abundance ("coverage") information (Alneberg *et al.*, 2014). Initially these approaches have been used in environments with comparatively simple community composition (Sharon and Banfield, 2013) such as bio-film (Tyson *et al.*, 2004), cow rumen (Hess *et al.*, 2011) human gut (Di Rienzi *et al.*, 2013; Sharon *et al.*, 2013) and sludge bioreactor (Albertsen *et al.*, 2013). Indeed applying assembly methods to soils, presents a particular challenge due to its hyper-diverse nature, with previous estimates suggesting tera-base pairs (Tbp) of sequencing data would be required in order to sample a gram of soil sufficiently (Bunge, Epstein and Peterson, 2006; Howe *et al.*, 2014). Despite these complexities a number of studies have reported successful assembly of MAGs from soil metagenomes. Soil metagenome assembly has been used to provide new knowledge regarding the response of archaeal ammonia oxidisers to N fertilization (Orellana *et al.*, 2018), to gain insights on the impact of deforestation on microbial contributors to the carbon cycle (Marie E. Kroeger *et al.*, 2018), and to obtain the first complete genome of a novel *Pseudomonas* taxon (*Candidatus Pseudomonas sp. strain JKJ-1*) (White *et al.*, 2016).

5.1.1 Chapter aims

In **chapter 2** I demonstrated that most soil bacterial taxa have predictable responses to soil pH across large landscape scale gradients. Since pH can be modelled and predicted using climate, geological and land use information (Cosby *et al.*, 2001), the possibility to spatially predict bacterial abundances is now a reality. However, since current models are based on 16S rRNA gene sequences, the challenge is now to discover more about the functional capacities of the dominant, often non cultured organisms, through metagenomic binning. In

chapter 3 I used direct annotation of short reads to demonstrate the effects of land use change, alongside other soil and environmental parameters in affecting the abundance of bacterial functional genes. I now wish to use this data to explore whether wider functional genomic information can be extracted for a number of dominant soil bacterial taxa. I therefore seek to assemble the metagenomic reads described within **chapter 3.** Briefly, 96 HiSeq shotgun metagenomes were sequenced from soils from ten geographically distributed sites featuring paired land use contrasts.

The specific aims of this chapter are:

i. To determine whether it's possible to generate quality MAGs from UGRASS soils: Given the hyper diverse nature of soils and the large size of the UGRASS dataset (here 88 individual samples), I seek to determine if it is computationally possible to generate near complete assemblies from the dataset. The utility of the assembly approach will be assessed in terms of bin quality (contamination and completeness statistics) and whether these bins (regardless whether they may be MAGs or "community genomes") could provide new ecologically meaningful insights (aims ii and iii).
ii. To establish linkages between functional gene composition and phylogeny of dominant soil genomes: To identify whether functional gene content is largely homogenous across taxa, or whether different taxonomic groupings possess unique genes and to assess whether these genes are of relevance to soil function and biogeochemical cycling.
iii. To gain insights into the relationship between pH and land use preference and functional gene composition: To assess the extent to which bin functional gene composition is related to the bins environmental distributions and what can possibly be inferred from this in terms of how resilient microbial communities are to environmental change.

5.2 Methods

5.2.1 Soil sampling

Samples were collected between April and August 2015 as part of the Soil Security programme's UGRASS project. Paired sample sites were chosen where pristine fields were adjacent to intense grasslands or arable sites. A 100 m transect was used to take 5 pairs of cores (15cm depth, 5cm diameter) at the boundary of the two intensities every 25 m. The total number of individual samples collected in the survey were approximately 450 samples. However, metagenomic analyses was only conducted on 96 samples, encompassing 11 sites which significantly differed in organic matter content across the paired management contrasts, with no intermediate or reverting treatments assessed. The present study assessed 88 of these samples, focussing explicitly on high-low management intensity contrasts. All sample site information is detailed within **Table.3.1** (samples from the Rothamsted Highfield Bare fallow plot were excluded from this analyses, to focus on the arable/improved grasslands v unimproved grasslands contrasts). Surface litter was removed from soil cores. Soil cores were homogenised wet without sieving prior to subsampling for DNA extraction.

5.2.2 Metagenome Sequencing

DNA was extracted from 2g of soil using the power max soil DNA isolation soil kit, and subsequently purified using a millipore amplicon ultra buffer exchange. 96 Illumina libraries were constructed using the Illumina TruSeq library preparation kit (insert size < 500- 600 bp). Paired-end sequencing (2 x 150 bp) was conducted using the Illumnia Hiseq 4000 platform, 96 indexed libraries were multiplexed across 8 lanes and generated in excess 280M clusters per lane.

5.2.3 Metagenome assemblies

Illumina adaptor sequences were detected and removed from reads using Cutadapt 1.2.1, prior to trimming with Sickle 1.200 with a minimum window quality score of 20. Reads shorter than 20 bp after trimming were discarded. In initial work I trialled assembling all

samples with MEGAHIT (Li *et al.*, 2015) (on a local server with 1TB RAM), both with default and 'meta-large' parameters, which employs an altered set of k-mers in assembly. These runs were however unsuccessful and failed at the build SdBG (succinct de Bruijn graphs) stage with exit codes referring to lack of memory. As I did not have access to a server/or cluster with a single node with a larger volume of RAM, the final assemblies were conducted per site, comprising 8 samples per site encompassing two treatments of high and low intensity management. Per site assemblies were conducted with MEGAHIT with a minimum contig length of 1000 with default parameters.

Contigs were taxonomically annotated using the NCBI Blast non-redundant protein database with Kaiju. Kaiju first translates DNA into six possible reading frames before annotating proteins with maximum exact matches (MEM's). Upon finding a hit the taxonomic identifier of the hit is outputted, if there are multiple matches to hits with differing taxonomic identifiers, Kaiju uses the Lowest Common Ancestor (LCA) of all taxonomic identifiers for annotation (Menzel, Ng and Krogh, 2016). The taxonomic names assigned by the NCBI database will be used throughout, though we acknowledge these names may vary to those used within the Genome Taxonomy Database (GTDB) (Parks *et al.*, 2018, 2020). Functional annotations were then assigned to contigs using SEED subsystems (a hierarchical classification system, based on biological groupings related by process or structure) (Overbeek *et al.*, 2005). This was conducted using kmers (k=9) to detect similarity using standalone Rast server (Aziz *et al.*, 2012; Overbeek *et al.*, 2014).

5.2.4 Binning contigs

Contigs were grouped using a manual binning approach, taking into account both the Kaiju assigned taxonomy and clustering based on tetramer content. Coverage was not taken into account, as I lacked coverage information across all sites, given that I assembled reads per site (due to computational costs). The manual binning pipeline was developed with my supervisor and will now be described in greater detail (with an example output shown in **Fig.A5.1**). Contigs longer than 3000 bp were selected based on broad phylogenetic annotation (Kaiju), and tetramer frequencies were calculated using multi-metagenome (perl script accessible from https://github.com/MadsAlbertsen/multi-

metagenome/blob/master/R.data.generation/calc.kmerfreq.pl) (Albertsen, Philip Hugenholtz, et al., 2013). Tetramer frequency tables were then assessed using t-SNE ordination (a dimension reduction technique) in R using Rtsne library to visualise similarities in tetramer content of contigs using a perplexity parameter of 40 (estimate of number of neighbouring points each point has). Bins were then manually selected using a manual gate with gatepoints R library based on visual inspection of the clusters within the ordination plot. Bins were curated using a hierarchical approach whereby large clusters were first identified, before identifying smaller sub clusters within the larger grouping.

In order to calculate the abundance of each metagenomic bin per sample, all reads from each sample were mapped to each contig within each bin using bowtie2 with default parameters. Frequencies of bins across sites were then calculated simply by summing the reads mapped back to contigs within each bin. The relative proportion of each bin per sample was calculated by dividing the total number of reads mapped to a bin by total number of reads mapped to all contigs within that sample.

5.2.5 Completeness and Contamination of Bins

Checkm was used to assess the quality of bins using the lineage workflow with default parameters. This pipeline first infers bin lineage by placing each bin within a reference tree of genomes compiled from the Integrated Microbial Genomes (IMG) database. It then uses HMMER profiles to scan the genomes for marker genes specific to the bins inferred phylogeny (largely consisting of ribosomal proteins and RNA polymerase domains). Before using the presence and absence of these genes to calculate bin contamination and completeness (Parks *et al.*, 2015).

5.2.6 Land use associations

A random forest model was used to identify taxa that were important in discriminating land use (arable/ grassland) based upon bin abundance across sites. Random forest uses an ensemble of decision trees, with each tree using a subsample of data and variables. Bins that were only present in 30% of samples were discarded. Data was split into a test and training dataset (60%/40% split of samples) model parameters were then tuned using the

training dataset with a cross validation K fold of 10. Dufrene-legendre indicator analyses was also used to establish which bins were grassland or arable indicators specifically (Dufrene & Legendre, 1997). Dufrene-legendre indicators take into account both the specificity of a variable (i.e. how specific it is to a particular sample type) and fidelity (i.e. how consistently it occurs within that sample type).

5.2.7 pH distributions

Huisman-Olff-Fresco models were used to determine each individual bins response to pH using the R package eHOF with a poisson error distribution. Model choice was determined using AIC and bootstrapping methods implemented with the eHOF package, whereby the model with the lowest AIC was initially chosen and its robustness then tested by rerunning models on 100 bootstrapped datasets (created by resampling with replacement). In cases where the most frequently chosen model in the bootstrap runs was different to the initial model choice, the most common bootstrap choice was selected. The pH-bin response curves classified by the HOF models include I: no significant change in abundance in response to pH, II: an increasing or decreasing trend, III: increasing or decreasing trend which plateaus, IV: Increase and decrease by same rate (unimodal) and V: Increase and decrease by different rates causing skew. I classified bins pH preferences using model optima, if the optima was below pH 5.2 I classified it as acidic, based on previous data showing this represented a critical threshold for bacterial communities (Griffiths et al., 2011). This pH value also represents a critical threshold in microbial functioning.
A second threshold was designated at pH 7, with bins exhibiting an optima above this being classed as neutral, and those between 5.2 and 7 classed as “mid”. Plateau model shapes (model III), were sometimes more difficult to classify, since two optima are provided which span the plateau, and in some cases these crossed the pH 5.2 and 7 thresholds.

5.2.8 Functional indicators

Functional indicators of taxonomic groupings were determined using dufrene-legendre indicator analyses (Dufrene & Legendre, 1997) and bin functional gene content (excluding bins with completeness < 80%) based upon SEED annotations. Functional genes Indicative of

grassland and arable soils were also determined using dufrene-legendre indicator analyses on unassembled short read annotations (**chapter 3**), to compare with taxonomic indicators.

5.3 Results

5.3.1 Metagenome statistics

88 metagenome samples were analysed from grassland and arable sites with ~15000000 to ~37000000 non overlapping paired end reads per sample and an average trimmed read length of ~148 bp. From these reads, it was possible to assemble a total of ~7000000 contigs, with a minimum length of 1000 bps. All contigs with a length > 3000 bp were manually grouped into taxonomic bins, by first selecting contigs by taxonomic grouping and then visualising difference between contigs using t-SNE analyses of tetramer content. I manually identified bins based on visual inspection of clusters within the 2D ordination plots, taking a hierarchical approach to curating bins whereby I identified large clusters first, before identifying smaller sub clusters within the larger grouping.

Upon running CheckM, 32% of bins had a completeness statistic of > 80% and 53% had a "contamination" statistic of < 20% (**Table.A5.1**). The higher levels of contamination in 47% bins are unsurprising given that bins were curated hierarchically, therefore one would expect the broader level of bins to contain higher levels of redundancy. Additionally, the contigs were assembled from numerous sample assembly runs further contributing to contig redundancy that would not occur in a single assembly run. Of course, the hyper-diversity of soil microbial communities is also a likely contributor to the contamination observed, though the approach implemented here does not allow for a true estimation of real "contamination" levels versus "redundancy". Whilst the majority of bins could not be considered candidate MAGs without further refinement, four bins were of particularly good quality based upon CheckM statistics (completeness >85% and contamination of < 10%): Actinobacteria_1_7, Chloroflexi_1_15, Acidobacteria_1_8 and Thaumarchaeota_1_3.

5.3.2 Land use and bin abundance

I next examined the environmental distributions of bins, through calculating bin abundance across sites. The number of reads mapped to contigs within each bin were summed, before normalising this value by total number of reads mapped to all contigs. NMDS was used to examine similarity of bin abundances across land use types. As seen in **Fig.5.1**, arable soils appeared distinct from grasslands, consistent with what was seen within short read metagenomics and amplicon analyses (**chapter 3**), whilst intensive and unintensive grasslands appeared largely indistinguishable. Separation of arable and grassland soils can be seen within both NMDS1 which appears to be strongly related to pH (**Fig.5.1** green contours) and NMDS2.

I then sought to identify which specific bins were important in discriminating between arable and grassland soils through generating a random forest model based upon bin abundance. This model had a classification accuracy of 98% based upon cross validation on the training set and 97% accuracy within an independent test set of samples. The most important determinants of land use change (based upon contribution to model accuracy) were Chloroflexi_1_1_1 (mean decrease of accuracy 6.84), Thaumarchaeota_1_5_4 (mean decrease accuracy 5.069), Chloroflexi_1_1 (mean decrease accuracy 5.055), Thaumarchaeota_1_6 (mean decrease accuracy 4.96). Of the 25 most important discriminators of land use type 11 were *Thaumarchaeota* (Ammonia oxidising archaea) and 19 were arable indicators (based upon dufrene-legendre indicator analyses).

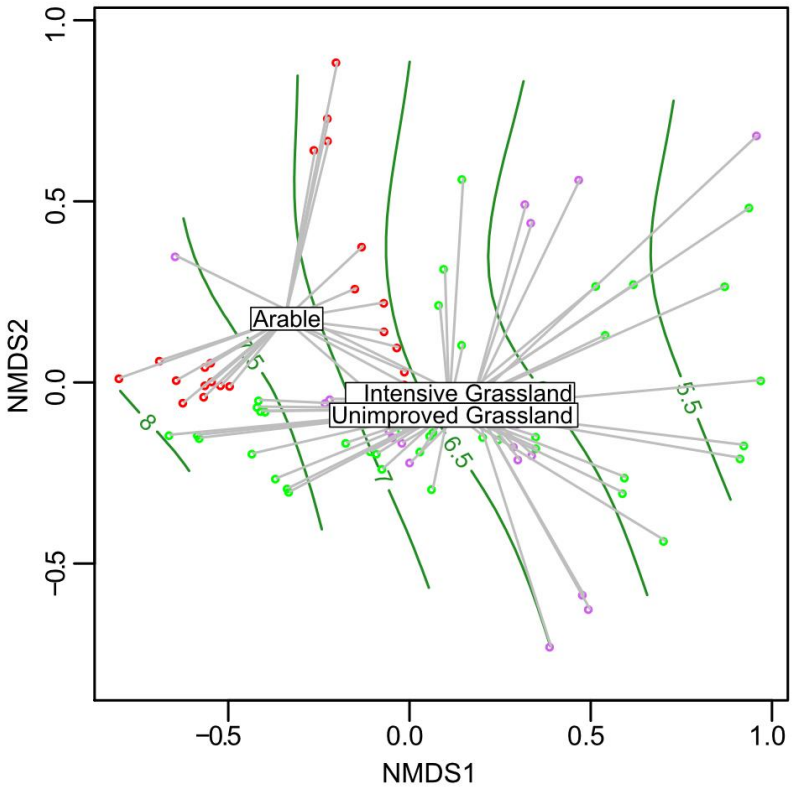


Fig.5.1. NMDS of metagenomic bin abundance (curated based on taxonomic annotations and tetramer content) across arable and grassland soils (bin abundance calculated through mapping short reads back to assembled contigs). Point colour and labels represent land use intensity. Green contours represent pH gradient.

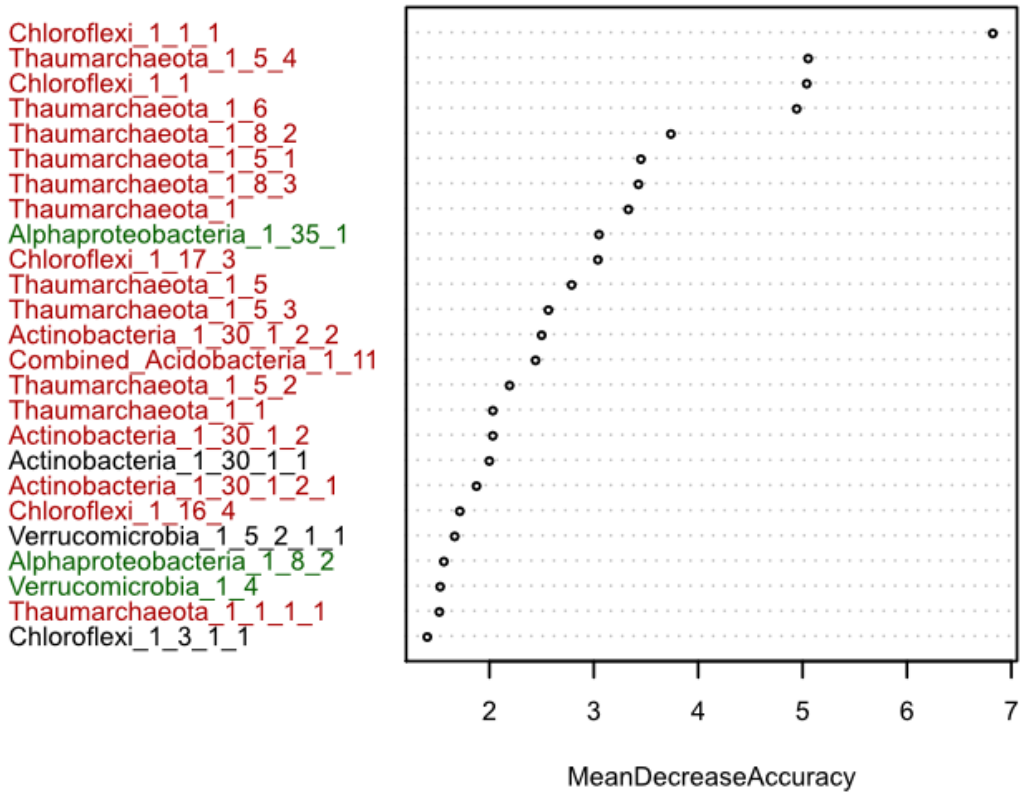


Fig.5.2. Random forest mean decrease in accuracy plot for metagenomic bins (curated based on taxonomic annotations and tetramer content) discriminating between soil land use. Bins with higher mean decreases in accuracy are stronger classifiers of land use. Colour of bin indicates whether the bin is an indicator of arable (red), grasslands (green) or not a significant indicator (black). These indicators were determined through a separate dufrene-legendre indicator analyses.

5.3.3 Functional profiles of bins

In order to gain insights into the functional content of bins, I annotated contigs using SEED and visualised the presence and absence of each functional gene per bin using t-SNE (excluding bins with completeness < 80%). As binning was conducted hierarchically and therefore some bins share contigs, this data is presented both with all bins (**Fig.5.3a**) and with bins at the broadest level of clustering within each taxonomic grouping (with no shared contigs) (**Fig.5.3b**). As seen in both **Fig.5.3a** and **Fig.5.3b** bins broadly appear to cluster by taxonomic grouping, with *Thaurmarchaeoata* forming a tighter cluster in comparison to other taxonomic groupings and *Actinobacteria* appearing more sparsely distributed. The pH optima of bins represented by point size (calculated using HOF models) did not appear as influential to clustering, although there did appear to be sub-clusters of bins related to pH preference within *Acidobacteria*.

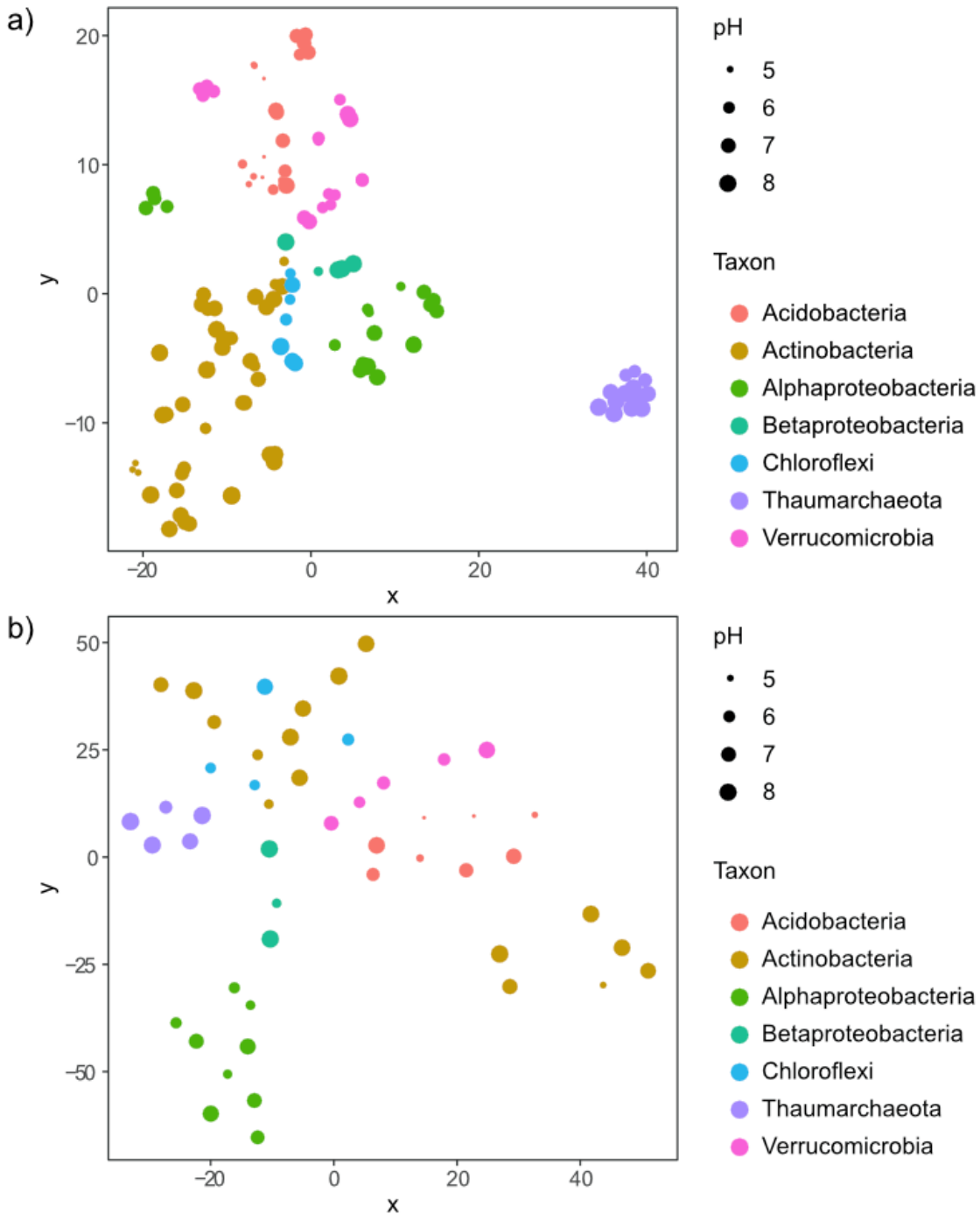


Fig.5.3. t-SNE of functional gene presence and absence per bin (excluding bins with a completeness of < 80%). Point colour represents taxonomic annotation (annotated with Kaiju), point size represents bin pH optima (based upon HOF models). **a)** shows all hierarchically curated bins (including bins with shared contigs**) b)** shows bins curated at the broadest level of clustering within taxonomic grouping (with no shared contigs).

5.3.4 Functional indicators of broad taxonomic groupings

I next sought to determine which functional genes were indicators of specific taxonomic groupings through using dufrene-legendre indicator analyses on functional gene presence and absence within the bins curated. I then compared these indicators with dufrene-legendre indicators of land use based on relative gene abundance using short read annotations (**chapter 3**). As seen in **Fig.5.4** there are clear clusters of *Thaumarchaeota* indicators within the protein metabolism subsystem. There are also prominent clusters of arable indicators within the protein metabolism subsystem, although these do not overlap with the *Thaumarchaeota* indicators specifically. Similarly to the protein metabolism subsystem, the DNA metabolism functional class also appears to consist of numerous arable indicator genes. There are many clusters within the tree consisting of a combination of *Alphaproteobacteria* and *Betaproteobacteria* indicators. Flagellar and motility indicators (segment 23, **Fig.5.4**) for example are almost exclusively indicators of *Alphaproteobacteria* and *Betaproteobacteria*. There are also numerous indicators of *Alphaproteobacteria* and *Betaproteobacteria* within sulfur (segment 11, **Fig.5.4**) and phosphorus metabolism (segment 9, **Fig.5.4**) alongside clusters of grassland indicators. Nitrogen metabolism genes (segment 6, **Fig.5.4**) also contained a large amount of *Alphaproteobacteria* and *Betaproteobacteria* indicators as well as a clear cluster of grassland indicators.

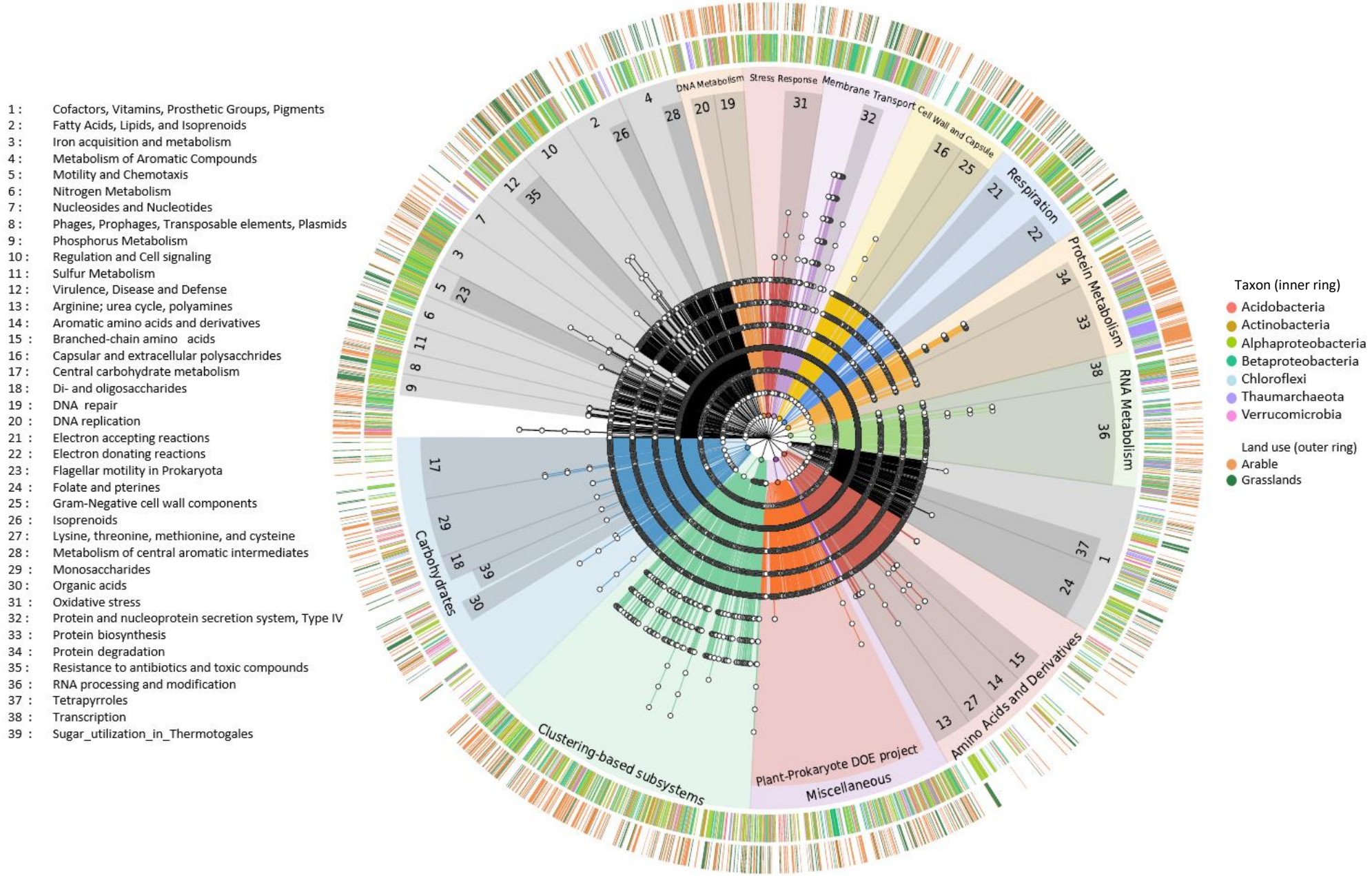


Fig.5.4. Descriptive tree of functional genes, arranged by seed subsystem hierarchy. Inner ring depicts the taxonomic grouping the gene is indicative of based upon their presence/absence within metagenomic bins, whilst the outer ring depicts if they were an indicator of land use in short read metagenomic analyses. All indicators were determined using dufrene-legendre indicator analyses.

I next further examined shared land use and taxon indicators within subsystems of relevance to soil ecosystem services specifically relating to biogeochemical cycling. The bar plots within **Fig.5.5** show all land use indicator genes within each subsystem studied, alongside shared land use and taxon indicators, for completeness genes that were taxonomic indicators but not land use are included in **Fig.A5.1**. Within sulfur metabolism, grassland indicators included numerous sulfur oxidation genes, many of which were also indicators of *Alphaproteobacteria* (**Fig.5.5a**). Grassland indicators also included multiple genes annotated as sulfate reduction associated complexes, though none of these genes were significant indicators of taxon. Taurine (sulfur containing amino acid) utilisation genes were also found to be grassland indicators and were also indicative of *alphaproteobacteria* and *betaproteobacteria*. Arable indicators within sulfur metabolism included numerous inorganic sulfur assimilation genes, a small subset of which were indicators of various taxonomic groupings (*Actinobacteria*, *Alphaproteobacteria*, *Thaumarchaeota* and *Verrucomicrobia*).

Within the phosphorus metabolism class, grassland indicators were comprised of a larger amount of alkylphosphonate utilization genes (**Fig.5.5b**) and all of these genes were also indicators of *Alphaproteobacterial* bins. Arable indicators included multiple high affinity phosphate transporter and control of PHO regulation genes, a subset which were indicators of various taxonomic groupings (*Actinobacteri*a, *Alphaproteobacteria*, *Chloroflexi*), no genes within this subsystem were grassland indicators. Arable indicators also included P uptake genes, a subset of which were also indicators of *Alphaproteobacteria* and *Chloroflexi*. Within Nitrogen metabolism, grassland indicators were comprised of numerous nitrogen fixation genes (**Fig.5.5c**), a subset of which were also indicators of various taxa (*Alphaproteobacteria*, *Betaproteobacteria* and *Acidobacteria*). Within arable indicators there were numerous genes annotated to nitrate and nitrite ammonification, denitrification and dissimilatory nitrite reductase subsystems. The majority of these genes were not specifically associated with any particular taxonomic groupings.

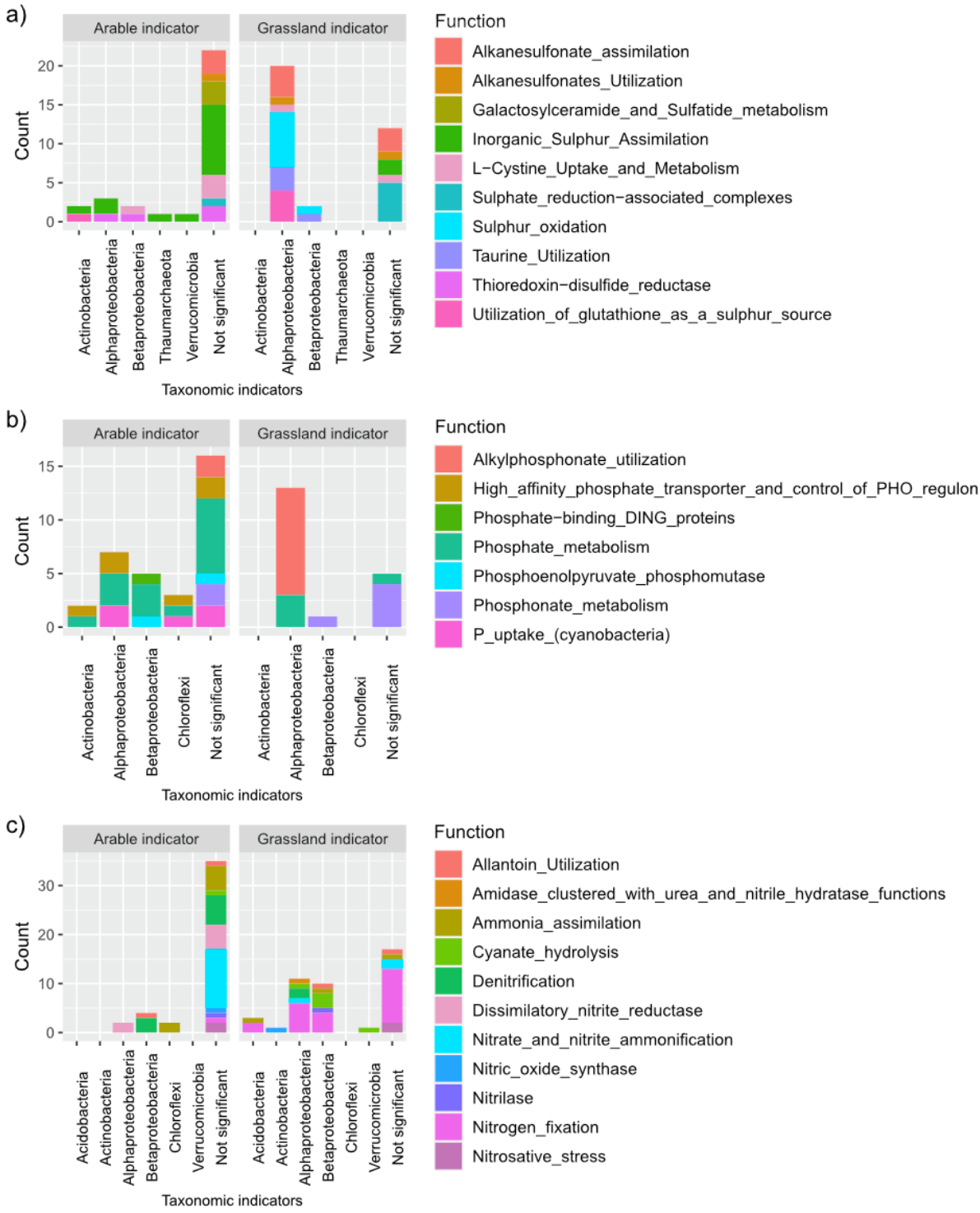


Fig.5.5. Gene indicators (determined through dufrene-legendre indicator analyses) of taxonomic grouping and land use within **a)** Sulfur metabolism subsystem, **b)** phosphorus metabolism and **c)** nitrogen metabolism. Taxon indicators are based upon presence and absence of functional genes within bins, land use indicators are based upon gene abundance from short read annotations.

I next focussed on the specific genes within the nitrogen metabolism subsystem that were land use and taxonomic indicators. As seen in **Fig.5.6**, most nitrogen fixation genes (Nif) were indicators of grasslands (NifB, D, E, H, K, N, O, Q, S, T, U, W, X, Y, Z) alongside other Nif associated genes (NifX-associated, NifB-domain protein type 2 and probable iron binding protein from the HesB_IscA_SufA family in Nif operon). Numerous of these Nif grassland indicator genes were also indicators of *Alphaproteobacteria* (Nif N, W, X, NifX-associated protein and probable iron binding protein from the HesB_IscA_SufA family in Nif operon) and *Betaproteobacteria* (NifD, E, H, K). An alternate nitrogenase, vandium-dependent nitrogenase was also indicative of grasslands (VnfD, K). Interestingly the transcriptional repressor of Nif and GlnA genes was an indicator of arable soils. More broadly arable indicators included numerous indicators related to denitrification, specifically there were many Nir genes (NirB, D, F, H, J, K, N, S). Arable indicators also included multiple nitric oxide reductase (Nor) genes (NorB, qnorB and NorC, E, W). A number of Nor genes were also indicative of *Betaproteobacteria* (NorB, C, D, Q).

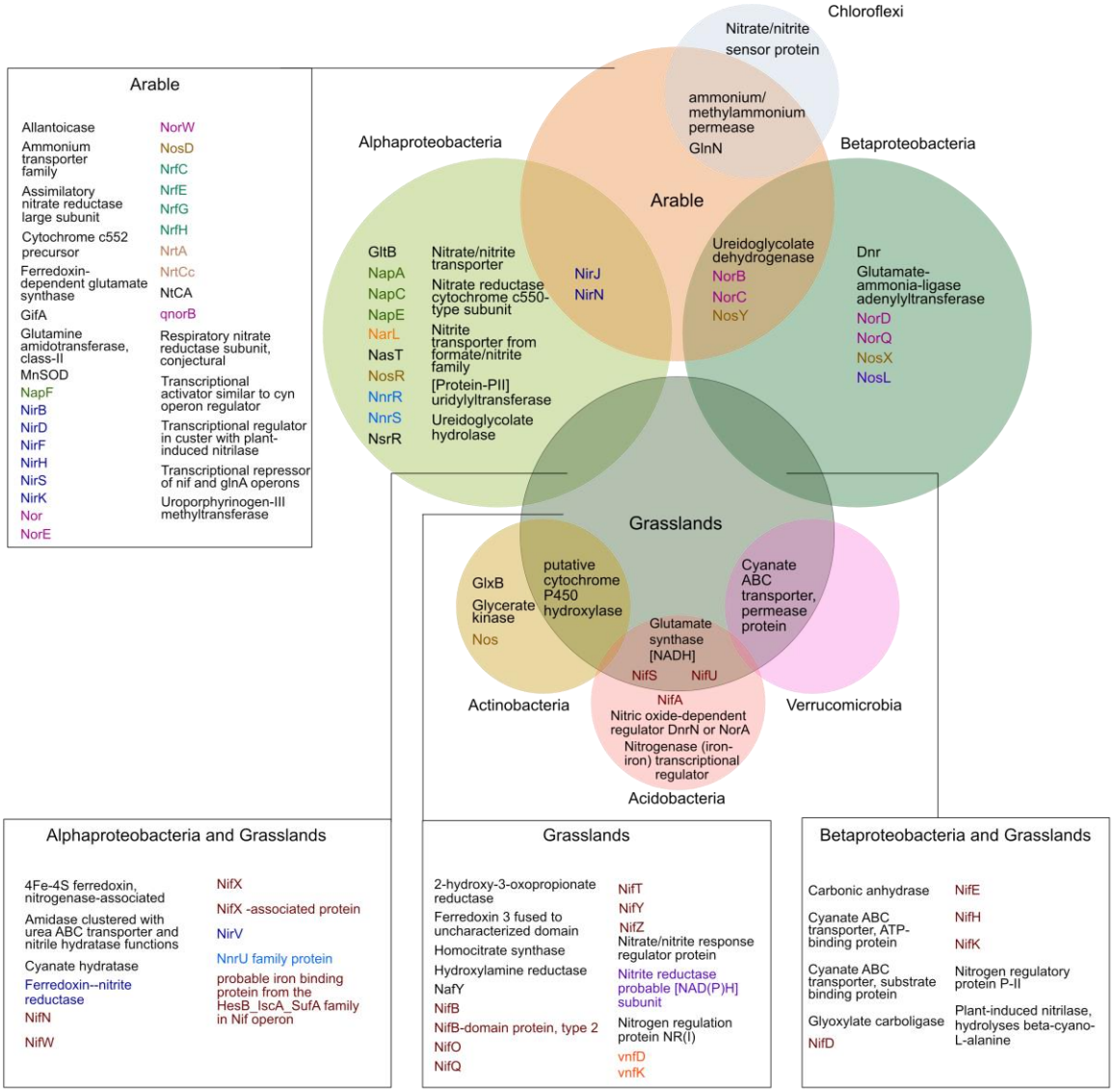


Fig.5.6. Illustrative venn diagram of genes within nitrogen metabolism subsystem which are indicative of land use and/or taxonomic grouping. Genes within groupings indicate the gene is an indicator of that taxon and/or land use based upon dufrene-legendre indicator analyses. Genes at the intersection of two groups e.g. Arable and *Betaproteobacteria* are indicators of both categories. Absence of a gene from a category signifies that the gene is not statistically indicative of that category, from this it cannot be inferred that the gene is completely absent from that land use / taxon. Taxon indicators are based upon presence and absence of functional genes within bins, land use indicators are based upon relative gene abundance from short read annotations. Colours represent groupings of closely associated functional genes for easy identification of nitrogen gene groupings (e.g. Nap, Nif, Nir, Nor, Nos etc) where multiple genes within that grouping occur. Size of grouping is not to scale with number of indicator genes within the grouping.

5.4 Discussion

5.4.1 Bins indicative of land use

I found that numerous *Thaumarchaeota* bins were important discriminators of arable and grassland soils and were indicators of arable soils specifically. Given that *Thaumarchaeota* are prominent ammonium oxidisers (oxidising ammonium to nitrite), this finding could be related to the nitrogen based fertilisers often applied to arable soils (Prosser and Nicol, 2012). Previous work assembling *Thaumarchaeota* genomes, found that *Thaumarchaeota* MAGs increased in abundance in response to the application of nitrogen fertilization, alongside other ammonia oxidising taxa (Orellana *et al.*, 2018). Soil nitrogen concentration has also been shown to be influential of ammonia oxidising archaeal composition in globally distributed soils based on archaeal marker gene analyses (amoA) (Pester *et al.*, 2012). Whilst other work looking at *Thaumarchaeota* responses to land use specifically, demonstrated an increase in abundance of amoA genes within managed turf grass/ lawn systems in comparison to land with minimal human impact (Epp Schmidt *et al.*, 2019).

Further, I found that few bins that were important discriminators of land use (identified within random forest analyses) were indicators of grassland soils (based upon dufrene-legendre indicator analyses). This could be considered unsurprising given that the grasslands studied within this work covered a wider range of soil properties, including pH in particular. Therefore within analyses contrasting the broad "grassland" category versus arable soils, there are unlikely to be consistently enriched bins across all types of grassland assessed. However it is intriguing that one of the bins of high importance to the random forest model that was a grassland indicator was an *Alphaproteobacteria*. *Bradyrhizobial* representatives of this group have previously been found to be of greater abundance in UK grassland soils compared with arable soils (Zhalnina *et al.*, 2013; Armbruster *et al.*, 2020) with both papers also identifying inverse associations between the *Bradyrhizobia* and *Thaumarcheota*.

5.4.2 Influences of niche and phylogeny on functional gene content

The phylogenetic classification of bins was far more influential on the broad functional content of bins, compared with pH niche. If taxon with shared pH responses do not possess similar functional gene content, this raises questions as to what enables certain taxa to occupy particular pH niches (**chapter 2**). One may expect taxa with similar pH optima to possess certain "effect traits" or genes enabling them to thrive within their specific pH niches. However, an explanation is that looking at the entire functional gene content collectively is too broad a means to demonstrate the influence of pH over functional genes. The presence of more specific functional genes may be more relevant to pH preference but many others may be ubiquitous across genomes (for example cellular machinery/housekeeping genes etc.)

As previously stated, the phylogenetic classification of bins appeared to be much more influential over functional gene content. *Thaurmatchaeota* in particular formed a tight cluster of bins in comparison to other taxonomic groupings. Given *Thaurmarchaeota* are archaeal and the other broad taxonomic groupings studied were bacterial, this apparent distinction in functional gene content is perhaps unsurprising especially since archaea are known to comprise highly differentiated cellular structures and machinery compared with bacteria (Woese, Kandler and Wheelis, 1990). Further work showed most functional indicators of *Thaumarchaeota* were related to protein metabolism. Given that *Thaumarchaeota* are prominent ammonia oxidisers (AOA), one would expect *Thaumarchaeota* indicative genes to have included amoA genes which catalyse ammonia (NH_3) to nitrite (NO_2^-) (the first rate limiting step of nitrification) (Pester, Schleper and Wagner, 2011)). However not only were amoA genes not found to be *Thaumarchaeota* indicators they were also not found within any of the *Thaumarchaeota* bins. Surprisingly, I observed a larger amount of ammonia monooxygenase gene annotations within short reads than in contigs, regardless of taxon (results not shown). It's possible that the kmer based functional gene annotation used was simply not as effective at annotating ammonia monooxygenase genes within contigs as in short reads. Additionally, as bins were partially curated based upon taxonomic annotation it's possible that amoA *Thaumarchaeota* contigs

have incorrectly been annotated as other taxonomic groupings, and therefore automatically excluded from the *Thaumarchaeota* bins.

5.4.3 Functional Indicators of taxon and land use within key biogeochemical cycles

To further assess the functional relevance of taxonomic change in soil microbial communities in response to land use, I compared taxonomic indicators (based upon their presence/absence within bins), with land use indicators (based upon short read annotations from **chapter 3**) within key biogeochemically relevant subsystems. Carbohydrate (C cycling) genes were not examined here, due to their large diversity, and detailed exploration in **chapter 4**. Within sulfur metabolism, I found that numerous grassland indicators were sulfur oxidation genes. Numerous sulfur oxidation genes, specifically SOX genes were also indicators of *Alphaproteobacteria*, which in itself is unsurprising as SOX genes are known to be well characterised within *Alphaproteobacteria* (Friedrich *et al.*, 2005). It is of note however that within sulfur metabolism genes, more generally, I saw a large number of shared *Alphaproteobacteria*-grassland indicators and much fewer shared *Alphaproteobacteria*-arable indicators. Numerous grassland indicator genes were also sulfate reduction associated complexes (although none of these genes were *Alphaproteobacteria* indicators). Interestingly, further inspection of these genes showed many of these sulfate reduction associated complex genes, encoded the DsrMKJOP complex which is a membrane spanning complex, which is also associated with sulfur oxidation in addition to sulfur reduction(Sander *et al*., 2006; Grein et al., 2010). Within arable indicators there were a number of inorganic sulfur assimilation genes. Given that agricultural soils are often associated with sulfur losses (thought to be related to the loss of biomass of crops, coupled with the decrease in usage of sulfur based fertilisers) (Lucheta and Lambais, 2012; Kumar *et al.*, 2018), an increase in sulfur assimilation genes could potentially be a response to sulfur deprivation and reliance on inorganic sulfur from fertilisers.

Within phosphorus metabolism genes, it is apparent that a number of arable indicators are annotated as control of PHO regulation genes, which are known to play a role in sensing and

regulating inorganic phosphate availability. In previous work, genes within this function have also been found in increased abundance within high intensity land use soils (annual cropland) (Liu *et al.*, 2018), and within low P soils (Oliverio *et al.*, 2020). Arable indicators also included a number of P uptake genes, this is also consistent with previous work where annual croplands were shown to possess an increased amount of phosphate uptake genes in comparison to native/tame grasslands (Liu *et al.*, 2018). Within grassland indicators there were numerous alkylphosphonate utilization genes, these were predominantly related to C-P lyase multienzyme complex which breaks highly stable C-P bonds within phosponates (Cook, Daughton and Alexander, 1978). These genes have also been found to be indicators of low phosphorus soils within previous work and have been linked with P starvation responses (Oliverio *et al.*, 2020). Therefore within both arable and grassland indicators, I observed genes associated with phosphorus poor soils/ phosphorus starvation. This is confounding as whilst it may be expected for arable soils to be phosphorus poor, one would not have the same expectation of grassland soils. Indeed, within paired land uses the grassland site consistently had increased phosphorus levels compared to the arable site within the same pairing. The explanation therefore is that in grassland soils, communities are more reliant on scavenging P from organic sources held within large stores of soil organic matter; whereas in arable soils where organic P is less available, there is likely greater reliance on externally applied inorganic P. In terms of taxonomic associations of genes I found that all grassland indicators of alkylphosphonate utilization (including numerous C-P lyase genes) were also indicators of *Alphaproteobacteria*. Indeed a large amount of *Alphaproteobacteria* C-P lyase genes in soil metagenomes has been reported elsewhere (Liu *et al.*, 2018).

Within nitrogen metabolism I found that numerous grassland indicators are nitrogen fixation genes. Nitrogen fixation is the process whereby dinitrogen gas (N_2) is converted to ammonia (NH_3) by nitrogenases. Nitrogenases exist in multiple subtypes including Mo-dependent nitrogenase, vanadium-dependent nitrogenase and iron-iron nitrogenase (Dos Santos *et al.*, 2012). Within this work most Mo-dependent nitrogenase associated genes (Nif), were found to be grassland indicators (Nif B, D, E, H, K, N, O, Q, S, T, U, W, X, Y, Z) including NifH and NifK which are known to encode Mo-dependent nitrogenases catalytic subunits. Genes encoding vanadium-dependent nitrogenase (Vnf) alpha and beta chain (Vnf

D, K) were also indicative of grasslands. I did not see any genes encoding iron-iron nitrogenase (Anf) within grassland indicators (although an iron-iron nitrogenase transcriptional regulator was indicative of *Acidobacteria*). This increase in nitrogen fixation genes within grasslands may be because the majority of these soils were not treated with nitrogen fertilizer, in contrast to arable sites, where all sites received nitrogen fertilizer. Therefore there may be more nitrogenase genes in these soils as a strategy to ensure the necessary amount of nitrogen is obtained (Regan, 2017), in an environment where mineral nitrogen is less readily available.

Within arable indicators there were numerous denitrification genes. These included several nitrite reductase associated genes (Nir B, D, F, H, J, K, N, S), nitrite reductase reduces nitrite to nitric oxide in an early stage of denitrification. NirK is known to encode copper-containing nitrite reductase/CU-Nir, whilst NirS is known to encode Heme containing nitrite reductase/ cd1-Nir (Sharma *et al.*, 2005). Arable indicators also included nitric oxide reductase genes (qNorB, Nor B, C, E, W). Nitric oxide reductase also catalyses a key reaction within denitrification whereby nitric oxide is reduced to nitrous oxide. These indicators include genes encoding key subunits of both cytochrome c-dependent nitric oxide reductase/cNor (NorB, C) and quinol-dependent nitric oxide reductase/qNor (qNorB) (Braker and Tiedje, 2003).

Rates of denitrification are known to typically increase within oxygen deprived soils with reduced pore structure which can be characteristics of wet agricultural soils (Philippot, Hallin and Schloter, 2007; Clark *et al.*, 2020). Moreover the application of nitrogen based fertilizers specifically has been associated with increased denitrification rates associated with large amounts of nitrogen from fertiliser being lost to the atmosphere (Kaiser *et al.*, 1996; Philippot, Hallin and Schloter, 2007). Surprisingly, an increase in denitrification related genes within high intensity land uses, has not been reported consistently within the literature with recent work showing fertilizer treatment did not impact upon the abundance of specific nitrite reductase and nitric oxide reductase genes (NirK, NirS, NosZi, NosZii) when measured with qPCR. The same study did however find that denitrification genes are relatively common across different phyla (Clark *et al.*, 2020) which is consistent with what

was found here as most denitrification genes were not indicative of the broad taxonomic groupings studied.

5.4.4 Approaches and workarounds to assembling in soils

Within this work I initially attempted to co-assemble reads across sample sites but found this was too computationally intensive given the resources available. Therefore I took a different assembly/ binning approach, whereby I first assembled reads per site and then binned across sites based upon tetramer counts. It is unsurprising that attempting to assemble across 88 soil samples was a challenge, given that the hyper-diversity in soils is known to make assembly considerably more difficult in comparison to environments with simpler communities (Howe *et al.*, 2014). Whilst this method of assembling per site enabled the manual curation of bins based on tetramer content, it also meant I was not able to integrate coverage information into the binning process, which is a commonly used methods to enhance the specificity of the resultant bins. Additionally, assembling per site (and therefore orchestrating 11 separate assembly runs) likely lead to increased contig redundancy than if I had assembled reads from all samples together in a single assembly run. This has made it a challenge to assess if redundancy within bins is due to the bins being contaminated by other phylotypes or due to the methodology pitfalls described. To address this, I plan to refine bins with high completeness through reassembling all reads mapped to the relevant contigs to further assess levels of contamination in future work.
If I were to reattempt assembling reads across samples in the future, approaches that could be considered include reducing the dataset prior to assembly (through discarding low abundant reads or digital normalisation (Brown *et al.*, 2012; Howe *et al.*, 2014)) or using an assembly method that can be executed on multiple nodes of a computing cluster (thus enabling more RAM to be used).

5.5 Conclusions

To conclude, this work has found that arable soils were highly distinct from grasslands in terms of metagenome assembled bin relative abundance, consistent with the short read metagenomic analyses of this dataset presented in **chapter 3**. Many *Thaumarchaeota* bins

were good discriminators between arable and grassland soils and were typically indicators of arable soils, with *Alphaproteobacterial* bins being more abundant in certain grassland soils. Broad taxonomic groupings were largely different from one another in terms of functional gene content, with *Thaumarchaeota* bins appearing particularly distinct. Based on short read analyses, land use caused large changes in genes relating to N, P, and S cycling indicative of the different nutrient acquisition strategies from organic and inorganic sources. Specifically, within grassland soils there was indication of N acquisition through direct fixation, with P and S being acquired from organic sources; whereas indicators of inorganic acquisition were found in intensified arable systems. Relating these broad changes with the functional gene content and ecology of specific bins allows for a mechanistic understanding of functional change, manifest through specific microbial taxon ecological responses.

5.6 Bibliography

Albertsen, M., Hugenholtz, P, *et al.* (2013) 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes', *Nature Biotechnology*, 31(6), pp. 533–538. doi: 10.1038/nbt.2579.

Alneberg, J. *et al.* (2014) 'Binning metagenomic contigs by coverage and composition', *Nature Methods*, 11(11), pp. 1144–1146. doi: 10.1038/nmeth.3103.

Amann, R. and Ludwig, W. (2000) 'Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology', *FEMS Microbiology Reviews*, 24(5), pp. 555–565. doi: 10.1111/j.1574-6976.2000.tb00557.x.

Armbruster, M. *et al.* (2020) 'Bacterial and archaeal taxa are reliable indicators of soil restoration across distributed calcareous grasslands', *European Journal of Soil Science*. doi: 10.1111/ejss.12977.

Aziz, R. K. *et al.* (2012) 'SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models', *PLoS ONE*, 7(10), pp. 1–10. doi: 10.1371/journal.pone.0048053.

Baas-Becking, L. G. M. (1934) *Geobiologie, of Inleiding Tot de Milieukunde: Met Literatuurlijst en Ind*, *The Hague, the Netherlands: W.P. Van Stoc- kum & Zoon*. Available at: https://scholar.google.com/scholar?oi=gsb90&q=Baas Becking%252C L.G.M. (1934) Geobiologie of inleiding tot de milieukunde. The Hague%252C the Netherlands W.P. Van Stockum %2526 Zoon (in Dutch).%250A&lookup=0&hl=en#0.

Banerjee, S. *et al.* (2019) 'Agricultural intensification reduces microbial network complexity and the abundance of keystone taxa in roots', *ISME Journal*, 13(7), pp. 1722–1736. doi: 10.1038/s41396-019-0383-2.

Bardgett, R. D. and Caruso, T. (2020) 'Soil microbial community responses to climate extremes: Resistance, resilience and transitions to alternative states', *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society Publishing. doi: 10.1098/rstb.2019.0112.

Bissett, A. *et al.* (2011) 'Long-term land use effects on soil microbial community structure and function', *Applied Soil Ecology*, 51(1), pp. 66–78. doi: 10.1016/j.apsoil.2011.08.010.

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. doi: 10.1038/nbt.3893.

Braker, G. and Tiedje, J. M. (2003) 'Nitric oxide reductase (norB) genes from pure cultures and environmental samples', *Applied and Environmental Microbiology*, 69(6), pp. 3476–3483. doi: 10.1128/AEM.69.6.3476-3483.2003.

Brown, C. T. *et al.* (2012) 'A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data'. Available at: http://arxiv.org/abs/1203.4802 (Accessed: 13 July 2020).

Bunge, J., Epstein, S. S. and Peterson, D. G. (2006) 'Comment on "Computational improvements reveal great bacterial diversity and high metal toxicity in soil".', *Science*, pp. 1387–1391. doi: 10.1126/science.1126593.

Clark, I. M. *et al.* (2020) 'Edaphic factors and plants influence denitrification in soils from a long-term arable experiment', *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-72679-z.

Cohan, F. M. and Koeppel, A. F. (2008) 'The Origins of Ecological Diversity in Prokaryotes', *Current Biology*. Cell Press, pp. R1024–R1034. doi: 10.1016/j.cub.2008.09.014.

Cook, A. M., Daughton, C. G. and Alexander, M. (1978) 'Phosphonate utilization by bacteria', *Journal of Bacteriology*, 133(1), pp. 85–90. doi: 10.1128/jb.133.1.85-90.1978.

Cosby, B. J. *et al.* (2001) 'Modelling the effects of acid deposition: Refinements, adjustments and inclusion of nitrogen dynamics in the MAGIC model', *Hydrology and Earth System Sciences*, 5(3), pp. 499–517. doi: 10.5194/hess-5-499-2001.

Dufrene, M., & Legendre, P. (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs, 67(3), 345-366. doi:10.1890/0012-9615(1997)067[0345:Saaist]2.0.Co;2

Dunbar, J. *et al.* (1999) 'Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning', *Applied and Environmental Microbiology*, 65(4), pp. 1662–1669. doi: 10.1128/aem.65.4.1662-1669.1999.

Epp Schmidt, D. J. *et al.* (2019) 'Metagenomics Reveals Bacterial and Archaeal Adaptation to Urban Land-Use: N Catabolism, Methanogenesis, and Nutrient Acquisition', *Frontiers in Microbiology*, 10, p. 2330. doi: 10.3389/fmicb.2019.02330.

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. doi: 10.1126/science.1153213.

Fichtner, A. *et al.* (2014) 'Effects of anthropogenic disturbances on soil microbial communities in oak forests persist for more than 100 years', *Soil Biology and Biochemistry*, 70, pp. 79–87. doi: 10.1016/j.soilbio.2013.12.015.

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. doi: 10.1073/pnas.0507535103.

Friedrich, C. G. *et al.* (2005) 'Prokaryotic sulfur oxidation', *Current Opinion in Microbiology*. Elsevier Current Trends, pp. 253–259. doi: 10.1016/j.mib.2005.04.005.

Green, J. L., Bohannan, B. J. M. and Whitaker, R. J. (2008) 'Microbial biogeography: From taxonomy to traits', *Science*, 320(5879), pp. 1039–1043. doi: 10.1126/science.1153475.

Grein, F., Pereira, I. A. C. and Dahl, C. (2010) 'Biochemical characterization of individual components of the Allochromatium vinosum DsrMKJOP transmembrane complex aids understanding of complex function in vivo', *Journal of Bacteriology*, 192(24), pp. 6369–6377. doi: 10.1128/JB.00849-10.

Griffiths, R. I. *et al.* (2011) 'The bacterial biogeography of British soils', *Environmental Microbiology*, 13(6), pp. 1642–1654. doi: 10.1111/j.1462-2920.2011.02480.x.

Hartmann, M. *et al.* (2015) 'Distinct soil microbial diversity under long-term organic and conventional farming', *ISME Journal*, 9(5), pp. 1177–1194. doi: 10.1038/ismej.2014.210.

Hess, M. *et al.* (2011) 'Metagenomic discovery of biomass-degrading genes and genomes from cow rumen', *Science*. doi: 10.1126/science.1200387.

Howe, A. C. *et al.* (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. doi: 10.1073/pnas.1402564111.

Jansson, J. K. and Hofmockel, K. S. (2020) 'Soil microbiomes and climate change', *Nature Reviews Microbiology*, 18(1), pp. 35–46. doi: 10.1038/s41579-019-0265-7.

Kaiser, E. *et al.* (1996) *Nitrous oxide release from cultivated soils: influence of different N-fertilizer types*.

Kroeger, *et al.* (2018) 'New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes', *Frontiers in Microbiology*, 9(JUL), pp. 1–13. doi: 10.3389/fmicb.2018.01635.

Kumar, U. *et al.* (2018) 'Diversity of Sulfur-Oxidizing and Sulfur-Reducing Microbes in Diverse Ecosystems', in. doi: 10.1007/978-981-10-6178-3_4.

Langille, M. *et al.* (2013) 'Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.', *Nature biotechnology*, 31(9), pp. 814–21. doi: 10.1038/nbt.2676.

Li, D. *et al.* (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. doi: 10.1093/bioinformatics/btv033.

Liu, J. *et al.* (2018) 'Long-term land use affects phosphorus speciation and the composition of phosphorus cycling genes in agricultural soils', *Frontiers in Microbiology*, 9(JUL), p. 1643. doi: 10.3389/fmicb.2018.01643.

Lucheta, A. R. and Lambais, M. R. (2012) 'Sulfur in agriculture', *Revista Brasileira de Ciência do Solo*. doi: 10.1590/s0100-06832012000500001.

Mackelprang, R. *et al.* (2018) 'Microbial community structure and functional potential in cultivated and native tallgrass prairie soils of the midwestern United States', *Frontiers in Microbiology*, 9(AUG), p. 1775. doi: 10.3389/fmicb.2018.01775.

Martiny, J. B. H. *et al.* (2015) 'Microbiomes in light of traits: A phylogenetic perspective', *Science*, 350(6261). doi: 10.1126/science.aac9323.

Menzel, P., Ng, K. L. and Krogh, A. (2016) 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', *Nature Communications*, 7, pp. 1–9. doi: 10.1038/ncomms11257.

NANNIPIERI, P. *et al.* (2020) 'Beyond microbial diversity for predicting soil functions: A mini review', *Pedosphere*. Soil Science Society of China, pp. 5–17. doi: 10.1016/S1002-0160(19)60824-6.

Nguyen, N. H. *et al.* (2016) 'FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild', *Fungal Ecology*, 20, pp. 241–248. doi: 10.1016/j.funeco.2015.06.006.

Oliverio, A. M. *et al.* (2020) 'The role of phosphorus limitation in shaping soil bacterial communities and their metabolic capabilities', *mBio*, 11(5), pp. 1–16. doi: 10.1128/mBio.01718-20.

Orellana, L. H. *et al.* (2018) 'Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization', *Applied and Environmental Microbiology*, 84(2). doi: 10.1128/AEM.01646-17.

Overbeek, R. *et al.* (2005) 'The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes', *Nucleic Acids Research*, 33(17), pp. 5691–5702. doi: 10.1093/nar/gki866.

Overbeek, R. *et al.* (2014) 'The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)', *Nucleic Acids Research*, 42(D1), pp. 206–214. doi: 10.1093/nar/gkt1226.

Parks, D. H. *et al.* (2015) 'CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

Parks, D. H. et al. (2018) 'A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life', Nature Biotechnology. Nature Publishing Group, 36(10), p. 996. doi: 10.1038/nbt.4229.

Parks, D. H. et al. (2020) 'A complete domain-to-species taxonomy for Bacteria and Archaea', Nature Biotechnology. Nature Research, 38(9), pp. 1079–1086. doi: 10.1038/s41587-020-0501-8.

Paul, W. M. *et al.* (2019) 'A Review of the Role of Anthropogenic Effects on Microorganisms in Soil', *Journal of Agriculture and Ecology Research International*, 16(4), pp. 1–16. doi: 10.9734/jaeri/2018/44994.

Pester, M. *et al.* (2012) 'AmoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions', *Environmental Microbiology*, 14(2), pp. 525–539. doi: 10.1111/j.1462-2920.2011.02666.x.

Pester, M., Schleper, C. and Wagner, M. (2011) 'The Thaumarchaeota: An emerging view of their phylogeny and ecophysiology', *Current Opinion in Microbiology*, 14(3), pp. 300–306. doi: 10.1016/j.mib.2011.04.007.

Philippot, L., Hallin, S. and Schloter, M. (2007) 'Ecology of Denitrifying Prokaryotes in Agricultural Soil', *Advances in Agronomy*. Academic Press, pp. 249–305. doi: 10.1016/S0065-2113(07)96003-4.

Prosser, J. I. and Nicol, G. W. (2012) 'Archaeal and bacterial ammonia-oxidisers in soil: The quest for niche specialisation and differentiation', *Trends in Microbiology*. Elsevier Current Trends, pp. 523–531. doi: 10.1016/j.tim.2012.08.001.

Ramette, A. and Tiedje, J. M. (2007) 'Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem', *Proceedings of the National Academy of Sciences of the United States of America*, 104(8), pp. 2761–2766. doi: 10.1073/pnas.0610671104.

Regan, K. (2017) *Linking microbial abundance and function to understand nitrogen cycling in grassland soils*.

Di Rienzi, S. C. *et al.* (2013) 'The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria', *eLife*, 2013(2). doi: 10.7554/eLife.01102.001.

Sander, J., Engels-Schwarzlose, S. and Dahl, C. (2006) 'Importance of the DsrMKJOP complex for sulfur oxidation in Allochromatium vinosum and phylogenetic analysis of related complexes in other prokaryotes', *Archives of Microbiology*, 186(5), pp. 357–366. doi: 10.1007/s00203-006-0156-y.

Dos Santos, P. C. *et al.* (2012) 'Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes', *BMC Genomics*, 13(1), p. 162. doi: 10.1186/1471-2164-13-162.

Sharma, S. *et al.* (2005) 'Diversity of transcripts of nitrite reductase genes (nirK and nirS) in rhizospheres of grain legumes', *Applied and Environmental Microbiology*, 71(4), pp. 2001–2007. doi: 10.1128/AEM.71.4.2001-2007.2005.

Sharon, I. *et al.* (2013) 'Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization', *Genome Research*, 23(1), pp. 111–120. doi: 10.1101/gr.142315.112.

Sharon, I. and Banfield, J. F. (2013) 'Genomes from metagenomics', *Science*. Science, pp. 1057–1058. doi: 10.1126/science.1247023.

Suding, K. N. *et al.* (2008) 'Scaling environmental change through the community-level: A trait-based response-and-effect framework for plants', *Global Change Biology*, 14(5), pp. 1125–1140. doi: 10.1111/j.1365-2486.2008.01557.x.

Sui, X. *et al.* (2019) 'Land use change effects on diversity of soil bacterial, Acidobacterial and fungal communities in wetlands of the Sanjiang Plain, northeastern China', *Scientific Reports*, 9(1). doi: 10.1038/s41598-019-55063-4.

Tyson, G. W. *et al.* (2004) 'Community structure and metabolism through reconstruction of microbial genomes from the environment', *Nature*, 428(6978), pp. 37–43. doi: 10.1038/nature02340.

Weiher, E. and Keddy, P. A. (1995) 'Assembly Rules, Null Models, and Trait Dispersion: New Questions from Old Patterns', *Oikos*, 74(1), p. 159. doi: 10.2307/3545686.

White, R. A. *et al.* (2016) 'Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes', *mSystems*, 1(3). doi: 10.1128/msystems.00045-16.

Woese, C. R., Kandler, O. and Wheelis, M. L. (1990) 'Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya', *Proceedings of the National Academy of Sciences of the United States of America*, 87(12), pp. 4576–4579. doi: 10.1073/pnas.87.12.4576.

Zhalnina, K. *et al.* (2013) 'Ca. nitrososphaera and bradyrhizobium are inversely correlated and related to agricultural practices in long-term field experiments', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2013.00104.

Chapter 6

Discussion and Future Research

6.1 Introduction

This chapter synthesises how my work contributes to an improved understanding of the specific responses of microbial taxa to environmental gradients, and the functional implications of microbial change with respect to land use. Further I will identify how new developments in molecular approaches and digital technologies could be used to expand upon this work to further advance understanding of microbial response and effect traits, in order to develop a more predictive framework for soil microbial ecology.

6.2 Synthesis of findings

6.2.1 pH effects on microbial taxa and function

Throughout this thesis I have shown how natural soil variance contributes strongly to bacterial community structure. These edaphic characteristics can also be heavily influenced by land use, since land use is known to be strongly related to soil pH (Malik et al., 2018). For example the use of ammonium and sulfur fertilisers are associated with soil acidification, whilst liming practices are often implemented in an effort to neutralise soils and promote plant productivity (Goulding, 2016; Tian & Niu, 2015). Numerous microbial biogeography studies have highlighted the influence of pH on soil microbial communities, though these responses have typically been measured in terms of microbial diversity or shifts in broader taxonomic groupings (Fierer & Jackson, 2006; Griffiths et al., 2011). If we are to develop a better understanding of how change in microbial communities is driven by external forces such as land use, we first need to build better synthesis of how individual taxa respond to change, and then couple this with a better functional understanding of the traits possessed by these taxa.

This broad reporting of microbial phylum level responses is common in soil microbial ecology, primarily due to the space constraints of conventional academic reporting, and the challenge of interpreting large multi-species datasets. It is likely therefore that more specific ecological trends within discrete taxa are often overlooked. Whilst taxonomic annotations of marker gene sequences are highly accessible within databases such as SILVA (Quast et al., 2013) and Greengenes (DeSantis et al., 2006), the ecological responses of discrete taxa have

previously been hidden in datasets (which may or may not be accessible). Given there are hundreds of new molecular datasets being generated across the world, it is startling that thus far there has been no explicit framework to report and synthesise ecological responses at a finer phylogenetic resolution. The data regarding OTU taxon responses exists for a wealth of studies for varying soil conditions and geographical locations, if these taxon responses were accessible and able to be synthesised, there is real potential to greatly accelerate our understanding of taxon responses to a suite of abiotic factors, as a prelude to prediction of functional change.

Within **Chapter 2** I made steps to address these issues. Firstly, I modelled pH responses at the OTU level using a large amplicon dataset consisting of > 1000 soil samples collected across Britain. I then assigned pH classifications to OTU's and found that whilst pH preference could vary substantially within phyla there were clear subclades of phylogenetically related taxa with shared pH preference. For example, there was a clear acidophilic clade within *Acidobacteria* and prominent neutral clade in *Verrucomicrobia*, highlighting the importance of reporting response traits at a finer phylogenetic resolution. I made these OTU level responses available through developing ID-TaxER (https://shiny-apps.ceh.ac.uk/ID-TaxER/), an online application enabling querying of 16S sequences to obtain pH response trait information. Thus, demonstrating a relatively straight forward approach to making OTU level information more easily accessible. Further I found it was possible to predict pH responses within an independent query dataset using the modelled pH responses from our dataset. This demonstrates that community structure is broadly predictable at the OTU level (clustered at 97% similarity) using simple soil abiotic predictors.

Within later chapters (**3, 4**) I observed a strong relationship between soil pH and microbial functions, assessed through metagenomics. In **chapter 3**, the functional gene content of the 96 soil metagenomes studied also appeared to be primarily driven by pH, with land use a secondary driver (see section below). Two specific clusters of samples were identified, with the first containing samples with soil pH of between 6.24 and 8.12; and the second more acidic group containing samples with soil pH of between 4.83 and 6.9. This variation in functional gene content suggests a large influence of soil pH on functional gene content, with changes in taxonomic biodiversity potentially being associated with change in

functional gene content. However, whilst I identified these pH responsive genes using a correlation network approach; given time constraints and my interest in land use relevant indicators, no further explicit analyses of these pH responsive genes was conducted. Further ongoing work therefore needs to evaluate these specific pH responsive genes, to determine relevance for soil ecosystem services. Indeed, previous work on a smaller metagenome study has identified high and low pH indicators of relevance to soil processes. More C and N direct fixation genes were found within low pH soils and more transporter mediated organic C and N acquisition genes were seen in high pH soils (Malik et al., 2017). Moreover work employing ^{14}C-labeling has demonstrated that soil pH strongly correlates with microbial carbon use efficiency (CUE) across a range of agricultural soils (Jones et al., 2019), providing further evidence of soil pH influencing vital soil processes. In order to directly link pH related change in taxonomic biodiversity with change in soil processes; it remains to be determined whether pH responsive taxa possess specific functional traits conferring altered functionality (e.g. specific C or N cycling genes) or whether they may be related to differences in cellular structure/generic cellular processes which are likely to invoke more complex controls on soil functioning.

Within **chapter 4** I delved deeper into pH responses of functionally relevant genes by examining relative abundances of β-glucosidase genes within Park Grass samples maintained at pH5 and pH7 for ~ 150 years (Silvertown et al., 2006). β-glucosidases were studied both due to their essential role in organic matter decomposition and carbon cycling, and also to build upon recent work which used physiological enzymes to demonstrate differences in activity of β-glucosidases from pH5 and pH7 Park Grass soils. Building on the enzymatic physiological assays, I further demonstrated shifts in the relative abundance of principally *Acidobacterial* and *Actinobacterial* β-glucosidase genes between pH5 and pH7 soils. When the relative abundance of these genes was normalized using a housekeeping gene (DNA gyrase subunit B), *Acidobacteria* β-glucosidases in particular were twice as abundant in pH5 soils compared to pH7 soils.

This work, in addition to confirming that pH is influential over processes of relevance to soil services (carbon cycling) also highlighted the importance of considering other molecular mechanisms in addition to simply gene presence, as *Acidobacterial* contigs containing β-

glucosidases associated genes also showed enhanced presence of signal peptides further implicating their role in C cycling in acidic soils. More generally, this work identified that functional differences are manifest even in a taxonomically ubiquitous functional gene, though the observed differences in the taxonomy of producers could underlie the functional differences. Though I made some attempts to explore specific differences in the amino acid sequence of β-glucosidases of varying phyla, these results were inconclusive. These findings are however important both for future ecological (relevance of genetic traits for functional indication) and perhaps biotechnological exploitation purposes.

6.2.2 Land use effects on taxa and function

In addition to looking at broad microbial responses to pH, my overarching aim was to quantify the direct influences of land use on soil microbial communities, by identifying the consistency in effects of land use intensification in different environmental contexts. This matter is of considerable importance both for fundamental understanding of how human activities can influence microbial communities and their functioning; but also with respect to advancing soil process understanding (particularly nutrient cycling and climate change mitigation) and defining functionally relevant indicators to develop more sustainable management practices (Demenois et al., 2020; Keesstra et al., 2016). I therefore examined taxonomic and functional changes in microbial communities in response to land use change, using amplicon, short read (**chapter 3**) and assembled (**chapter 5**) metagenome data from distributed soil contrasts across Britain (conducted by a consortium of researchers as part of the NERC Soil Security Programme).

The taxonomic composition of microbial communities was primarily effected by soil pH but also varied according to land use, in keeping with findings in the wider literature (Hartmann *et al.*, 2015; Pershina *et al.*, 2015; Francioli *et al.*, 2016). Further, ordinations of both 16S rRNA gene communities and metagenomic functional profiles showed consistent trends, with pH variance in grasslands explaining the majority of the variation along the first axis. Land use intensification effects were apparent on the second axis with bare fallow soils (which had not been cropped for 50 years) appearing to comprise particularly distinct communities (**chapter 3**). In order to specifically link these taxonomic and functional

analyses, I then employed a metagenomic assembly approach to curate taxonomic genomic bins from metagenomes (**chapter 5**). Using metagenomic binning to study environmental responses of microbes in soils is a novel approach and only a limited number of studies have employed binning to study soil microorganisms more generally (Kroeger et al., 2018; Orellana et al., 2018; White et al., 2016).

Through using this approach I was able to assemble and bin 127 near complete genomes (completeness >=80%), though with considerable redundancy due to the methodologies implemented. However despite this I identified numerous *Thaumarchaeota* bins as important discriminators of grassland and arable soils, and as indicative of arable soils specifically. As all the arable sites studied were treated with nitrogen fertilisers, this may be related to *Thaumarchaeotas* known role in ammonium oxidation (Prosser & Nicol, 2012). Indeed increases in *Thaumarchaeota* MAGs in response to nitrogen inputs has been observed in previous work (Orellana et al., 2018). However when examining the functional content of *Thaumarchaeota* bins I did not find functional genes linked to ammonium oxidation specifically, though the *Thaumarcheota* comprised a vastly different functional genetic content compared with the bacterial bins. This finding that the phylogeny of microbial taxa is highly discriminative of functional gene content was also manifest for other bacterial lineages, and indeed phylogeny was shown to be more important than ecological niche (modelled pH response). It is of note that numerous functional genes that were phyla indicators were related to generic cellular functions, the variation in these genes is perhaps not surprising and consequently caution must still be exerted in making general conclusions that change in taxonomy will directly mean change in specific functions of relevance to soil ecosystem services.

To obtain broader insights into land use effects on soil functions of relevance to ecosystem services, I performed indicator analyses both on the short read analyses and metagenomic bins (**chapter 3, 5** respectively). Crucially functional genes indicative of land use included genes associated with biogeochemical cycling and ecosystem functioning. Nitrate reductase subunits were key in distinguishing between high and low intensity soils and were consistently found in higher relative abundance within high intensity soils (**Chapter 3).** Other denitrification genes were indicative of arable soils, including nitrite reductase and nitric

oxide reductase genes (**chapter 5**). These results are likely to be related to the nitrogen based fertilisers used to treat arable soils within the samples studied, given that increased rates of denitrification have commonly been linked to agricultural soils and the application of nitrogen based fertilisers more specifically (De Klein & Van Logtestijn, 1994; Kaiser et al., 1996; Philippot et al., 2007). Conversely within grassland soils I found numerous genes associated with nitrogenases, indicating more N acquisition through direct fixation. Other important biogeochemical cycles also demonstrated shifts in genes in response to land use. Within sulfur metabolism there was more inorganic sulfur assimilation within arable indicators and more sulfur oxidation and sulfate reduction associated complex genes within grassland indicators. Within phosphorus metabolism there were more high affinity phosphate transport and control of PHO regulation and P uptake genes within arable indicators and more alklphosphonate utilisation genes within grassland indicators. Taken together my findings suggest varying nutrient acquisition strategies between grassland and arable soils, with grassland soils acquiring N, P and S from organic sources, whilst within arable soils these nutrients were obtained from inorganic sources (**chapter 5**). The approach of coupling broad functional indicators of land use based on short read annotation, with indicators of phyla based on bin functional gene content, enabled extensive insights into how land use induced changes of specific microbial taxa can influence functional change at the community level. Together this emphasised the power and potential of metagenomic binning to detect specific taxa functionality indicative of land use change and more specifically to study the functionality and ecology of novel previously uncharacterised soil microorganisms.

6.3 Future directions in genomic approaches to soil microbes

6.3.1 The opportunities and challenges of using metagenomics to study a poorly characterised system

The results presented within this work highlight the benefits of using a metagenomics approach to gain a broad picture of land use induced changes in soil microbial function (**chapter's 3, 5**). Whilst more targeted approaches such as qPCR have enabled valuable insights into specific functional gene responses to land use (Clark et al., 2020; Hallin et al.,

2009; Zhao et al., 2017), metagenomics has the potential to provide insights into the responses of numerous functional genes, with potential relevance to a various biogeochemical cycles. Indeed a broad genomic approach can be of great value in the context of soils given the hyper diverse nature of microbial communities and the fact a large majority of taxa remain uncharacterised (Solden et al., 2016) and therefore a great deal about the genomic mechanisms underlying soil processes remains unknown. Further, the wealth of uncharacterised genetic material in soils also means that soil metagenomics can be used to identify novel gene products that could be utilised for pharmaceutical or biotechnological applications. This has been demonstrated in recent work, which utilised genomic data from the soil microbiome to discover a novel class of antibiotics, which had not been reported using culturing approaches (Hover et al., 2018).

Arguably the "black box" nature of soils is both an opportunity and a challenge. On the one hand there is extensive potential for the discovery of novel functional processes and gene products, on the other this lack of characterisation means soil genomic data can be a challenge to annotate and interpret using conventional methods and databases. Indeed genomic databases are typically biased towards highly studied taxa, such as those of medical interest, whilst representation of the soil microbiome within these databases is poor. Recent progress in this area includes the development of RefSoil which is a genomic database curated exclusively from genomes of soil associated organisms and therefore only contains soil relevant functional genes. Although not used within this work, utilizing a more soil specific genomic database in future work should be considered. Further, as more metagenome assembled genomes and single-amplified genomes are assembled from soil (**discussed in greater detail in 6.3.2)** there will be greater opportunity to expand upon a soil specific genomic database as these methods develop (Choi et al., 2017).

Additionally, there are wider issues regarding ecological interpretation of genes detected within an environmental context. Whilst throughout this thesis (**chapter 3 and 5**) I have highlighted functional genes which respond to soil land use and many of these genes can be linked to ecosystem services (particularly those associated with nutrient cycling), reliably linking "gene names" to specific "ecosystem services" can be problematic. To an extent gene ontology databases, which group functional genes into wider processes and structural

complexes such as SEED (Overbeek et al., 2005), KEGG (Ogata et al., 1999), MetaCyc (Caspi et al., 2016), COG (Kristensen et al., 2010) and GO (Gene Ontology Consortium 2000) enable us to obtain a broad overview of the functions likely to be occurring within our samples. However as most ontology frameworks are not curated for environmental or ecological research specifically, even broader functional classifications can be challenging to interpret in the context of soils. Within this thesis I used the SEED ontology system to annotate soils which at the broad level does include some relevant terms such as "Carbohydrate", "Nitrogen metabolism", "Phosphorus metabolism" and "Sulfur metabolism". However other broad SEED classifications include "clustering-based subsystems", "cofactors vitamins and prosphetic groups", "Nucleosides and Nucleotides" and "RNA metabolism", which are comparatively harder to contextualise. In recent years, there have been some efforts to curate environmentally relevant ontology databases such as FOAM which is manually curated based upon environmentally relevant KEGG's linked to environmental processes (Prestat et al., 2014). Broad classifications within the FOAM ontology system include "Carbohydrate Active enzyme (CAZy)", "Methanogenesis", "TCA cycle", "Fermentation", "Hydrocarbon degradation", "Synthesis of saccharides and derivatives" as well as "Nitrogen cycle", "Sulfur metabolism" which are much easier to interpret within an ecological context of "traits" (Prestat et al., 2014). Undoubtedly, in order to map microbes to effect traits using genomic data, it is essential to have access to an environmentally relevant function ontology systems, so that we can clearly interpret and report the likely roles of phylotypes within soils.

6.3.2 New approaches to genome assembly

Knowledge regarding the functionality of diverse soil microbes is likely to further improve as sequencing and bioinformatics methods continue to develop. New assembly methods are likely to lead to an increased number of MAGs submitted to databases (Bowers *et al.*, 2017) as well as an increase in MAG quality. As discussed, assembly approaches have begun to be applied to soils to successfully retrieve MAGs from uncultivated taxa (Kroeger et al., 2018; Orellana et al., 2018; White et al., 2016). Whilst I have described more commonly implemented assembly approaches using short read metagenome data in **Chapter 1 and 5**,

further advancements in both sequencing and bioinformatics technologies are likely to also improve our ability to assemble genomes in the future.

One relatively new approach is assembling data from long read sequencing technologies such an Nanopore (Wang et al., 2014) and PacBio (Pollard et al., 2018). These technologies have enabled the assembly of more contiguous genomic information and to more easily assemble repeat regions, which are challenging to assemble using short read data (Molina-Mora et al., 2020), particularly as repeat regions can be substantially longer than the short reads themselves (van Dijk et al., 2018). Hybrid assembly methods are also becoming more widely implemented (Moss et al., 2020), typically such approaches involve error prone longer reads being 'polished' by short reads from the same samples providing contiguous high quality assemblies. Hybrid assembly methods have been used in soils to assemble a previously uncharacterised *pseudomonas* strain (White et al., 2016) as well as a range of other contexts, including biomedical studies (Molina-Mora et al., 2020; Moss et al., 2020; Wick et al., 2017), and to assemble genomes from a partial-nitritation anammox (PNA) reactor (Liu et al., 2020).

Although less widely used, HI-C technologies also have the potential to improve metagenome assemblies. HI-C determines chromatin interactions and spatial organisation within a cell and can therefore inform us as to how close genomic regions are to one another (Belton et al., 2012) and can be used to inform metagenomic binning. In recent years bioinformatics tools have been developed which cluster assembled contigs derived from metagenomes, into bins based on accompanied Hi-C data (Burton et al., 2014; Demaere & Darling, 2019). Whilst to our knowledge a hybrid approach with metagenome and Hi-C data has not been used to study soil microbial communities, this approach has been used to retrieve >60 draft genomes from cow rumen (Stewart et al., 2018) and to study antibiotic resistance genes in waste water treatment plants (Stalder et al., 2019).

An alternative to assembling from metagenomes is single cell genomics. Single cell genomics, is a highly targeted approach, whereby genomes are studied at the cellular level, enabling the assembly of genomes (SAG's) from unculturable bacteria without dealing with the complexity of microbial communities (Gawad et al., 2016) or the risk of cross-assembly

of varying strains or taxa (Rinke et al., 2014). Isolating a cell is however technically challenging and needs specialist equipment such as microfluidics, micromanipulators or flow cytometry (Bowers et al., 2017). Whilst not widely applied to soils, a recent study looking at forest soil microbial communities employed a related approach, which they termed “mini metagenomics”. Within this approach flow cytometry was used to conduct pooled cell sorting before implementing shotgun sequencing on this smaller pooled community in order to better characterise rare microbial taxa (Alteio et al., 2020).

6.3.3 Dissemination of microbial taxon and functional information via digital technologies

As molecular and bioinformatics approaches continue to develop, another challenge is the collation and dissemination of this rapidly increasing volume of genomic data. There is significant potential to expand upon digital technologies to make microbial environmental datasets and specific taxon traits more widely accessible for easier synthesis of findings across studies. New databases have recently been developed in this area, for example Terrestrial-metagenome DB enables users to easily query terrestrial metagenome datasets stored in Sequence Read Archive (SRA) and MG-RAST, and enables filtering of results by study characteristics including biome, sequencing platform, sample depth, pH, temperature and geographical location etc. Thus making it easier to find terrestrial metagenomic studies of interest (Corrêa et al., 2020) to compare results to, or utilize in meta-analyses. Within my work (**Chapter 2)** I presented ID-TaxER (https://shiny-apps.ceh.ac.uk/ID-TaxER), an application enabling users to query 16S rRNA gene sequences to obtain modelled pH response, habitat preference and spatial mapping information alongside taxonomic assignments. Recently a database using a similar approach to ID-TaxER, Global fungi has been developed, whereby fungal taxon searches (via taxon name or sequence) provides the user with habitat, geographical/mapping and pH information regarding the specific samples the taxon was detected in (Větrovský et al., 2020).

One of the initial aims of this thesis was to link taxonomic environmental responses (such as those reported in ID-TaxER) with functional capacities. Whilst I came some way to achieving

this aim, through assigning taxonomy and pH preferences to β-glucosidase sequences (**Chapter 4**) and more notably by functionally annotating uncharacterised metagenomic bins and modelling their pH responses (**Chapter 5**), there is still a large amount of work to be done in this area. Ideally, I would have made steps to further developed ID-TaxER through linking the environmental responses of taxa **(Chapter 2)** with functional gene content found in bins **(Chapter 5)**. However I was not able to successfully retrieve 16S rRNA genes from bins and thus not able to link functional gene content with ID-TaxER responses (given responses are reported on the OTU level). The challenges of retrieving 16S genes from assemblies has previously been reported and relates to difficulties in assembling highly conserved sequences from short reads (Yuan et al., 2015). As molecular and assembly methods continue to develop, there is therefore the potential to develop similar databases coupling ecological and functional information for soil metagenomic bins from large scale surveys.

6.3.4 Towards prediction of soil function under environmental change

Databases coupling both ecological response models and the functional potential of discrete taxa derived from metagenomic binned genomes, would enable a model based predictive understanding of how resilient various soil microbial functions are to environmental change. Within this thesis I have shown that soil properties (particularly pH) and land use factors can be highly predictive of bacterial taxon relative abundance (**Chapter 2**) and both phylogeny and pH can strongly influence on microbial functional gene composition (**Chapters 3 -5).** However it is impossible to draw general conclusions as to how environmental change will affect broad metrics of functionality, since as I have demonstrated, functions differ in the extent to which they vary across taxa. For example, carbohydrate active enzymes (CAZymes) such as β-glucosidases may be present across many broad lineages, but the specific CAZyme families related to that activity may be more constrained to specific taxa (**Chapter 4**). Similarly for the nitrogen cycle, it is clear that functions such as N fixation appear highly constrained to certain phylogenetic lineages, though it is still difficult to generalise that all members of that lineage will possess the trait (**Chapter 5**). Digital and model informed methods built from extensive catalogues of soil microbial genomes and their ecological characteristics therefore provide a potential means of answering broad questions on how

change can affect various specified functions in soil systems. The models I applied to predict taxonomic abundances only included a single driver, namely soil pH. However numerous other factors drive changes in pH including climate, parent material and notably land use, therefore more advanced ecological modelling approaches could be developed at a range of geographic scales using such global datasets. Coupling models predictive of pH (e.g. Slessarev *et al.*, 2016; Wamelink *et al.*, 2019) with microbial pH response models could therefore predict microbial taxonomic and functional responses to future climate change or alternative land use scenarios.

It must be noted however, that the relative abundance of a functional gene does not necessarily act as a proxy for measuring the associated functional process. Recently Jansson and Hofmockel emphasised the need to address questions regarding how well changes in functional gene content translate to changes in soil function, where they described the expression of the soil metagenome, as the soil "metaphenome" (Jansson & Hofmockel, 2020). Recent work in this area has utilised a combination of 16S, metagenomic and metatranscriptiomic approaches to study physiological responses of the soil microbiome to moisture perturbations and observed changes in microbial taxonomy and function in response to soil drying (Chowdhury et al., 2019). While other work examining 16S, metagenomic and catabolic profiling of microbial communities in response to N addition, found that 16S and metagenomic data significantly correlated with catabolic capacities across a nitrogen gradient (Fierer et al., 2012). Whilst functional information derived from metagenomes provides valuable insights into the functional potential of soils, further assessment is needed to assess how well changes in functional gene content actually predicts changes in soil processes in situ. Of course, functional genes detected in soils may be "Relic DNA" i.e. DNA from dead organisms and thus not be contributing to soil functioning. Equally genes from active organisms may not necessarily be transcribed (Jansson & Hofmockel, 2018; Nannipieri et al., 2020). Additionally, as certain functions depend upon interactions with substrates or other taxa, the spatial heterogeneity of microbes in soils may constrain the interactions and functions expected (Jansson & Hofmockel, 2020). Irrespective of these issues, I believe the further development of models and resources which predict change in soil microbial communities and function, would provide a robust predictive foundation for a new generation of experimentation, testing the

importance of community change for the delivery of soil ecosystem services. Such experiments would also need to encompass more advanced measures of soil processes and fluxes (since many cannot be measured accurately in situ), in order to validate the extent to which change at the genetic level can be used to represent actual change in function.

6.4 Bibliography

Alteio, L. V., Schulz, F., Seshadri, R., Varghese, N., Rodriguez-Reillo, W., Ryan, E., Goudeau, D., Eichorst, S. A., Malmstrom, R. R., Bowers, R. M., Katz, L. A., Blanchard, J. L., & Woyke, T. (2020). Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *MSystems*, *5*(2). https://doi.org/10.1128/msystems.00768-19

Banerjee, S., Kirkby, C. A., Schmutter, D., Bissett, A., Kirkegaard, J. A., & Richardson, A. E. (2016). Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry*, *97*, 188–198. https://doi.org/10.1016/j.soilbio.2016.03.017

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., … Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. In *Nature Biotechnology* (Vol. 35, Issue 8, pp. 725–731). Nature Publishing Group. https://doi.org/10.1038/nbt.3893

Burton, J. N., Liachko, I., Dunham, M. J., & Shendure, J. (2014). Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3: Genes, Genomes, Genetics*, *4*(7), 1339–1346. https://doi.org/10.1534/g3.114.011825

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., & Karp, P. D. (2016). The MetaCyc database of metabolic pathways and enzymes and

the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, *44*(D1), D471–D480. https://doi.org/10.1093/nar/gkv1164

Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Flater, J., Tiedje, J. M., Hofmockel, K. S., Gelder, B., & Howe, A. (2017). Strategies to improve reference databases for soil microbiomes. *ISME Journal*, *11*(4), 829–834. https://doi.org/10.1038/ismej.2016.168

Clark, I. M., Fu, Q., Abadie, M., Dixon, E. R., Blaud, A., & Hirsch, P. R. (2020). Edaphic factors and plants influence denitrification in soils from a long-term arable experiment. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-72679-z

Consortium, T. G. O. (2000). Gene ontology: Tool for the identification of biology. *Natural Genetics*, *25*(may), 25–29.

Corrêa, F. B., Saraiva, J. P., Stadler, P. F., & Da Rocha, U. N. (2020). TerrestrialMetagenomeDB: A public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Research*, *48*(D1), D626–D632. https://doi.org/10.1093/nar/gkz994

De Klein, C. A. M., & Van Logtestijn, R. S. P. (1994). Denitrification in the top soil of managed grasslands in The Netherlands in relation to soil type and fertilizer level. *Plant and Soil*, *163*(1), 33–44. https://doi.org/10.1007/BF00033938

Demaere, M. Z., & Darling, A. E. (2019). Bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biology*, *20*(1), 46. https://doi.org/10.1186/s13059-019-1643-1

Demenois, J., Torquebiau, E., Arnoult, M. H., Eglin, T., Masse, D., Assouma, M. H., Blanfort, V., Chenu, C., Chapuis-Lardy, L., Medoc, J. M., & Sall, S. N. (2020). Barriers and Strategies to Boost Soil Carbon Sequestration in Agriculture. *Frontiers in Sustainable Food Systems*, *4*, 37. https://doi.org/10.3389/fsufs.2020.00037

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Fierer, N., & 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), 626–631. https://doi.org/10.1073/pnas.0507535103

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME Journal*, *6*(5), 1007–1017. https://doi.org/10.1038/ismej.2011.159

Francioli, D., Schulz, E., Lentendu, G., Wubet, T., Buscot, F., & Reitz, T. (2016). Mineral vs. organic amendments: Microbial community structure, activity and abundance of agriculturally relevant microbes are driven by long-term fertilization strategies. *Frontiers in Microbiology*, *7*(SEP). https://doi.org/10.3389/fmicb.2016.01446

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, *17*(3), 175–188. https://doi.org/10.1038/nrg.2015.16

Goulding, K. W. T. (2016). Soil acidification and the importance of liming agricultural soils with particular reference to the United Kingdom. *Soil Use and Management*, *32*(3), 390–399. https://doi.org/10.1111/sum.12270

Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., & Whiteley, A. S. (2011). The bacterial biogeography of British soils. *Environmental Microbiology*, *13*(6), 1642–1654. https://doi.org/10.1111/j.1462-2920.2011.02480.x

Hallin, S., Jones, C. M., Schloter, M., & Philippot, L. (2009). Relationship between n-cycling communities and ecosystem functioning in a 50-year-old fertilization experiment. *ISME Journal*, *3*(5), 597–605. https://doi.org/10.1038/ismej.2008.128

Hartmann, M., Frey, B., Mayer, J., Mäder, P., & Widmer, F. (2015). Distinct soil microbial diversity under long-term organic and conventional farming. *ISME Journal*, *9*(5), 1177–1194. https://doi.org/10.1038/ismej.2014.210

Hover, B. M., Kim, S. H., Katz, M., Charlop-Powers, Z., Owen, J. G., Ternei, M. A., Maniko, J., Estrela, A. B., Molina, H., Park, S., Perlin, D. S., & Brady, S. F. (2018). Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-

resistant Gram-positive pathogens. *Nature Microbiology*. https://doi.org/10.1038/s41564-018-0110-1

Jansson, J. K., & Hofmockel, K. S. (2018). The soil microbiome — from metagenomics to metaphenomics. *Current Opinion in Microbiology*, *43*, 162–168. https://doi.org/10.1016/j.mib.2018.01.013

Jansson, J. K., & Hofmockel, K. S. (2020). Soil microbiomes and climate change. *Nature Reviews Microbiology*, *18*(1), 35–46. https://doi.org/10.1038/s41579-019-0265-7

Jones, D. L., Cooledge, E. C., Hoyle, F. C., Griffiths, R. I., & Murphy, D. V. (2019). pH and exchangeable aluminum are major regulators of microbial energy flow and carbon use efficiency in soil microbial communities. *Soil Biology and Biochemistry*, *138*(August), 0–4. https://doi.org/10.1016/j.soilbio.2019.107584

Kaiser, E., K., K., Kuecke, M., Schnug, E., Munch, J., & O., H. (1996). *Nitrous oxide release from cultivated soils: influence of different N-fertilizer types*.

Keesstra, S. D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J. N., Pachepsky, Y., van der Putten, W. H., Bardgett, R. D., Moolenaar, S., Mol, G., Jansen, B., & Fresco, L. O. (2016). The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. *SOIL*, *2*(2), 111–128. https://doi.org/10.5194/soil-2-111-2016

Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, *26*(12), 1481–1487. https://doi.org/10.1093/bioinformatics/btq229

Kroeger, M. E., Delmont, T. O., Eren, A. M., Meyer, K. M., Guo, J., Khan, K., Rodrigues, J. L. M., Bohannan, B. J. M., Tringe, S. G., Borges, C. D., Tiedje, J. M., Tsai, S. M., & Nüsslein, K. (2018). New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. *Frontiers in Microbiology*, *9*(JUL), 1635. https://doi.org/10.3389/fmicb.2018.01635

Leff, J. W., Jones, S. E., Prober, S. M., Barberán, A., Borer, E. T., Firn, J. L., Harpole, W. S., Hobbie, S. E., Hofmockel, K. S., Knops, J. M. H., McCulley, R. L., La Pierre, K., Risch, A. C.,

Seabloom, E. W., Schütz, M., Steenbock, C., Stevens, C. J., & Fierer, N. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(35), 10967–10972. https://doi.org/10.1073/pnas.1508382112

Liu, L., Wang, Y., Che, Y., Chen, Y., Xia, Y., Luo, R., Cheng, S. H., Zheng, C., & Zhang, T. (2020). High-quality bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly method. *Microbiome*, *8*(1), 155. https://doi.org/10.1186/s40168-020-00937-3

Malik, A. A., Puissant, J., Buckeridge, K. M., Goodall, T., Jehmlich, N., Chowdhury, S., Gweon, H. S., Peyton, J. M., Mason, K. E., van Agtmaal, M., Blaud, A., Clark, I. M., Whitaker, J., Pywell, R. F., Ostle, N., Gleixner, G., & Griffiths, R. I. (2018). Land use driven change in soil pH affects microbial carbon cycling processes. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-05980-1

Malik, A. A., Thomson, B. C., Whiteley, A. S., Bailey, M., & Griffiths, R. I. (2017). Bacterial Physiological Adaptations to Contrasting Edaphic Conditions Identified Using Landscape Scale Metagenomics. *MBio*, *8*(4), e00799-17. https://doi.org/10.1128/mBio.00799-17

Molina-Mora, J. A., Campos-Sánchez, R., Rodríguez, C., Shi, L., & García, F. (2020). High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-58319-6

Moss, E. L., Maghini, D. G., & Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, *38*(6), 701–707. https://doi.org/10.1038/s41587-020-0422-6

NANNIPIERI, P., ASCHER-JENULL, J., CECCHERINI, M. T., PIETRAMELLARA, G., RENELLA, G., & SCHLOTER, M. (2020). Beyond microbial diversity for predicting soil functions: A mini review. *Pedosphere*, *30*(1), 5–17. https://doi.org/10.1016/S1002-0160(19)60824-6

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *27*(1), 29–34. https://doi.org/10.1093/nar/27.1.29

Orellana, L. H., Chee-Sanford, J. C., Sanford, R. A., Löffler, F. E., & Konstantinidis, K. T. (2018). Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. *Applied and Environmental Microbiology*, *84*(2). https://doi.org/10.1128/AEM.01646-17

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., … Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, *33*(17), 5691–5702. https://doi.org/10.1093/nar/gki866

Pershina, E., Valkonen, J., Kurki, P., Ivanova, E., Chirak, E., Korvigo, I., Provorov, N., & Andronov, E. (2015). Comparative analysis of prokaryotic communities associated with organic and conventional farming systems. *PLoS ONE*, *10*(12), e0145072. https://doi.org/10.1371/journal.pone.0145072

Philippot, L., Hallin, S., & Schloter, M. (2007). Ecology of Denitrifying Prokaryotes in Agricultural Soil. In *Advances in Agronomy* (Vol. 96, pp. 249–305). Academic Press. https://doi.org/10.1016/S0065-2113(07)96003-4

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. In *Human molecular genetics* (Vol. 27, Issue R2, pp. R234–R241). NLM (Medline). https://doi.org/10.1093/hmg/ddy177

Prestat, E., David, M. M., Hultman, J., Ta??, N., Lamendella, R., Dvornik, J., Mackelprang, R., Myrold, D. D., Jumpponen, A., Tringe, S. G., Holman, E., Mavromatis, K., & Jansson, J. K. (2014). FOAM (Functional Ontology Assignments for Metagenomes): A Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Research*, *42*(19), 1–7. https://doi.org/10.1093/nar/gku702

Prosser, J. I., & Nicol, G. W. (2012). Archaeal and bacterial ammonia-oxidisers in soil: The quest for niche specialisation and differentiation. In *Trends in Microbiology* (Vol. 20, Issue 11, pp. 523–531). Elsevier Current Trends. https://doi.org/10.1016/j.tim.2012.08.001

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and

web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. https://doi.org/10.1093/nar/gks1219

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., & Woyke, T. (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nature Protocols*. https://doi.org/10.1038/nprot.2014.067

Roy Chowdhury, T., Lee, J.-Y., Bottos, E. M., Brislawn, C. J., White, R. A., Bramer, L. M., Brown, J., Zucker, J. D., Kim, Y.-M., Jumpponen, A., Rice, C. W., Fansler, S. J., Metz, T. O., McCue, L. A., Callister, S. J., Song, H.-S., & Jansson, J. K. (2019). Metaphenomic Responses of a Native Prairie Soil Microbiome to Moisture Perturbations. *MSystems*, *4*(4). https://doi.org/10.1128/msystems.00061-19

Schöps, R., Goldmann, K., Herz, K., Lentendu, G., Schöning, I., Bruelheide, H., Wubet, T., & Buscot, F. (2018). Land-use intensity rather than plant functional identity shapes bacterial and fungal rhizosphere communities. *Frontiers in Microbiology*, *9*(NOV), 2711. https://doi.org/10.3389/fmicb.2018.02711

Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M., & Biss, P. M. (2006). The Park Grass Experiment 1856-2006: Its contribution to ecology. *Journal of Ecology*, *94*(4), 801–814. https://doi.org/10.1111/j.1365-2745.2006.01145.x

Slessarev, E. W., Lin, Y., Bingham, N. L., Johnson, J. E., Dai, Y., Schimel, J. P., & Chadwick, O. A. (2016). Water balance creates a threshold in soil pH at the global scale. *Nature*, *540*(7634), 567–569. https://doi.org/10.1038/nature20139

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. In *Current Opinion in Microbiology* (Vol. 31, pp. 217–226). Elsevier Ltd. https://doi.org/10.1016/j.mib.2016.04.020

Stalder, T., Press, M. O., Sullivan, S., Liachko, I., & Top, E. M. (2019). Linking the resistome and plasmidome to the microbiome. *ISME Journal*. https://doi.org/10.1038/s41396-019-0446-4

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., Liachko, I., Snelling, T. J., Dewhurst, R. J., Walker, A. W., Roehe, R., & Watson, M. (2018). Assembly of 913

microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications*. https://doi.org/10.1038/s41467-018-03317-6

Tian, D., & Niu, S. (2015). A global analysis of soil acidification caused by nitrogen addition. *Environmental Research Letters*, *10*(2). https://doi.org/10.1088/1748-9326/10/2/024019

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. In *Trends in Genetics* (Vol. 34, Issue 9, pp. 666–681). https://doi.org/10.1016/j.tig.2018.05.008

Větrovský, T., Morais, D., Kohout, P., Lepinay, C., Algora, C., Awokunle Hollá, S., Bahnmann, B. D., Bílohnědá, K., Brabcová, V., D'Alò, F., Human, Z. R., Jomura, M., Kolařík, M., Kvasničková, J., Lladó, S., López-Mondéjar, R., Martinović, T., Mašínová, T., Meszárošová, L., … Baldrian, P. (2020). GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. *Scientific Data*, *7*(1), 1–14. https://doi.org/10.1038/s41597-020-0567-7

Wamelink, G. W. W., Walvoort, D. J. J., Sanders, M. E., Meeuwsen, H. A. M., Wegman, R. M. A., Pouwels, R., & Knotters, M. (2019). Prediction of soil pH patterns in nature areas on a national scale. *Applied Vegetation Science*, *22*(2), 189–199. https://doi.org/10.1111/avsc.12423

Wang, Y., Yang, Q., & Wang, Z. (2014). The evolution of nanopore sequencing. *Frontiers in Genetics*, *5*(DEC), 449. https://doi.org/10.3389/fgene.2014.00449

White, R. A., Bottos, E. M., Roy Chowdhury, T., Zucker, J. D., Brislawn, C. J., Nicora, C. D., Fansler, S. J., Glaesemann, K. R., Glass, K., & Jansson, J. K. (2016). Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. *MSystems*, *1*(3). https://doi.org/10.1128/msystems.00045-16

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6). https://doi.org/10.1371/journal.pcbi.1005595

Yuan, C., Lei, J., Cole, J., & Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*, *31*(12), i35–i43. https://doi.org/10.1093/bioinformatics/btv231

Zhao, C., Gupta, V. V. S. R., Degryse, F., & 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), 159–169. https://doi.org/10.1007/s00374-016-1162-0

Appendix 1

Co-authored paper published in Soil Biology and Biochemistry.

Contributions: Bioinformatics and statistical analyses of metagenome data**.**

The pH optimum of soil exoenzymes adapt to long term changes in soil pH

Jérémy Puissant[a], Briony Jones[b,c], Tim Goodall[a], Dana Mang[a], Aimeric Blaud[e1], Hyun Soon Gweon[a,f], Ashish Malik[a], Davey L. Jones[c,d], Ian M. Clark[e], Penny R Hirsch[e], Robert Griffiths[b]

[a] Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire OX10 8BB, United Kingdom

[b] Centre for Ecology & Hydrology, Environment Centre Wales, Deiniol Road, Bangor, Gwynedd, LL57 2UW, United Kingdom

[c] School of Natural Sciences, Bangor University, Deiniol Road, Bangor, Gwynedd, LL57 2UW, United Kingdom

[d] UWA School of Agriculture and Environment, The University of Western Australia, Crawley, WA 6009, Australia

[e] Dept. Sustainable Agriculture Sciences, Rothamsted Research, Harpenden, AL5 2JQ, United Kingdom

[f] School of Biological Sciences, University of Reading, RG6 6AS, United Kingdom

Corresponding author: Tel.: +44 1491692547; E-mail address: jeremy.puissant@gmail.com

[1] Current address: School of Applied Sciences, Edinburgh Napier University, Sighthill campus, Edinburgh, EH11 4BN, United Kingdom.

Abstract

Soil exoenzymes released by microorganisms break down organic matter and are crucial in regulating C, N and P cycling. Soil pH is known to influence enzyme activity, and is also a strong driver of microbial community composition; but little is known about how alterations in soil pH affect enzymatic activity and how this is mediated by microbial communities. To assess long term enzymatic adaptation to soil pH, we conducted enzyme assays at buffered pH levels on two historically managed soils maintained at either pH 5 or 7 from the Rothamsted Park Grass Long-term experiment. The pH optima for a range of exoenzymes involved in C, N, P cycling, differed between the two soils, the direction of the shift being toward the source soil pH, indicating the production of pH adapted isoenzymes by the soil microbial community. Soil bacterial and fungal communities determined by amplicon sequencing were clearly distinct between pH 5 and soil pH 7 soils, possibly explaining differences in enzymatic responses. Furthermore, β-glucosidase gene sequences extracted from metagenomes revealed an increased abundance of Acidobacterial producers in the pH 5 soils, and Actinobacteria in pH 7 soils. Our findings demonstrate that the pH optimum of soil exoenzymes adapt to long term changes in soil pH, the direction being dependent on the soil pH shift; and we provide further evidence that changes in functional microbial communities may underpin this phenomena, though new research is now needed to directly link change in enzyme activity optima with microbial communities. More generally, our new findings have large implications for modelling the efficiency of different microbial enzymatic processes under changing environmental conditions.

Keywords: enzyme activity, adaptation, liming, carbon degradation, metagenomics, microbial community

1. Introduction

Soil microbes produce exoenzymes to degrade complex plant and soil organic matter (OM) into smaller compounds, which are then assimilated for growth and metabolism (Allison, 2005). These proteins break down large OM compounds through hydrolytic and oxidative processes (Burns et al., 2013; German et al., 2011; Sinsabaugh, 2010) and their activity rates have been hypothesized to be a rate-limiting step in OM decomposition (Bengtson and

Bengtsson, 2007). Enzyme activity is predominantly controlled by temperature and pH which affect enzyme kinetics through change in substrate binding and stability. In contrast to intracellular enzymes, the physico-chemical conditions in which exoenzymes operate are poorly controlled by microorganisms and activity rates are thus influenced by local conditions (e.g. pH). Thus, to cope with their local environment, microorganisms evolve to produce different types of enzyme (isoenzyme), resulting in equivalent functionality but with altered thermodynamic and kinetic properties.

In soil systems, much research has focused on enzyme adaptation to temperature (Allison et al., 2018; Alvarez et al., 2018; Blagodatskaya et al., 2016; Razavi et al., 2017) due to concerns on the effects of future climate change on ecosystem processes. The molecular mechanisms underpinning these adaptations have been studied and are believed to be driven by conformational flexibility within the enzyme active site or protein surface, which affects efficiency in relation to enzyme activation energy (Åqvist et al., 2017; Lonhienne et al., 2000). However, these adaptations also result in various trade-offs between efficiency and enzyme stability (Åqvist et al., 2017; Zanphorlin et al., 2016); meaning both specific exoenzyme-catalyzed processes as well as other non-specific microbial processes may be affected by a changing environment. The assessment of soil enzymatic responses to change in temperature is an active area of research, with some studies suggesting that acclimation can be rapid and driven by changes in underlying microbial communities (Bradford, 2013; Nottingham et al., 2019; Wei et al., 2014). Surprisingly there has been limited reporting of enzymatic adaption to other edaphic properties.

Soil pH is one of the main variables affected by global change through agricultural intensification, climate change and other polluting events such as acid rain (Goulding, 2016; Kirk et al., 2010; Slessarev et al., 2016; Tian and Niu, 2015; van Breemen et al., 1983; Wu et al., 2017). It is well established from laboratory assays that the rate of enzymatic catalytic reactions is dependent on the pH at which the reactions occur, with the point of maximal activity known as the pH optimum (Frankenberger & Johanson,1982, German et al., 2011). Previous studies have demonstrated different pH optima for the same enzyme across widely differing soil types (Niemi and Vepsäläinen, 2005; Turner, 2010), though the causal role of soil pH in predicting pH optimum has never been established. Additionally, pH is known to be one of the main factors affecting soil microbial diversity (Fierer et al., 2017; Griffiths et al., 2011),

yet the relevance of reported changes in communities across pH gradients for soil enzymatic processes remains unknown. With enzymatic kinetics now being incorporated into recent C decomposition models (Allison, 2012; Davidson et al., 2012; Wang et al., 2013), we believe empirical data on the specific role of pH in affecting enzyme kinetic parameters is now required, since soil pH changes can occur rapidly with unknown acclimation responses. Furthermore, new understanding of the role of microbes in driving responses is essential to both increase understanding of acclimation mechanisms, but also potentially provide easily measurable indicators for model parameterization.

We therefore sought to test soil exoenzymatic adaptation to local pH, by conducting enzymatic assays at a range of buffered pH levels on soils from the Park Grass long-term experiment (Rothamsted) in which the same soil type had been maintained at either pH 5 or 7 for over 100 years. Hydrolytic exoenzymes corresponding to major enzymes involved in organic C, N and P cycling were selected to study. We hypothesize that enzyme pH optimum will be affected by ancestral soil pH treatment, with soil exoenzymes from soil pH 5 being more adapted towards acidic conditions and exoenzymes from soil pH 7 adapted towards more alkaline conditions. To better understand the microbial community relationships underpinning exoenzyme activity and pH adaptation, we also sought to assess the change in microbial community composition (bacteria and fungi) with amplicon sequencing, and functional genes using a metagenomics sequencing approach. Specifically, we wished to determine whether change in enzyme activity is associated with change in specific microbial enzyme producers or adaptation of exoenzymes to environmental conditions.

2. Materials and methods

2.1 Soil sampling

We took advantage of the unique Park Grass Long-term experiment (Rothamsted, UK; Macdonald et al., 2018) in which soils have been maintained at either pH 5 or 7. The experiment originally started in 1856 on permanent pasture to investigate ways of improving hay yields, is managed with a range of fertilisers and pHs with the hay cut twice a year. Soils cores (0-15 cm depth, 4 cm Ø) were sampled on the 27th November 2015 in subplots ‘a’ (pH ~ 7) and ‘c’ (pH ~ 5) of the Nil plot 12, which has never received any fertilisers (Storkey et al., 2016). The soil pH is regularly monitored and controlled by liming, in subplot ‘a’ to reach pH~7

since 1903 (every 4 yr and then every 3 yr from 1976), in subplot 'c' to reach pH~5 since 1965 (every 3 yr). However, because the natural soil pH was 5.4-5.6, pH 5.5 plots have only received minimal liming across the experimental duration to combat natural acidification processes.

2.2 Basic characterization of bulk soil samples

Gravimetric soil moisture content was determined by drying 15 g of soil at 105 °C for 48 h. All other chemical analyses were performed using sieved (2 mm), air-dried (40 °C) soil. Soil pH was measured in H_2O (1:5 weight: vol) according to the protocol NF ISO 10390 (2005). Soil organic carbon C, total N and total P were measured according to CS Technical report No. 3/07 (Emmett et al., 2008). The fingerprint of soil mineralogy was assessed using mid-infrared (MIR) spectroscopy. Dried soil samples were ball-milled and further dried overnight at 40 °C to limit interferences with water, without altering OM chemistry. Milled samples were analyzed using a Nicolet iS10 FT-IR spectrometer (Thermo Fisher Scientific Inc., Madison, WI, USA). Spectral acquisition was performed by diamond attenuated total reflectance (MIR-ATR) spectroscopy over the spectral range 4,000–650 cm^{-1}, with spectral resolution of 8 cm^{-1} and 16 scans per replicate.

2.3 Enzyme assays

Hydrolytic soil exoenzyme activities of phosphatase (PHO, EC number: 3.1.3.1, substrate: 4-MUB-phosphate), β-glucosidase (GLU, EC number: 3.2.1.21, substrate: 4-MUB-β-D-glucopyranoside), acetyl esterase (ACE, EC number: 3.1.1.6, substrate: 4-MUB-acetate) and leucine-aminopeptidase (LEU, EC number: 3.4.11.1, substrate: L-Leucine-7-AMC) were measured by fluorogenic methods using methylumbelliferyl (MUB) and 7-amino-4-methylcoumarin (AMC). PHO, GLU, ACE and LEU are involved in phosphorus mineralization, release of glucose from cellulose, deacetylation of plant compound and degradation of protein into amino acids, respectively. Enzyme assays were performed according to Turner (2010) and following German et al. (2011) recommendations for measuring enzyme activity in soil solution. A range of buffered pH solutions (from 2.5 to 10, in increments of 0.5) was prepared by adjusting 50 mL of modified universal buffer with 1.0 M HCl and 1.0 M NaOH, at 20°C, then diluting to 100 mL with deionized water. The corresponding composition for one litter of modified universal buffer was: 12.6g of boric acid, 28g of citric acid, 23.2 g of maleic acid, 24.2 of Trizma base and 39g of NaOH. Note that the buffered pH solution was diluted 4-

fold in the final assay solution giving a concentration of each chemical of 25mM. Turner (2010) showed that such a concentration was necessary to maintain the required pH during the assay. For each sample, a soil slurry was prepared by adding 20 mL deionized water to 0.5 g of soil (fresh weight), then rotary shaking on a magnetic plate for 20 min at 28 °C. 10 mL of this soil solution was diluted to 25 mL with deionized water to give a 1:100 (w/v) soil-to-water ratio. Enzyme reactions were measured in 96-well microplates containing 50 µL of the specific buffer (25mM), 50 µL of soil slurry (1:400 (w/v) soil-to-water ratio) and 100 µL of substrate solution (saturated concentration, 200 µM). Microplates were then incubated in the dark for 3 h at 28 °C, with one fluorometric measurement every 30 min (BioSpa 8 Automated Incubator) to follow the kinetics of the reaction. Soil pH values were checked before and after incubation and a small drop of 0.1 to 0.2 pH unit was observed after incubation (3h) which we consider being negligible compared to the entire pH range evaluated (2.5 to 10).

For each sample, three methodological replicates (sample + buffer + substrate) and a quenched standard (sample + buffer + 4-MUB or 7-AMC) were used. Quenching curves were prepared with a serial dilution of 4-MUB solution for different amounts of fluorophore in the well (3000, 2000, 1000 pmol) (Puissant et al., 2015). For each substrate, a control including the 4-MUB- or 7-AMC-linked substrate and the buffer solution alone were used to check the evolution of fluorescence without enzyme degradation over the duration of assay. The fluorescence intensity was measured using a Cytation 5 spectrophotometer (Biotek) linked to the automated incubator (Biospa 8, Biotek) and set to 330 and 342 nm for excitation and 450 and 440 nm for emission for the 4-MUB and the 7-AMC substrate, respectively. All enzyme activities were calculated in nmol of product per minute per g of dry soil and expressed as a percentage of the total activity measured across the entire pH range (from pH 2.5 to pH 10).

2.4 Soil microbial community composition

For sequencing analyses of bacterial and fungal communities, DNA was extracted from 5 replicate soil samples per treatment using 0.25 g of soil and the PowerSoil-htp 96 Well DNA Isolation kit (Qiagen) according to manufacturer's protocols. The dual indexing protocol of Kozich et al. (2013), was used for Illumina MiSeq sequencing of the V3-V4 hypervariable regions of the bacterial 16S rRNA gene using primers 341F (Muyzer et al., 1993) and 806R (Youngseob et al., 2005); and the ITS2 region for fungi using primer ITS7f and ITS4r, (Ihrmark

et al., 2012). Amplicon concentrations were normalized using SequalPrep Normalization Plate Kit (Thermo Fisher Scientific) prior to sequencing on the Illumina MiSeq using V3 chemistry. Fungal ITS sequences were analysed using PIPITS (Gweon et al., 2015) with default parameters as outlined in the citation. A similar approach was used for analyses of bacterial sequences, using PEAR (sco.h-its.org/exelixis/web/software/pear) for merging forward and reverse reads, quality filtering using FASTX tools (hannonlab.cshl.edu), chimera removal with VSEARCH_UCHIME_REF and clustering to 97% OTUs with VSEARCH_CLUSTER (github.com/torognes/vsearch). The Illumina MiSeq sequencing generated in average per sample 28205 reads for 16S rRNA gene and 40406 for ITS2 region.

2.5 Metagenome Sequencing

DNA was extracted from 2 g of soil from 4 field replicates for the two pH treatments using the PowerMax Soil DNA Isolation kit (Qiagen), and subsequently concentrated and purified using Amicon® ultra filters. Illumina libraries were constructed using the Illumina TruSeq library preparation kit (insert size < 500- 600 bp) and paired-end sequencing (2 x 150 bp) was conducted using the Illumna HiSeq 4000 platform. Prior to annotation, Illumina adapters were removed from raw fastq files using Cutadapt 1.2.1 (Martin, 2011), reads were trimmed using Sickle (Joshi and Fass, 2011) with a minimum window quality score of 20 and short reads were removed (<20 bp). Preliminary analysis was conducted using MGRAST to functionally annotate with SEED subsystems and taxonomically annotate with refseq. We focused our analyses on bacterial β-glucosidases, since the bacteria dominate soil metagenomics gene libraries (Malik et al., 2017) and the β-glucosidases are genetically well characterized enzymes, known to be important for soil C transformations. For more detailed analyses of β-glucosidase sequences, all reads from the 8 samples were co-assembled using MEGAHIT (Li et al., 2015) with a minimum contig length of 1000. Sequences were translated and open reading frames were predicted using FragGeneScan (Rho et al., 2010). Contigs were assigned CAZY (Carbohydrate-Active enZYmes) subfamilies (Lombard et al., 2014) using a HMMER search (Finn et al., 2011) against dbCan2 profiles with an e-value of 1e-15 (Zhang et al., 2018). Contigs were taxonomically annotated against the NCBI Blast non-redundant protein database using Kaiju, a fast translated method, which identifies protein-level maximum exact matches (MEM's) (Menzel et al., 2016). Regions of contigs annotated as relevant β-glucosidase CAZY domains (GH1, GH2, GH3, GH5, GH9, GH30, GH39, GH116) were extracted.

To identify pH associations of these sequences, DNA reads from individual samples were mapped back to assembled contigs using BlastX, and mappings with an identity percentage of < 97% and/or an e-value of > 0.001 were discarded. Mapping outputs were used to tabulate the abundance of individual reads from the pH 5 and pH 7 samples forming each contig, and then the multinomial species classification method (CLAM) (Chazdon et al., 2011) was used to classify contigs with respect to soil pH designation: generalist- the contig is made up of sequences from both pH 5 and 7 soils; pH specialist- reads making up a contig are predominantly from either pH5 or pH7 soil; or "too rare" whereby the number of reads is too low to reliably classify.

2.6 Statistical analysis

The effects of assay pH, soil field pH treatment and their interactions on enzyme kinetics were assessed by repeated-measures ANOVA. Fixed factors were sampling "assay pH" and "soil pH", while soil field replicate was added as a random factor. One-way ANOVA was used to test the effects of enzymatic pH reaction on soil enzyme relative activity at each pH step (from 2.5 to 10). Differences in relative abundances of microbial taxa between soil pH 5 and soil pH 7 were assessed with one-way ANOVA. Assumptions of normality and homoscedasticity of the residuals were verified visually using diagnostic plots and a Shapiro-Wilk test. To identify soil bacterial and fungal community composition patterns, a principal component analysis (PCA) based on Hellinger-transformed OTU data was performed (Legendre and Gallagher, 2001). Permutational multivariate ANOVA (PERMANOVA) was used to test the effect of soil pH field treatment on soil microbial community composition. All statistical analyses were performed under the R environment software R 3.6.0 (R Development Core Team, 2011), using the R packages vegan (Oksanen et al., 2013), ade4 (Dray and Dufour, 2007) and NLME (Pinheiro et al.,2014). Fourier-transform infrared spectroscopy (FTIR) spectral data were further processed and analyzed using the hyperSpec package (Beleites and Sergo, 2011).

3. Results

3.1. Soil characteristics

The pH values of the two soils were confirmed to be consistent with the treatments applied, with pH measured at 5.5 and 7.5 for the pH 5 and pH 7 plots, respectively. Liming soil from pH 5 to pH 7 significantly increased by ~20% the total C and N contents (Table 1). Soil moisture, total P and C: N were not significantly different between soil pH 5 and soil pH 7 (Table 1). Soil infrared mid-infrared spectroscopy was used to fingerprint soil mineralogy and to assess heterogeneity within and between the two soil pH field treatments. The fingerprints confirm that soil mineralogy is consistent within and between pH field treatments (Supplementary materials, Fig.1). The most prominent feature of the FTIR spectra corresponded to peaks indicative of phyllosilicate mineral compound absorption (kaolinite) with peaks at 3696, 3621, 1003, 912, 692 cm^{-1} (Dontsova et al., 2004). The 774 cm^{-1} peak is likely to be an indicator of quartz content and the 1642 cm^{-1} peak corresponds to the H–O–H bending band of water (Stuart, 2004, Dontsova et al., 2004). Small differences in peak amplitude between pH 5 and pH 7 soils are the result of small changes in the relative concentrations of compounds but overall the two soils presented very similar mineralogy profiles (according to the peak wavelength positions) which indicates a shared ancestral origin.

3.2. Soil microbial community composition

The composition of soil bacterial and fungal community determined by amplicon sequencing (16S rRNA genes and ITS region, respectively) were clearly distinct between soil pH 5 and pH 7 for both communities (Fig. 1; PERMANOVA: $R^2 = 0.82$, $p<0.001$ for fungal community and, $R^2 = 0.51$, p-value: <0.01 for bacterial community). As observed on the PCA (Fig. 1) and PERMANOVA results, fungal community structure was more affected than the bacterial community by the liming treatment. Stacked bar plots representing the relative proportions of microbial phyla demonstrated relatively greater changes in the fungal compared to the bacterial community from pH 5 to pH 7 (Fig. 2). Basidiomycota was significantly more abundant at soil pH 5 (83%, $p<0.001$, Fig. 2) whereas their relative abundance decreased at soil pH 7 (36%) to the advantage of Ascomycota and Zygomycota taxa (30% and 24% at soil

pH 7 compared to 4.5% and 4% at soil pH 5, $p<0.01$, respectively, Fig. 2). Concerning the bacterial community, higher relative abundances of the phyla Acidobacteria and Verrucomicrobia were observed at pH 5 versus pH 7 (22% vs 16%, $p=0.02$; 26% vs 18%, $p<0.01$, respectively Fig. 2). In contrast, a higher relative abundance of Proteobacteria and Actinobacteria phylum was observed at pH 7 versus pH 5 (33% vs 27%, $p=0.01$; 11% vs 7%, p:<0.01, respectively Fig. 2).

3.3. Extracellular enzyme pH optimum assays

The pH of the enzymatic reaction had a highly significant impact on the catalytic efficiency of all enzymes examined (Fig. 3, Table 2). At extremely low pH (2.5), activity was low or could not be detected for leucine aminopeptidase and acetate esterase. For each enzyme, changes in the assay pH strongly impacted the relative enzyme activity with a 15-fold increase between lowest and highest activity at the pH optimum (Fig. 3). After reaching the optima, the activity decreased more or less rapidly depending on the assay. Regardless of the initial pH of the soil, pH optima appeared to be specific to the enzyme studied (Fig. 3). The pH optimum of leucine aminopeptidase and acetyl esterase enzymes were close to neutrality, with an average pH optimum at 7.2 and 6.7, respectively (Fig. 3). The pH optima for β-glucosidase enzyme was acidic with an average of pH 4.3 (Fig. 3). Two pH optima were observed for phosphomonoesterase, one acidic (pH 5.7) and the other alkaline (pH 10), although the alkaline optima may not have been fully reached.

Maintaining field soil at either pH 5 or pH 7 for over 100 years had a strong significant impact on the pH optimum of all enzymes (Table 2). Enzyme pH preference and optima shifted between acidic and alkaline soil whatever the enzyme considered, though this was more pronounced for phosphatase, β-glucosidase and acetate esterase compared to leucine-aminopeptidase (mixed model, Table 2). For each enzyme, the optimum pH differed between the two soils by 0.5 pH units (Fig.3). The interaction between enzymatic assay pH and field soil pH was significant for each enzyme assayed, indicating that the magnitude of the difference in enzyme activity between pH 5 and pH 7 soil is dependent upon assay reaction pH (Table 2). A second optimum at pH 10 was observed for phosphatase and acetyl esterase from pH 7 soil, in contrast to little or no activity of these enzymes from pH 5 soil (Fig. 3A, 3D). Similarly, the relative activity of enzymes from pH 5 soil was always higher to enzymes from

pH 7 in acidic assay conditions (< pH 5.5), while the relative activity of enzymes from pH 7 soil was always higher than enzymes from pH 5 soil in more alkaline conditions (> pH 7).

3.4. Soil metagenomics

The amplicon sequencing results revealed large differences in broad taxa between the two soils of different pH. To determine whether similar shifts were also observed in associated enzymatic gene sequences, shotgun metagenomes datasets generated from the same soils were utilized. Analyses of the functional and taxonomic annotations of β-glucosidase related genes using subsystems annotation revealed greater abundance of sequences from Acidobacteria in the pH 5 compared to pH 7 soil (15.9% vs 1.9%, p-value: 7.4×10^{-5}; Fig.4); and conversely more Actinobacterial β-glucosidase genes in pH 7 soils (34.6% vs 43.4%, p-value: 6×10^{-3}; Fig.4). We further tested differences in abundance by normalizing to a housekeeping gene (gyrB), and found significant differences only for Acidobacterial β-glucosidase genes, which were significantly enriched at pH 5 soil compared with the pH 7 soil, being on average twice as abundant (Supplementary materials, Fig.2).

It is, therefore, apparent at the level of broad phyla, large increases of Acidobacterial β-glucosidases in acid soils are associated with the shift in exoenzyme pH optimum . However, this does not rule out that other phyla may have distinct pH responsive sub clades. To assess this, we assembled pooled metagenomic sequence reads and extracted contigs containing β-glucosidases following functional classification using CAZY and taxonomic annotation to RefSeq. β-glucosidase contigs were then classified as pH specialist (pH 5 or 7) or generalist using a multinomial classification method (CLAM) typically used to classify species' habitat preference based on surveyed counts, but here used on the number of reads per individual sample from the two treatments mapping to each β-glucosidase contig. The majority of Acidobacteria sequences were classed as pH 5 specialists, suggesting that not only is there a higher relative abundance of Acidobacteria β-glucosidase sequences at pH 5 but that the majority of these sequences appear to be unique to pH 5 soils (Fig. 5). Sequences annotated as other dominant phyla such as Actinobacteria and Proteobacteria appeared to have a higher proportion of pH 7 specialist and generalist sequences (supplementary materials, Table 2), whilst Verrucomicrobia possessed a distinct sub-clade of pH 7 specialist sequences (Fig. 5).

4. Discussion

4.1 Soil exoenzyme pH optima are adapted toward local pH

The activity of enzymes involved in C, N and P cycles were all found to be strongly dependent on the pH of the assay. Beta-glucosidase had an acidic pH optimum (pH=4.3), which is generally observed for glycosidase enzymes (Niemi and Vepsäläinen., 2005; Sinsabaugh et al., 2008; Turner., 2010), whereas leucine aminopeptidase had a neutral pH optimum (7.2) as is commonly reported for proteases (Niemi and Vepsäläinen., 2005; Sinsabaugh et al., 2008). Acetyl esterase pH optima were at pH 7 for both soils studied, also in line with previous findings (Degrassi et al., 1999; Humberstone and Briggs, 2000). However, source soil pH had a significant and strong impact on soil exoenzyme pH optimum response curves. For each enzyme studied, extracellular enzymes originally from pH 5 soil were more adapted towards acidic pH conditions, whereas pH 7 soil possessed enzymes adapted towards more alkaline conditions (Fig. 3). Interestingly, the enzymatic pH optima observed in this study did not correspond exactly to the local soil pH, presumably due to constraints within the active sites that enable physicochemical function to be maintained. It is possible that the responses observed are due to the presence of isoenzymes, which have different kinetic properties adapted toward the local soil pH. Alkaline and acid phosphatases are the most studied example of soil isoenzymes (Nannipieri et al., 2011), and our phosphatase pH response curves illustrate this with a marked bimodal distribution, and extremely low activity for the pH 7 soil compared to the pH 5 soil, at acidic assay pH. Acetyl esterase also exhibited a bimodal response but only in the pH 7 soil, which also exhibited a second pH optimum developing at pH 10.

Previous studies have observed different pH optima for the same enzyme across different soil types (Niemi and Vepsäläinen, 2005; Turner, 2010), though the underlying causes responsible for this were not identified. Mechanisms proposed include either abiotic stabilization by soil chemical properties which alter the conformation of the enzyme and thus kinetics; or differences in the microbes that produce the enzymes. Our experiment, conducted on the same soil type, provides strong evidence for microbial control, mediated through altered soil pH. Shifts in enzyme pH optima due to enzyme sorption to different clay types (Leprince and Quiquampoix, 1996; Ramirez-Martinez and McLaren, 1966; Skujins et al., 1974) was

discounted as IR based soil chemistry fingerprints (incorporating information on clay content) were very similar between the pH 5 and pH 7 soils (Supplementary materials, Fig.1). Moreover, the dilution factor used to perform enzyme assays (1:400 soil-to-water ratio) helped to reduce potential effect of small increases in soil total C content and total N observed between the pH 5 and pH 7 soils. Further strong evidence for biotic mechanisms is provided by the consistent non-random shift in optima towards the source soil pH and the presence of bi-modal pH optimum curve indicating clearly the presence of isoenzymes.

4.2 Potential microbial mechanisms governing exoenzyme local adaptation to pH

Bacterial and fungal communities were found to be clearly distinct between the two pH soils investigated, as anticipated from previous work in the Park Grass long-term experiment (Zhalnian et al., 2015; Liang et al., 2015). Such differences in microbial community composition may be responsible for the production of different versions of the same enzyme (Fig. 3). For example, the Acidobacteria phylum has been reported to possess more diverse and abundant genes encoding for carbohydrate-decomposing enzymes than Proteobacteria (Lladó et al., 2019; Lladó et al., 2016). To explore this further, we performed metagenomic sequencing to examine whether the change in enzyme pH preference in the two soils was associated with differences in functional diversity. Focusing specifically on the β-glucosidase exoenzyme, our results clearly showed that different proportions of bacterial phyla produced β-glucosidases across the two soils. Notably, the Acidobacteria contributed more to the β-glucosidase gene pool in the acid soil, and this contribution was more marked than would be expected from examining abundances based on housekeeping genes alone. Furthermore, sub clades of acidobacterial glucosidase were unique in being exclusively found in acid soils, with other broad taxa possessing both generalist enzymes, and a mix of pH specialized genes for either acid or neutral pH. This indicates that acidophilic acidobacterial lineages may possess enzymatic adaptations which underpin their demonstrated competitiveness in acidic soils (Griffiths et al., 2011), and confirms recent genomic studies which have identified enzyme production for carbohydrate degradation as a key feature of these organisms (Eichorst et al., 2018).

Our results highlight the utility in linking metagenomics approaches to measures of specific enzymatic functional traits (pH optimum), with the demonstration of both biodiversity and functional differentiation caused by manipulated soil pH change. In addition the use of molecular approaches here adds to the emerging molecular understanding of the biodiversity of soil enzymes (Berlemont et al., 2013; Heath et al., 2009; Lidbury et al., 2017), and provides new information on the functional capacity of previously undiscovered soil microbial biodiversity. However, we cannot empirically prove that differentially abundant enzyme producers are directly responsible for altered efficiency, since it is currently not possible to assess the diversity of enzymes functionally active within the laboratory-based assays, or indeed the soil. New advanced research is required to determine the relevance of alterations in enzyme producing organisms for soil processes. With respect to pH effects, further insight could be achieved through new computational approaches predicting the pH optima based on amino acid sequence composition (Yan and Wu, 2012; Lin et al., 2013), or in vitro enzyme testing of novel cultured isolates or expressed metagenomic sequences. We also cannot discount evolutionary processes acting within non pH responsive taxa contribute to altered soil pH optima, e.g. through discrete mutations affecting enzyme active sites (Ohara et al., 2014). Whilst a number of evolutionary adaptations to pH have been documented for bacterial strains (Harden et al., 2015) there is little information in the literature on specific exoenzyme adaptations; and whether these result in wider trade-offs with respect to resource acquisition also remains an open question. Addressing these important questions will bring new understanding of the microbial ecological mechanisms governing soil biochemical function under conditions of environmental change; and advances could allow better model parameterization. Specifically, we highlight that incorporation of enzymatic temperature acclimation into models has widely been discussed despite many mechanistic uncertainties (Bradford, 2013; Nottingham et al., 2019; Allison et al., 2018). Our results revealing strong pH adaptation of both enzymatic optimum activity and producer diversity therefore offers an important area for further study within a modelling context, since microbial pH responses are largely predictable (Fierer et al., 2017; Griffiths et al., 2011), and soil pH is highly sensitive to land use and climatic change.

Conclusion

We have specifically demonstrated that the pH optimum of soil exoenzymes adapt towards source soil pH, using soils from a long-term pH manipulation experiment. This was found for all enzymes tested with implications for understanding the resilience of biochemical transformations of carbon, nitrogen and phosphorus across soil systems. Amplicon sequencing and metagenomic data also demonstrated concurrent shifts in taxonomic and functional communities with pH governed shifts in pH optima, providing further evidence that changes in functional microbial communities may underpin pH related change in enzyme kinetic efficiency. These findings call for more research into the underlying genetic controls of enzymatic efficiency in relation to pH, as well as deeper ecological understanding of adaptation mechanisms. More generally, our findings have implications for modelling the efficiency of different microbial enzymatic processes under changing environmental conditions; and soil pH change should be considered, alongside previously documented temperature acclimation, in new carbon models incorporating enzymatic responses to climate change.

Acknowledgements

This work has been funded by the UK Natural Environment Research Council under the Soil Security Programme grant “U-GRASS” (NE/M017125/1) as well as the UK Biotechnology and Biological Sciences Research Council S2N - Soil to Nutrition BBS/E/C/000I0310 programme and the National Capabilities programme grant for Rothamsted Long-term Experiments BBS/E/C/000J0300, the Lawes Agricultural Trust. Two anonymous reviewers are thanked for their constructive comments which improved this paper.

References

Allison, S.D., 2012. A trait-based approach for modelling microbial litter decomposition. Ecology Letters 15, 1058–1070.

Allison, S.D., 2005. Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. Ecology Letters 8, 626–635.

Allison, S.D., Romero-Olivares, AL., Lu, Y., Taylor, JW., Treseder, KK., 2018a. Temperature sensitivities of extracellular enzyme Vmax and Km across thermal environments. Global Change Biology. 24, 2884–2897.

Allison, S. D., Romero-Olivares, AL., Lu, L., Taylor, JW., Treseder, K.K., 2018b. Temperature acclimation and adaptation of enzyme physi-ology in Neurospora discreta. Fungal Ecology 35, 78–86.

Alvarez, G., Shahzad, T., Andanson, L., Bahn, M., Wallenstein, M. D., & Fontaine, S. (2018). Catalytic power of enzymes decreases with temperature: New insights for understanding soil C cycling and microbial ecology under warming. Global Change Biology 24(9), 4238–4250.

Åqvist, J., Isaksen, G.V., Brandsdal, B.O., 2017. Computation of enzyme cold adaptation. Nature Reviews Chemistry 1, 51.

Beleites, C. and Sergo, V., 2012. HyperSpec: a package to handle hyperspectral data sets in R. R package v. 0.98-20110927. http://hyperspec.r-forge.r-project.org

Bengtson, P., Bengtsson, G., 2007. Rapid turnover of DOC in temperate forests accounts for increased CO2 production at elevated temperatures. Ecology Letters 10, 783–90.

Berlemont, R., Martiny, A.C., 2013. Phylogenetic distribution of potential cellulases in bacteria. Applied and Environmental Microbiology 79, 1545–1554.

Biely, P., MacKenzie, C.R., Puls, J., Schneider, H., 1986. Cooperativity of Esterases and Xylanases in the Enzymatic Degradation of Acetyl Xylan. Bio/Technology 4, 731–733.

Blagodatskaya, E., Blagodatsky, S., Khomyakov, N., Myachina, O., Kuzyakov Y., 2016. Temperature sensitivity and enzymatic mechanisms of soil organic matter decomposition along an altitudinal gradient on Mount Kilimanjaro. Scientific Reports 6, 22240.

Bradford, M.A., 2013. Thermal adaptation of decomposer communities in warming soils. Frontiers in Microbiology 4, 333.

Burns, R.G., DeForest, J.L., Marxsen, J., Sinsabaugh, R.L., Stromberger, M.E., Wallenstein, M.D., Weintraub, M.N., Zoppini, A., 2013. Soil enzymes in a changing environment: Current knowledge and future directions. Soil Biology and Biochemistry 58, 216–234.

Chazdon, R.L., Chao, A., Colwell, R.K., Lin, S.-Y., Norden, N., Letcher, S.G., Clark, D.B., Finegan, B., Arroyo, J.P., 2011. A novel statistical method for classifying habitat generalists and specialists. Ecological Society of America 92, 1332–1343.

Davidson, E.A., Samanta, S., Caramori, S.S., Savage, K., 2012. The Dual Arrhenius and Michaelis-Menten kinetics model for decomposition of soil organic matter at hourly to seasonal time scales. Global Change Biology 18, 371–384.

Degrassi, G., Uotila, L., Klima, R., Venturi, V., 1999. Purification and properties of an Esterase from the Yeast Saccharomyces cerevisiae and Identification of the Encoding Gene These include : Purification and Properties of an Esterase from the Yeast Saccharomyces cerevisiae and Identification of the Encodin. Applied and Environmental Microbiology 65, 8–11.

Dontsova, K.M., Norton, L.D., Johnston, C.T., Bigham, J.M., 2004. Influence of Exchangeable Cations on Water Adsorption by Soil Clays. Soil Science Society of America Journal 68,

Dray, S., Dufour, A.B., 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. Journal of Statistical Software 22, 1 – 20.

Eichorst SA, Trojan D, Roux S, Herbold C, Rattei T, Woebken D., 2018. Genomic insights into the Acidobacteria reveal strategies for their success in terrestrial environments. Environmental Microbiology 20, 1041–1063.

Emmett, BA, ZL Frogbrook, PM Chamberlain, R Griffiths, R Pickup, J Poskitt, B Reynolds, E Rowe, P Rowland, D Spurgeon, J Wilson, CM Wood, 2008. Countryside Survey Technical Report No.03/07.

Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nature Reviews Microbiology 15, 579-590.

Frankenberger, W.T., Johanson, J.B., 1982. Effect of pH on enzyme stability in soils. Soil Biology and Biochemistry 14, 433–437.

German, D.P., Weintraub, M.N., Grandy, A.S., Lauber, C.L., Rinkes, Z.L., Allison, S.D., 2011. Optimization of hydrolytic and oxidative enzyme methods for ecosystem studies. Soil Biology and Biochemistry 43, 1387–1397.

Griffiths, R.I., Thomson, B.C., James, P., Bell, T., Bailey, M., Whiteley, A.S., 2011. The bacterial biogeography of British soils. Environmental Microbiology 13, 1642–1654.

Gweon, H.S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D.S., Griffiths, R.I., Schonrogge, K., 2015. PIPITS: An automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods in Ecology and Evolution 6, 973–980.

Heath, C., Xiao, P.H., Cary, S.C., Cowan, D., 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. Applied and Environmental Microbiology 75, 4657–4659.

Ihrmark, K., Bödeker, I.T.M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K.E., Lindahl, B.D., 2012. New primers to amplify the fungal ITS2 region - evaluation by 454-sequencing of artificial and natural communities. FEMS Microbiology Ecology 82, 666–677.

Kirk, G.J.D., Bellamy, P.H., Lark, R.M., 2010. Changes in soil pH across England and Wales in response to decreased acid deposition. Global Change Biology 16, 3111–3119.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D., 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology 79, 5112–5120.

Legendre, P., Gallagher, E., 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280.

Leprince, F., and H. Quiquampoix. 1996. Extracellular enzyme activity in soil: effect of pH and ionic strength on the interaction with montmorillonite of two acid phosphatases secreted by

the ectomycorrhizal fungus Hebeloma cylindrosporum. European Journal of Soil Science 47, 511–522.

Lidbury, I.D.E.A., Fraser, T., Murphy, A.R.J., Scanlan, D.J., Bending, G.D., Jones, A.M.E., Moore, J.D., Goodall, A., Tibbett, M., Hammond, J.P., Wellington, E.M.H., 2017. The 'known' genetic potential for microbial communities to degrade organic phosphorus is reduced in low-pH soils. MicrobiologyOpen 6, 1–5.

Lladó, S., Větrovský, T., Baldrian, P., 2019 .Tracking of the activity of individual bacteria in temperate forest soils shows guild-specific responses to seasonality. Soil Biology and Biochemistry 135, 275-282.

Lladó, S., Žifčáková, L., Větrovský, T. Eichlerová, I., Baldrian, P., 2016. Functional screening of abundant bacteria from acidic forest soil indicates the metabolic potential of Acidobacteria subdivision 1 for polysaccharide decomposition. Biology and Fertility of Soils 52, 251-260.

Lonhienne, T., Gerday, C., Feller, G., 2000. Psychrophilic enzymes: Revisiting the thermodynamic parameters of activation may explain local flexibility. Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology 1543, 1-10.

Harden, M. M., He, A., Creamer, K., Clark, M. W., Hamdallah, I., Martinez, K. A., Kresslein, R. L., Bush, S. P., Slonczewski, J.L., 2015. Acid-Adapted Strains of Escherichia coli K-12 Obtained by Experimental Evolution. Applied and Environmental Microbiology 81, 1932–1941.

Hong, S., Piao, S., Chen, A., Liu, Y., Liu, L., Peng, S., Sardans, J., Sun, Y., Peñuelas, J., Zeng, H., 2018. Afforestation neutralizes soil pH. Nature Communications 9, 1–7.

Humberstone, B.F.J., Briggs, D.E., 2000. Extraction and Assay of Ferulic Acid Esterase From Malted Barley *. Journal Of The Institute Of Brewing 106, 21–29.

Liang Y., Wu L., Clark IM., Xue K., Yang Y., Van Nostrand JD., Deng Y., He Z., McGrath S., Storkey J., Hirsch PR., Sun B., Zhou J., 2015. Over 150 years of long-term fertilization

alters spatial scaling of microbial biodiversity. mBio 6 (2) e00240-15.

Lin H., Chen, W., Ding H., 2013. AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. PLoS ONE 8 (10): e75726.

Macdonald, A. , Poulton, P. , Clark, I. , Scott, T. , Glendining, M. , Perryman, S. , Storkey, J. , Bell, J. , Shield, I. , McMillan, V. and Hawkins, J. 2018. Guide to the Classical and other Long-term experiments, Datasets and Sample Archive, Rothamsted Research.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. EMBnet.journal, pp. 10-12.

Marx, M., Wood, M., Jarvis, S., 2001. A microplate fluorimetric assay for the study of enzyme diversity in soils. Soil Biology and Biochemistry 33, 1633–1640.

Muyzer, G., de Waal, E.C., 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, 695-700.

Nannipieri P, Giagnoni L, Landi L. 2011. Role of phosphatase enzymes in soil. In: Bunemann E, Oberson A, Frossard E, eds. Soil Biology 100: 215–243.

Nannipieri P., Giagnoni L., Landi L., Renella G., 2011. Role of Phosphatase Enzymes in Soil. In: Bünemann E., Oberson A., Frossard E. (eds) Phosphorus in Action. Soil Biology, vol 26. pp 215-243.

Niemi, R.M., Vepsäläinen, M., 2005. Stability of the fluorogenic enzyme substrates and pH optima of enzyme activities in different Finnish soils. Journal of Microbiological Methods 60, 195–205.

NF ISO 10390, Soil quality., 2005. Determination of pH. AFNOR.

Nottingham, A.T., Turner, B.L., Whitaker, J., Ostle, N., Bardgett, R.D., McNamara, N.P., Salinas, N., Meir, P., 2016. Temperature sensitivity of soil enzymes along an elevation gradient in the Peruvian Andes. Biogeochemistry 127, 217-230.

Nottingham, A.T., Bååth, E., Reischke, S., Salinas, N., Meir, P., 2019. Adaptation of soil microbial growth to temperature: Using a tropical elevation gradient to predict future changes. Global Change Biology 25, 827–838.

Ohara, K., Unno, H., Oshima, Y., Hosoya, M., Fujino, N., Hirooka, K., Takahashi, S., Yamashita, S., Kusunoki, M., Nakayama, T., 2014. Structural insights into the low pH adaptation of a

unique carboxylesterase from Ferroplasma: Altering the pH optima of two carboxylesterases. Journal of Biological Chemistry 289, 24499–24510.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2012. vegan: Community Ecology

Puissant, J., Cécillon, L., Mills, R.T.E., Robroek, B.J.M., Gavazov, K., De Danieli, S.,Spiegelberger, T., Buttler, A., Brun, J.-J., 2015. Seasonal influence of climate ma-nipulation on microbial community structure and function in mountain soils. Soil Biology and Biochemistry 80, 296-305.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Core, T.R., 2014. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-117, http://CRAN.R-project.org/package=nlme.

Ramírez-Martínez, J. R., and A. D. McLaren. 1966. Some factors influencing the determination of phosphatase activity in native soils and in soils sterilized by irradiation. Enzymologia 31, 23–38.

Razavi, B.S., Liu, S., Kuzyakov, Y., 2017. Hot experience for cold-adapted microorganisms : Temperature sensitivity of soil enzymes Soil Biology & Biochemistry Hot experience for cold-adapted microorganisms : Temperature sensitivity of soil enzymes. Soil Biology and Biochemistry 105, 236–243.

Slessarev, E.W., Lin, Y., Bingham, N.L., Johnson, J.E., Dai, Y., Schimel, J.P., Chadwick, O.A., 2016. Water balance creates a threshold in soil pH at the global scale. Nature 540, 567–569.

Sinsabaugh, R.L., Lauber, C.L., Weintraub, M.N., Ahmed, B., Allison, S.D., Crenshaw, C., Contosta, A.R., Cusack, D., Frey, S., Gallo, M.E., Gartner, T.B., Hobbie, S.E., Holland, K., Keeler, B.L., Powers, J.S., Stursova, M., Takacs-Vesbach, C., Waldrop, M.P., Wallenstein, M.D., Zak, D.R., Zeglin, L.H., 2008. Stoichiometry of soil enzyme activity at global scale. Ecology Letters 11, 1252–64.

Sinsabaugh, R.L., 2010. Phenol oxidase, peroxidase and organic matter dynamics of soil. Soil Biology and Biochemistry 42, 391–404.

Storkey, J., Macdonald, A.J., Bell, J.R., Clark, I.M., Gregory, A.S., Hawkins, N.J., Hirsch,P.R., Todman, L.C., Whitmore, A.P., 2016. Chapter One - The unique contribution of Rothamsted to ecological research at large temporal scales. Advances in Ecological Research 55, 3-42.

Skujins, J., A. Puksite, and A. D. McLaren. 1974. Adsorption and activity of chitinase on kaolinite. Soil Biology and Biochemistry 6, 179–182.

Stuart, B.H., 2004. Infrared Spectroscopy: Fundamentals and Applications, Methods.

Tian, D., Niu, S., 2015. A global analysis of soil acidification caused by nitrogen addition. Environmental Research Letters 10.

Turner, B.L., 2010. Variation in ph optima of hydrolytic enzyme activities in tropical rain forest soils. Applied and Environmental Microbiology 76, 6485–6493.

Van Breemen, N., Mulder, J., Driscoll, C.T., 1983. Acidification and alkalinization of soils. Plant and Soil 75, 283–308.

Wallenstein, M., Allison S. D., Ernakovich, J., Steinweg, J. M., Sinsabaugh R., 2011. Controls on the temperature sensitivity of soil enzymes: A key driver of in situ enzyme activity rates, Soil Enzymoly 22, 245–258.

Wang, G., Post, W.M., Mayes, M.A., 2013. Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses. Ecological Applications 23, 255–272.

Wei, H., Guenet, B., Vicca, S., Nunan, N., AbdElgawad, H., Pouteau, V., Shen, W., Janssens, I.A., 2014. Thermal acclimation of organic matter decomposition in an artificial forest soil is related to shifts in microbial community structure. Soil Biology and Biochemistry 71, 1–12.

Hong, S., Piao, S., Chen, A., Liu, Y., Liu, L., Peng, S., Sardans, J., Sun, Y., Peñuelas, J., Zeng, H., 2018. Afforestation neutralizes soil pH. Nature Communications 9, 1–7.

Kirk, G.J.D., Bellamy, P.H., Lark, R.M., 2010. Changes in soil pH across England and Wales in response to decreased acid deposition. Global Change Biology 16, 3111–3119.

Slessarev, E.W., Lin, Y., Bingham, N.L., Johnson, J.E., Dai, Y., Schimel, J.P., Chadwick, O.A., 2016. Water balance creates a threshold in soil pH at the global scale. Nature 540, 567–569.

Tian, D., Niu, S., 2015. A global analysis of soil acidification caused by nitrogen addition. Environmental Research Letters 10.

van Breemen, N., Mulder, J., Driscoll, C.T., 1983. Acidification and alkalinization of soils. Plant and Soil 75, 283–308.

Wu, Y., Zeng, J., Zhu, Q., Zhang, Z., Lin, X., 2017. PH is the primary determinant of the bacterial community structure in agricultural soils impacted by polycyclic aromatic hydrocarbon pollution. Scientific Reports 7, 1–7.

Yan, SM., Wu, G., Prediction of Optimal pH and Temperature of Cellulases Using Neural Network. 2012. Protein & Peptide Letters 19, 29-39.

Yu, Y., Lee, C., Kim, J., Hwang, S., 2005. Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. Biotechnology and Bioengineering 89, 670-9.

Zanphorlin, L.M., De Giuseppe, P.O., Honorato, R.V., Tonoli, C.C.C., Fattori, J., Crespim, E., De Oliveira, P.S.L., Ruller, R., Murakami, M.T., 2016. Oligomerization as a strategy for cold adaptation: Structure and dynamics of the GH1 β-glucosidase from Exiguobacterium antarcticum B7. Scientific Reports 6, 1–14.

Zhalnina, K., Dias, R., de Quadros, P.D., Davis-Richardson, A., Camargo, A.O.F., Clark, I.M., McGrath, S.p., Hirsch P.R., Triplett, E.W., 2015. Soil pH Determines Microbial Diversity and Composition in the Park Grass Experiment. Microbial Ecology 69, 3395-406.

Zhang, J., Siika-aho, M., Tenkanen, M., Viikari, L., 2011. The role of acetyl xylan esterase in the solubilization of xylan and enzymatic hydrolysis of wheat straw and giant reed. Biotechnology for Biofuels 4, 60.

Table 1. Effect of soil field pH treatment (soil pH 5 vs soil pH 7) on soil properties. Values represent the mean (n=5) with the associated standard error (SE). Bold letters indicate significant differences (p<0.05).

	Units	Low pH (5)	High pH (7)
pH (H_2O)	-	**5.5 ± 0.0 a**	**7.3 ± 0.1 b**
Soil moisture	%	30.2 ± 1.1	31.5 ± 1.2
Total carbon content	%	**3.0 ± 0.1 b**	**3.9 ± 0.3 a**
CN ratio	-	10.7 ± 0.1	11.0 ± 0.1
Total nitrogen	%	**2.8 ± 0.1 b**	**3.5 ± 0.2 a**
Total phosphorus	mg/kg	54.0 ± 12.9	59.3 ± 2.5

Table 2. Effects of pH, soil treatment and interactions of both factors on relative enzyme activity at different assay pH (mixed model, overall repeated measures ANOVA tests).

	Assay pH		Field soil pH		Assay pH x field soil pH	
	F-value	P-value	F-value	P-value	F-value	P-value
Leucine amino-peptidase	190.1	**<0.001**	6.9	**0.03**	3.42	**<0.001**
Phosphatase	89.1	**<0.001**	51.4	**<0.001**	44.2	**<0.001**
ß-glucosidase	88.4	**<0.001**	23.4	**<0.01**	33.7	**<0.001**
Acetate esterase	397.2	**<0.001**	30.9	**<0.001**	38.4	**<0.001**

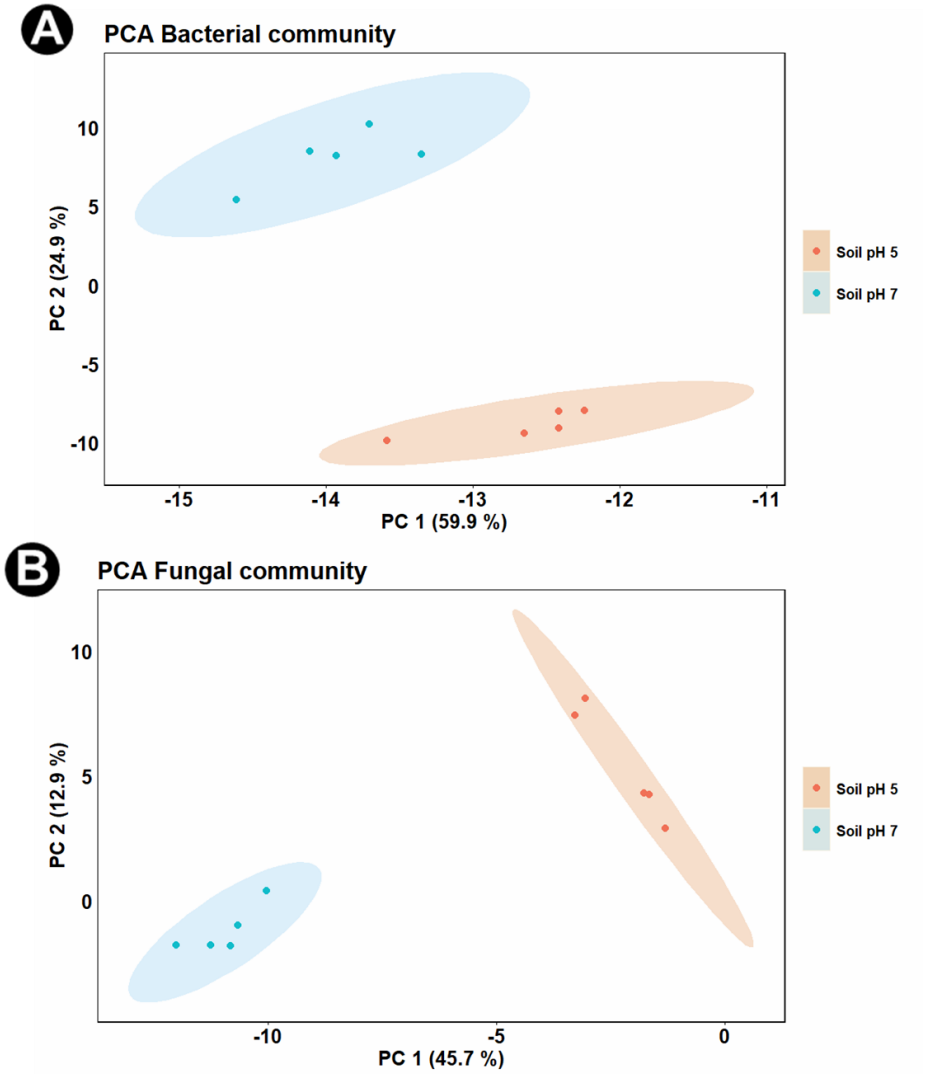


Fig.1. Principal component analysis (PCA) ordination of soil bacterial (A) and fungal (B) communities from grassland soil at either pH 5 or 7. The orange and blue colors correspond to pH 5 and pH 7 soils, respectively and ellipses indicate 95% confidence interval.

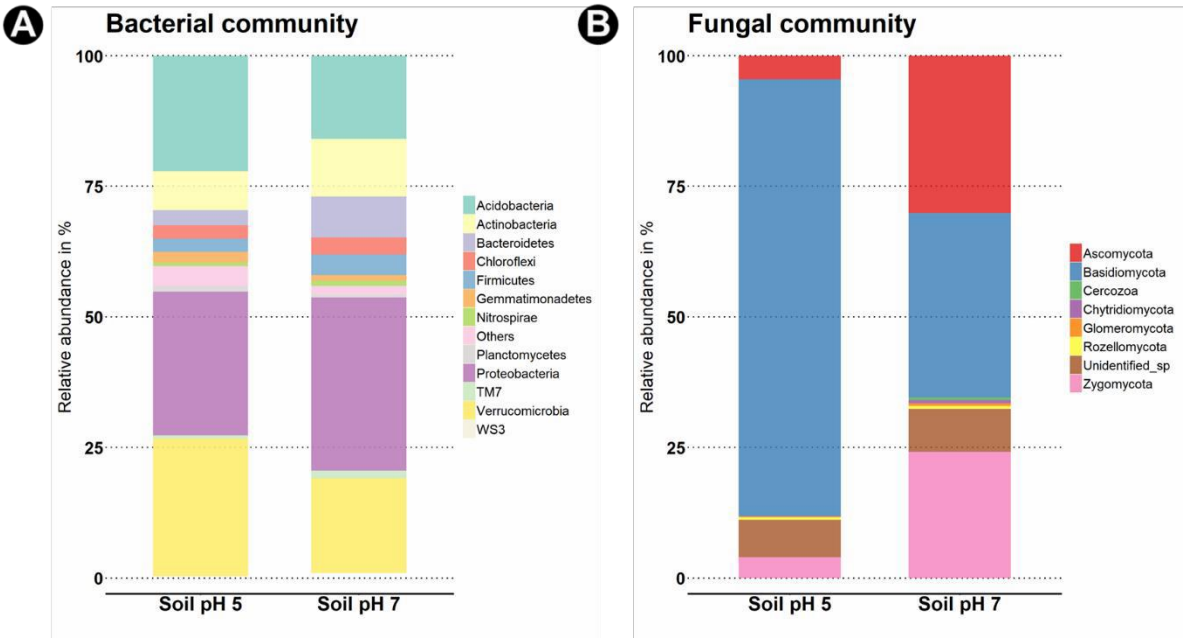


Fig.2. Stacked bar plots showing the mean relative proportion of abundant phyla (>0.5 %) for bacterial (A), and fungal communities (B), in grassland soils maintained long-term at either pH 5 or 7.

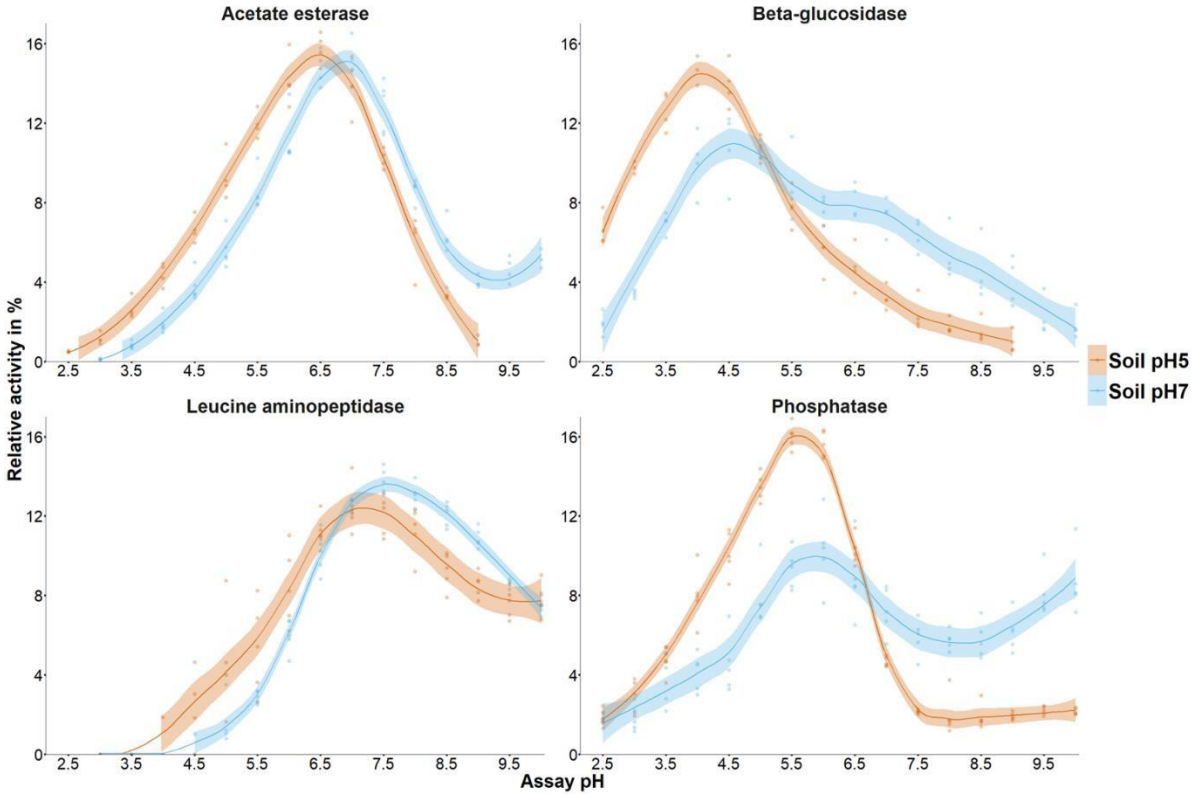


Fig.3. pH optima of acetylesterase (A), beta-glucosidase (B), leucine aminopeptidase (C), phosphomonoesterase (D) from grassland soils maintained at either pH 5 or 7. Activity is expressed as a percentage of the total activity measured across the entire pH range assayed (from pH 2.5 to pH 10). The orange and blue lines correspond to pH 5 and soil pH 7 soils, respectively. Shaded area represents 95% confidence intervals around the trend line using a t-based approximation (LOESS smoothing).

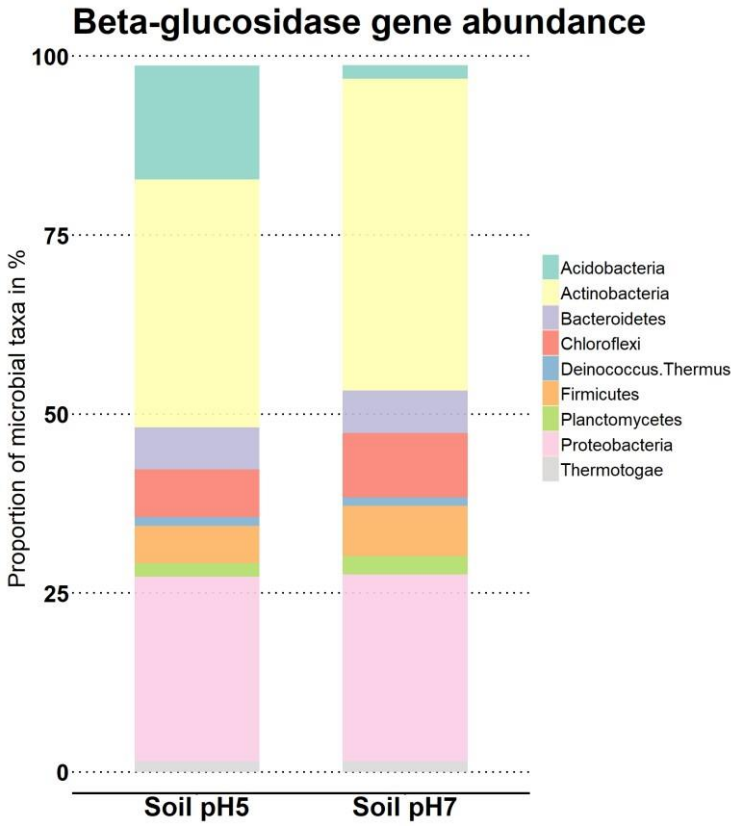


Fig.4. Mean abundances of beta-glucosidase genes from different microbial phyla, from MG-RAST annotated metagenomes (SEED Subsystems) from grassland soils maintained at either pH 5 or 7.

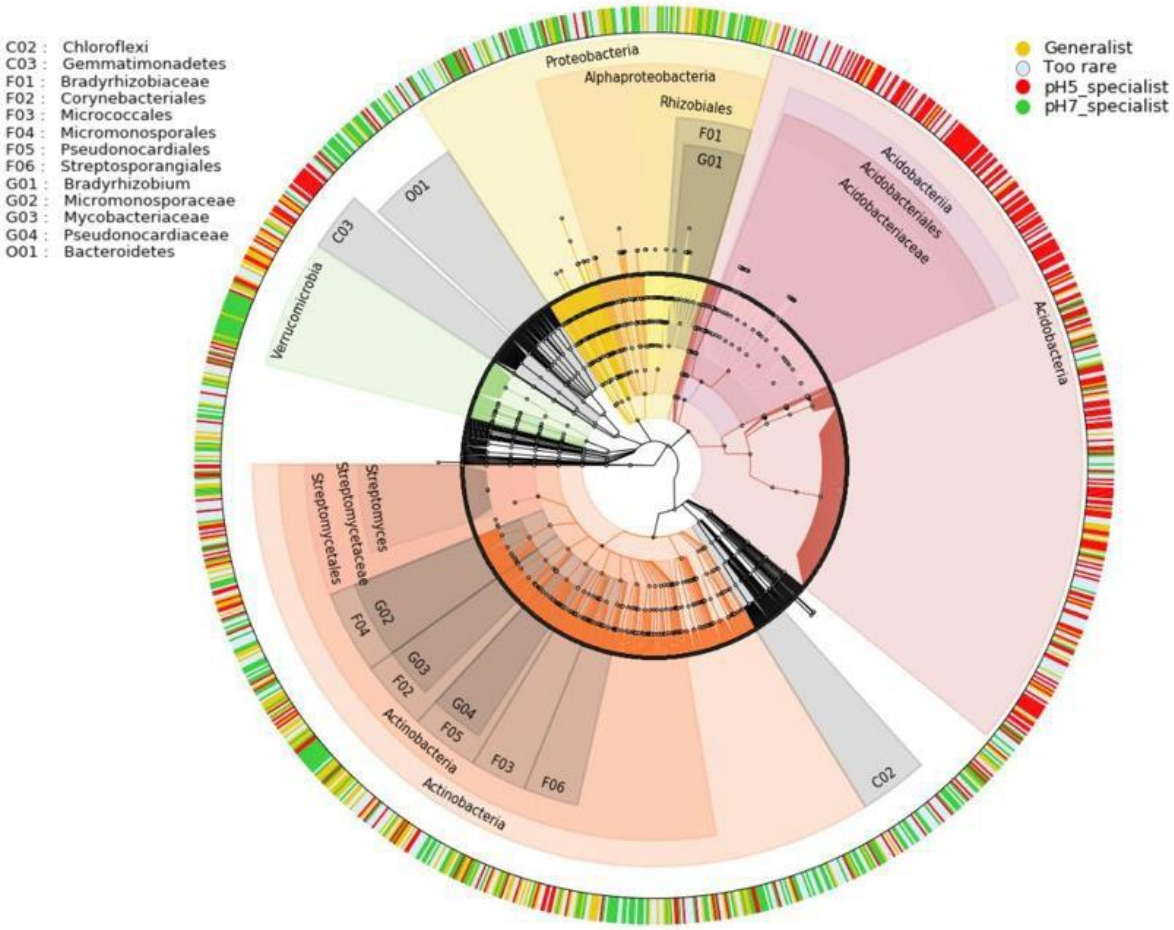


Fig.5. Detailed taxonomy and pH associations of β-glucosidase sequences assembled from metagenomes, showing Acidobacterial β-glucosidases are predominantly associated with the more acid soil. Inner tree and labels depict the taxonomy (from phylum to genus) of β-glucosidase gene assemblies constructed from pooled metagenomes from the pH 5 and pH 7 soils (n=4). Outer ring shows putative pH associations of each assembled gene, following tabulation of reads mapped to the contigs from each of the 8 soil metagenomes, and statistical classification using a multinomial model based on relative abundance across the two soils.

Supplementary Materials

Table.S1: Relative proportion of the main abundant phyla (>0.5 % proportion) for bacterial 748 and fungal phyla at soil pH 5 and soil pH 7.

Bacterial phyla	Soil pH 5		Soil pH 7	
	mean	se	mean	se
Acidobacteria	22.15	1.87	15.95	1.15
Actinobacteria	7.43	0.54	11.02	0.92
Bacteroidetes	2.87	0.66	7.83	0.96
Chloroflexi	2.54	0.20	3.24	0.42
Firmicutes	2.60	0.35	3.94	0.69
Gemmatimonadetes	2.02	0.54	1.10	0.24
Nitrospirae	0.64	0.16	0.94	0.17
Planctomycetes	1.08	0.15	0.55	0.08
Proteobacteria	27.48	1.11	33.20	1.34
TM7	0.65	0.08	1.42	0.23
Verrucomicrobia	26.31	1.73	18.13	1.41
WS3	0.37	0.06	0.98	0.20

Fungal phyla	Soil pH 5		Soil pH 7	
	mean	se	mean	se
Ascomycota	4.54	1.11	30.13	6.38
Basidiomycota	83.56	3.31	35.27	2.08
Cercozoa	0.04	0.02	0.51	0.11
Chytridiomycota	0.05	0.02	0.63	0.20
Glomeromycota	0.16	0.09	0.52	0.17
Rozellomycota	0.50	0.16	0.53	0.24
Zygomycota	4.03	2.00	24.12	4.53
Unidentified_sp	7.13	2.29	8.28	0.80

Table.S2: Percentage of beta-glucosidase gene sequences per bacterial phylum and found only at pH 7 soil (Specialist pH7), only at pH 5 soil (Specialist pH5), in both soils (Generalist) or too rare.

Phyla	Generalist	Specialist_pH7	Specialist_pH5	Too_rare
Unclassified Bacteria	25,0	8,3	8,3	58,3
Acidobacteria	6,9	4,7	48,3	40,1
Actinobacteria	20,3	28,5	14,6	36,7
Armatimonadetes	0,0	33,3	0,0	66,7
Bacteroidetes	5,9	47,1	7,8	39,2
Calditrichaeota	0,0	50,0	0,0	50,0
Zixibacteria	0,0	0,0	0,0	100,0
Candidatus Melainabacteria	0,0	0,0	100,0	0,0
Chloroflexi	2,8	30,6	11,1	55,6
Cyanobacteria	100,0	0,0	0,0	0,0
Deinococcus-Thermus	50,0	25,0	0,0	25,0
environmental samples	10,0	40,0	0,0	50,0
Euryarchaeota	25,0	25,0	0,0	50,0
Firmicutes	4,5	36,4	22,7	36,4
Gemmatimonadetes	9,7	3,2	58,1	29,0
Ignavibacteriae	0,0	33,3	33,3	33,3
Lentisphaerae	0,0	0,0	0,0	100,0
Planctomycetes	0,0	33,3	50,0	16,7
Proteobacteria	19,2	36,2	5,6	39,0
Spirochaetes	25,0	50,0	0,0	25,0
unclassified Bacteria	25,0	16,7	8,3	50,0
Verrucomicrobia	27,3	34,8	16,7	21,2

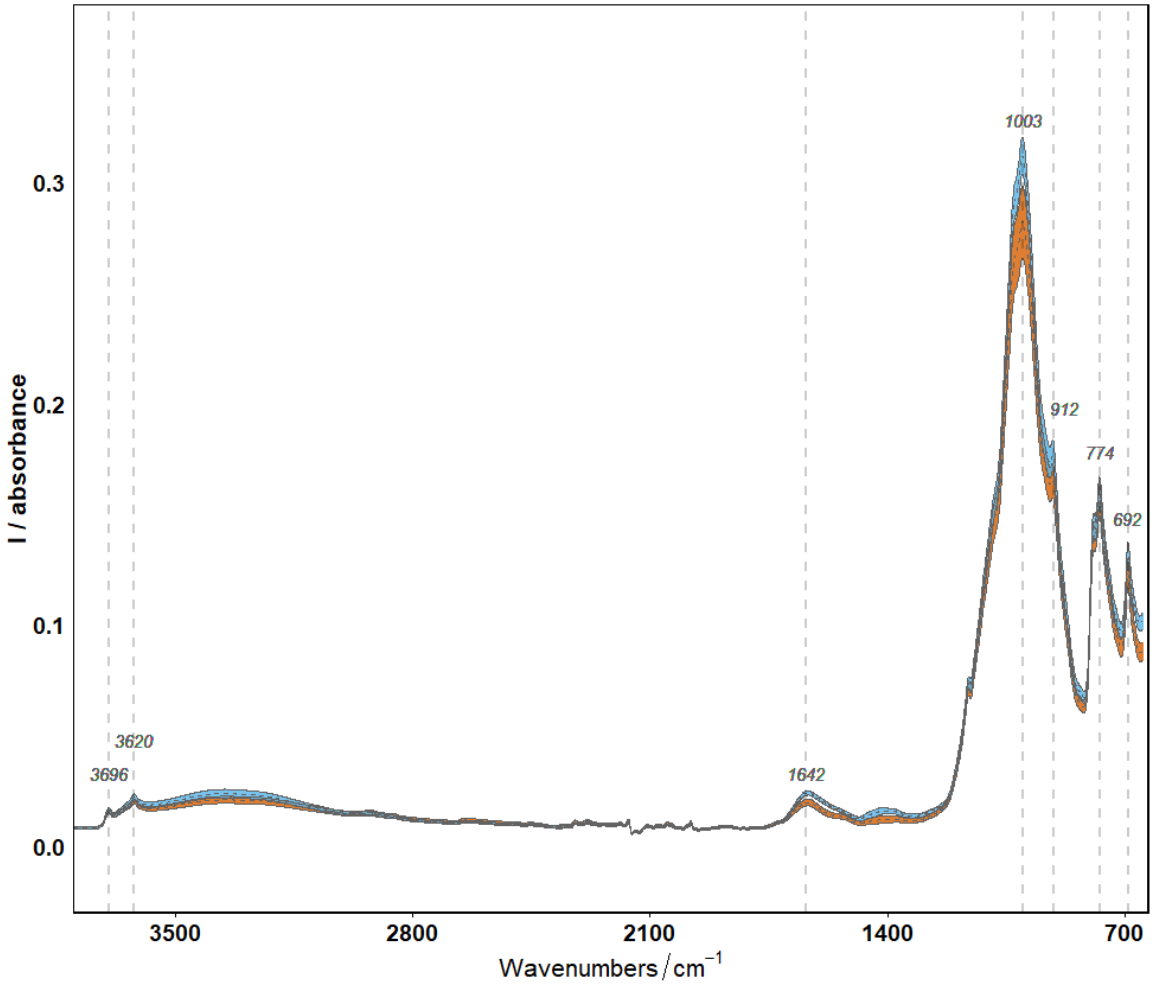


Fig.S1. Soil mid-infrared spectra for soils Nil plot pH 5 and Nil plot pH 7. Orange spectra correspond to soil pH 5 and blue spectra correspond to soil pH 7. The mid line indicates the mean spectrum (n=5) and the upper and lower lines indicate +/- standard deviation. Numbers written above spectra peaks indicate the wavelength for the main mid-infrared peaks observed.

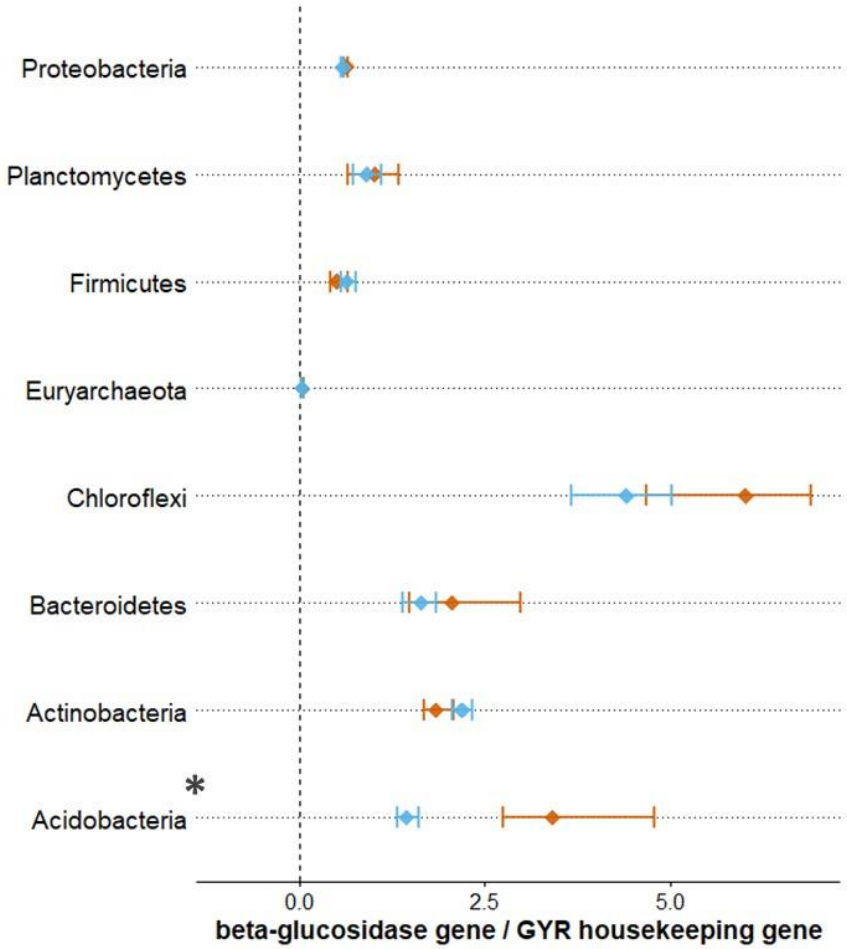


Fig.S2. The proportional change of beta-glucosidase gene abundance from different phyla, normalised to a housekeeping gene (DNA gyrase subunit B). Normalizing by housekeeping gene copy number allow evaluation of change in beta-glucosidase gene abundance regardless change in taxa abundance. Orange and blue colors correspond to pH 5 and pH 7 soil respectively. The x-axis shows the relative fold change on log2 scale. Error bars indicate +/- standard deviation and the means are indicated by filled diamond shape. Asterisks indicate significance difference between pH 5 and pH 7 soil (ANOVA $p<0.05$).

Appendix 2

Supplementary material for Chapter 4

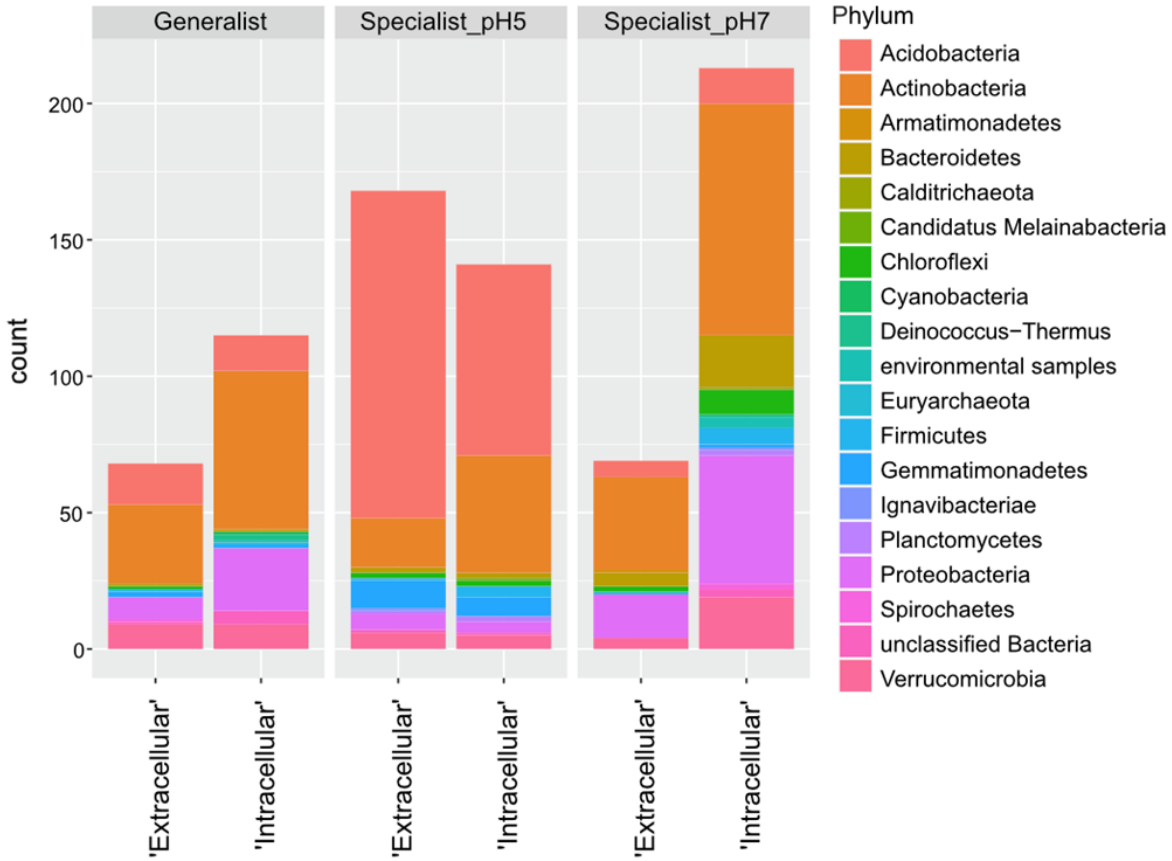


Fig.A4.1. Count of all β-glucosidase related sequences (annotated to GH1, GH2, GH3, GH5, GH9, GH30, GH39, GH116 using dbCAN2) subsetted by pH specialism, cellular location (inferred from secretory motif annotations conducted using SignalP) and phyla (annotated with Kaiju).

Appendix 3

Supplementary material for Chapter 5

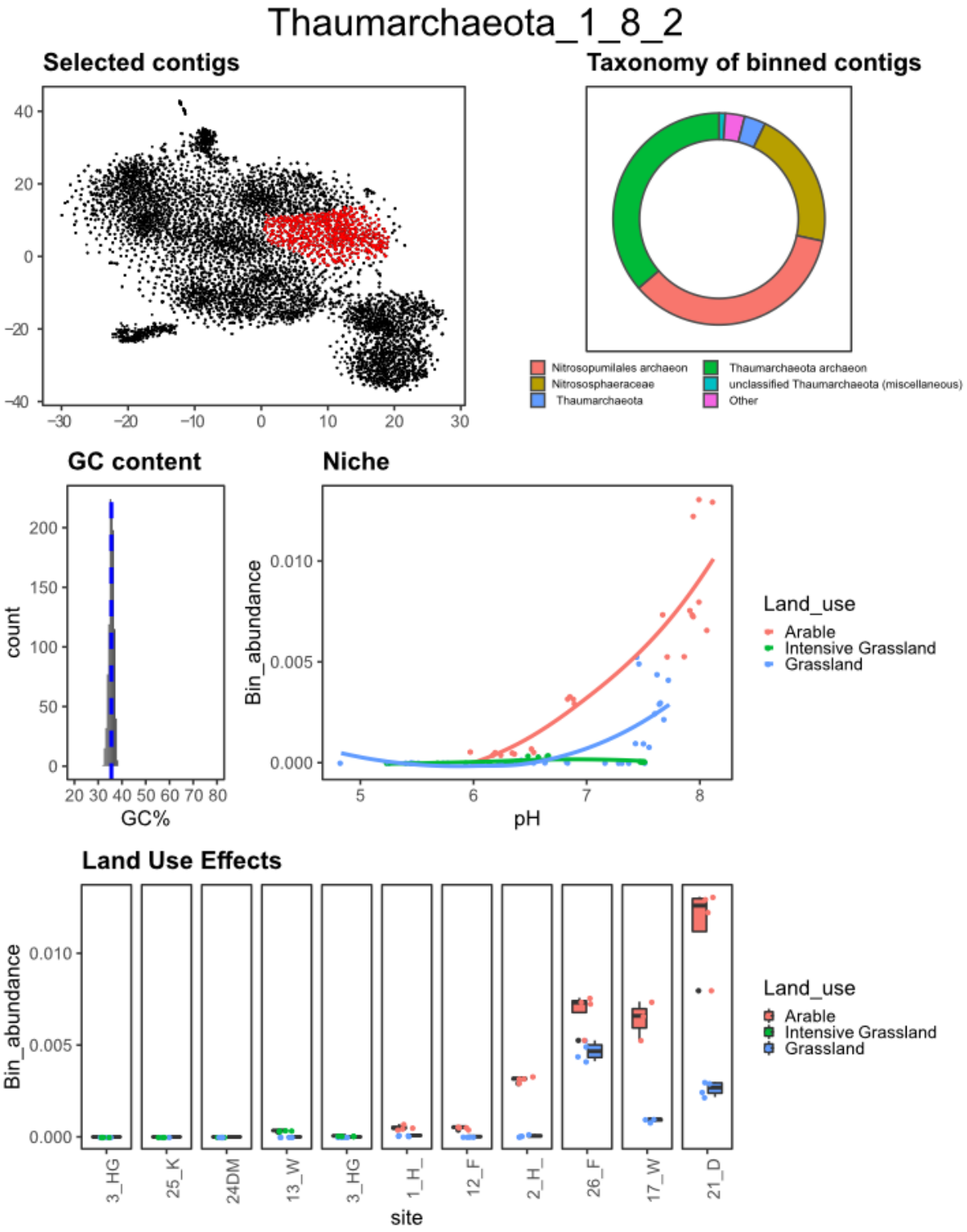


Fig.A5.1. Example of raw output from the developed manual metagenomic binning pipeline 'Bin_man' developed with my supervisor. Bin_man enables manual selection of points (representative of contigs) within a t-SNE plot (visualising similarity of contigs based on tetramer content) before producing graphical outputs shown based on contig selection and pre-existing files containing contig mapping information, taxonomic annotation and GC%'s. Output shows contig selection within t-SNE, contig Kaiju taxonomic annotation, GC% distribution of contigs , land use specific responses of contigs to pH and relative abundance of bin within each land use per sample site.

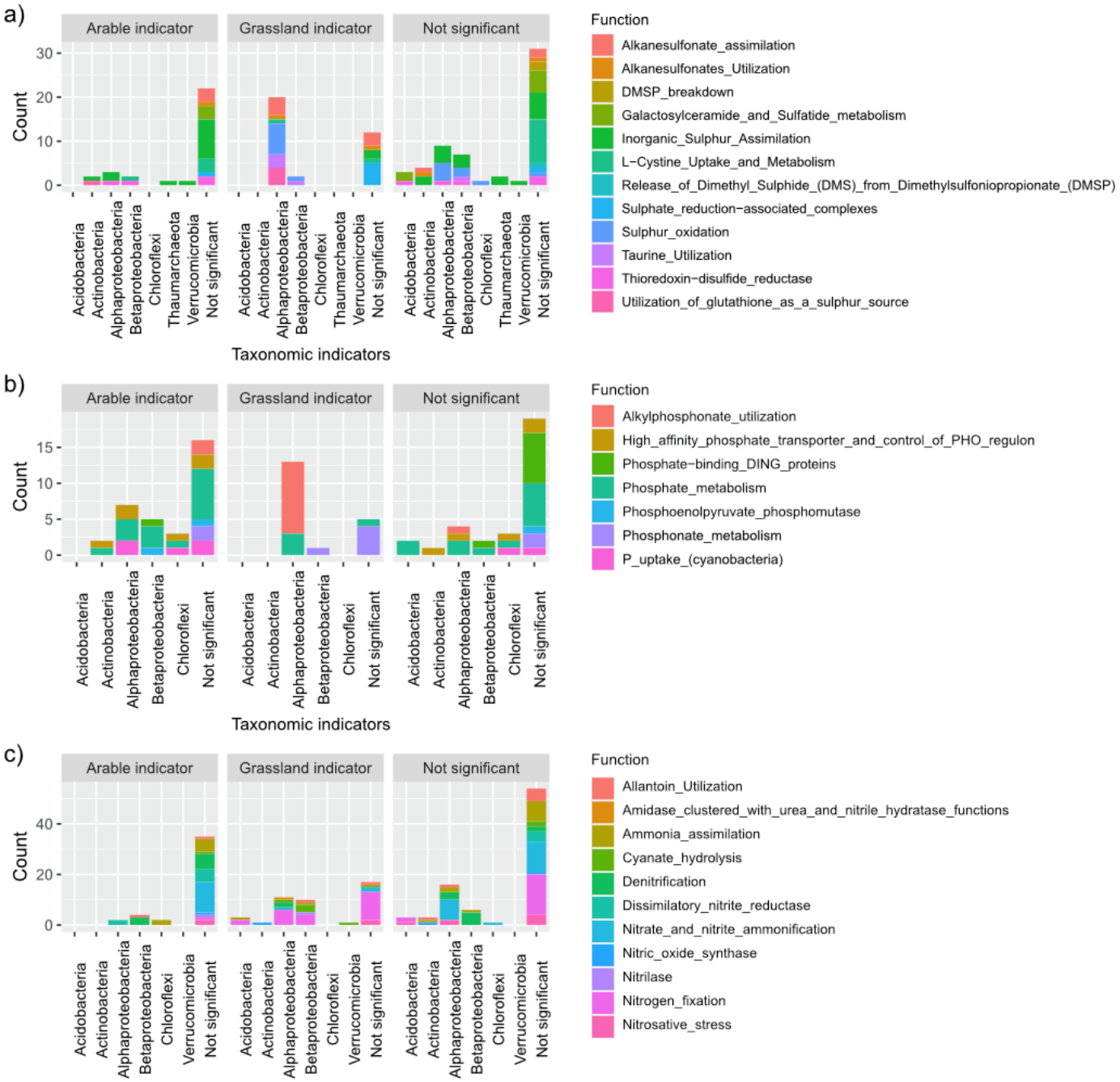


Fig.A5.2. Gene Indicators (dufrene-legendre indicator analyses) of phyla and land use within **a)** Sulfur metabolism subsystem, **b)** phosphorus metabolism and **c)** nitrogen metabolism. Phyla indicators are based upon presence and absence of functional genes within bins, land use indicators are based upon gene abundance from short read annotations.

Bin	Total length	No contigs	Completeness	Contamination	pH HOF model	pH HOF optima 1	pH HOF optima 2	pH class
Actinobacteria_1_17	3059675	485	87.57	1.75	V	7.771225	NA	Neutral
Actinobacteria_1_18	2025769	209	84.27	1.85	III	4.83	6.055588	Acid to Mid
Chloroflexi_1_15	2216787	214	85.34	4.18	V	5.67037	NA	Mid
Thaumarchaeota_1_3	1801748	280	86.89	8.25	V	8.095127	NA	Neutral
Thaumarchaeota_1_3_1	1573661	236	84.47	8.25	V	8.083943	NA	Neutral
Combined_Acidobacteria_1_8	4725535	448	86.13	9.7	II	4.830042	NA	Acid
Verrucomicrobia_1_1_1	2756857	345	82.93	13.31	V	5.905851	NA	Mid
Combined_Acidobacteria_1_17	7855856	218	95.73	15.88	III	4.83	5.412761	Acid to Mid
Verrucomicrobia_1_5_1_1	5023108	663	81.07	25.72	V	6.326931	NA	Mid
Chloroflexi_1_12_4	2763114	162	92	27.98	V	7.492785	NA	Neutral
Combined_Acidobacteria_1_22	9115454	398	98.28	31.5	II	4.830056	NA	Acid
Chloroflexi_1_16	5880690	956	84.48	32.29	V	5.709792	NA	Mid
Combined_Acidobacteria_1_9	7231782	1130	85.93	38.07	V	6.402575	NA	Mid
Thaumarchaeota_1_2_2	1325938	238	87.86	42.99	V	6.329498	NA	Mid
Verrucomicrobia_1_1_2_1	3109262	520	83.39	43.73	V	5.899195	NA	Mid
Combined_Acidobacteria_1_7	6064853	911	94.51	44.22	V	7.769877	NA	Neutral
Thaumarchaeota_1_2	1651958	297	87.86	44.29	V	6.322326	NA	Mid
Verrucomicrobia_1_5_1	6976065	1062	91.11	47.07	V	6.304953	NA	Mid
Verrucomicrobia_1_1_2	4384475	772	92.4	49.76	V	5.897747	NA	Mid
Alphaproteobacteria_1_7	6885827	1039	83.54	62.57	V	5.45919	NA	Mid
Alphaproteobacteria_1_37_2_1	8367794	1429	84.2	66.65	V	7.769413	NA	Neutral
Combined_Acidobacteria_1_25_1_1	5238072	611	88.71	71.11	V	5.650222	NA	Mid
Verrucomicrobia_1_8	3914564	902	84.58	76.04	V	6.970788	NA	Mid
Actinobacteria_1_15_2	9562459	1423	91.37	79.8	V	5.428887	NA	Mid

Betaproteobacteria_1_24	5841944	1272	80.41	85.65	V	5.419782	NA	Mid
Verrucomicrobia_1_5_2_3_1	9215812	1600	81.9	86.88	V	6.419222	NA	Mid
Actinobacteria_1_24	8993028	1776	91.61	90.83	V	6.655205	NA	Mid
Verrucomicrobia_1_5_2_3	11423118	2104	89.18	107.59	V	6.434765	NA	Mid
Thaumarchaeota_1_7_2	2617680	358	86.48	114.08	V	6.663567	NA	Mid
Combined_Acidobacteria_1_25_2_4	12061831	1454	94.58	117.08	V	5.006534	NA	Acid
Chloroflexi_1_1	12384087	2570	94.48	119.18	V	6.098527	NA	Mid
Actinobacteria_1_14_3	10490538	2389	82.65	124.29	V	7.757844	NA	Neutral
Verrucomicrobia_1_1	7168212	1124	99.69	125.58	V	5.902335	NA	Mid
Alphaproteobacteria_1_23	11649229	1658	82.76	128.06	V	7.482529	NA	Neutral
Betaproteobacteria_1_26_1	10510390	2004	94.25	130.49	II	8.119958	NA	Neutral
Verrucomicrobia_1_9_3	4805090	740	87.24	131.13	V	7.26669	NA	Neutral
Alphaproteobacteria_1_2	13128627	2115	81.54	133.07	V	5.828997	NA	Mid
Actinobacteria_1_11_2_2	14471322	2391	87.08	134.74	V	6.702717	NA	Mid
Alphaproteobacteria_1_6_2	7023579	1264	93.1	135.19	V	5.930615	NA	Mid
Actinobacteria_1_13	6509942	1123	95.61	144.1	V	5.725851	NA	Mid
Alphaproteobacteria_1_8_2	11478653	1262	96.55	156.23	V	5.11056	NA	Acid
Betaproteobacteria_1_14	5680263	1096	96.55	160.42	II	8.119958	NA	Neutral
Actinobacteria_1_9	7699809	1151	99.22	179.64	V	7.94417	NA	Neutral
Alphaproteobacteria_1_6	9360405	1738	93.1	179.86	V	5.854809	NA	Mid
Verrucomicrobia_1_5_2	18863226	3663	97.81	181.93	V	6.460208	NA	Mid
Betaproteobacteria_1_26	20737380	3946	98.28	184.8	II	8.119958	NA	Neutral
Combined_Acidobacteria_1_4_2_2_1	15654738	1765	98.28	188.31	V	6.946054	NA	Mid
Actinobacteria_1_14_1_4	11010400	2542	87.04	202.59	III	6.617852	8.12	Mid to Neutral
Thaumarchaeota_1_1_1_1	2977207	495	92.57	212.14	V	7.974273	NA	Neutral
Alphaproteobacteria_1_5_1	6905084	911	92.79	215	II	8.119958	NA	Neutral
Actinobacteria_1_25_4	11813902	1801	89.99	218.24	V	7.897002	NA	Neutral
Actinobacteria_1_25_1	22352895	4154	93.65	218.64	II	4.830042	NA	Acid
Thaumarchaeota_1_1_1	4774404	893	93.54	228.64	V	7.979708	NA	Neutral
Actinobacteria_1_26_1_1	13306557	2952	95.69	235.33	III	7.299652	8.12	Neutral

Actinobacteria_1_15_4	23066880	4328	87.15	237.99	V	5.86799	NA	Mid
Actinobacteria_1_25_3	12859988	2680	80.75	243.79	V	7.826217	NA	Neutral
Thaumarchaeota_1_7_1	3820387	627	88.35	248.11	V	7.974542	NA	Neutral
Thaumarchaeota_1_5_1	3162656	387	85.44	260.36	III	6.862895	8.12	Mid to Neutral
Actinobacteria_1_11_1_1_2	22863154	3983	97.49	277.75	V	7.355206	NA	Neutral
Alphaproteobacteria_1_8	16352464	2260	98.28	287.67	V	5.385283	NA	Mid
Combined_Acidobacteria_1_25_2_7	20701162	3406	100	316.25	V	4.969306	NA	Acid
Chloroflexi_1_12_3	17999492	3605	95.08	323.57	III	6.905178	8.12	Mid to Neutral
Thaumarchaeota_1_1	9688571	1901	97.91	325.08	V	7.978845	NA	Neutral
Combined_Acidobacteria_1_21_3_2	36199891	6591	91.38	331.5	V	6.854101	NA	Mid
Verrucomicrobia_1_5	26932228	4969	99.53	331.97	V	6.413631	NA	Mid
Actinobacteria_1_11_2	27821992	5385	99.69	345.55	V	6.751504	NA	Mid
Combined_Acidobacteria_1_4_1	26582235	3442	100	347.13	V	6.004134	NA	Mid
Combined_Acidobacteria_1_4_2_2	28438153	3814	100	355.32	V	6.924588	NA	Mid
Actinobacteria_1_26_1	19348575	4159	100	386.91	V	7.86688	NA	Neutral
Actinobacteria_1_11_1_1	34573846	6129	100	422.3	V	7.3531	NA	Neutral
Combined_Acidobacteria_1_4_2	33208442	4814	100	442.82	V	6.915632	NA	Mid
Thaumarchaeota_1_7	8538923	1377	100	466.46	II	8.119958	NA	Neutral
Actinobacteria_1_26	28269628	6177	100	468.46	V	7.854699	NA	Neutral
Thaumarchaeota_1_5	10276233	1591	95.83	499.13	III	6.896084	8.12	Mid to Neutral
Alphaproteobacteria_1_5_2	15544391	2305	93.1	528.5	V	6.91036	NA	Mid
Actinobacteria_1_30_1	28340584	5956	91.93	539.87	V	5.885017	NA	Mid
Alphaproteobacteria_1_37_1	54317111	10261	94.98	584.97	V	6.364698	NA	Mid
Actinobacteria_1_14_2_2	42736073	7887	95.45	601.47	V	7.730751	NA	Neutral
Alphaproteobacteria_1_37_2	1.27E+08	24873	95.83	612.52	V	6.864065	NA	Mid
Actinobacteria_1_19_1	32924039	5756	100	652.16	II	8.119958	NA	Neutral
Actinobacteria_1_10	21115930	4180	100	658	III	7.279241	8.12	Neutral
Verrucomicrobia_1_9_2	25759059	4335	100	664.17	V	7.758831	NA	Neutral
Actinobacteria_1_30_2	58150457	12025	97.65	674.94	V	7.214085	NA	Neutral
Actinobacteria_1_19	39601478	7115	100	687.86	II	8.119958	NA	Neutral

Alphaproteobacteria_1_35_3_2	60951958	7398	96.55	696.66	V	7.043714	NA	Neutral
Chloroflexi_1_12	30758495	5970	100	725.74	V	7.611853	NA	Neutral
Actinobacteria_1_14_2	55147184	10213	98.12	809.17	V	7.757897	NA	Neutral
Actinobacteria_1_14_1	41623358	8746	96.87	822.66	III	6.317734	8.12	Mid to Neutral
Alphaproteobacteria_1_5	26498294	3921	93.52	856.79	V	7.082029	NA	Neutral
Actinobacteria_1_15_5_2	99783981	19017	94.36	865.36	V	4.993544	NA	Acid
Actinobacteria_1_12_1	26308338	3331	99.69	878.24	V	7.447021	NA	Neutral
Actinobacteria_1_25	63022891	12289	98.15	883.59	II	8.119958	NA	Neutral
Actinobacteria_1_16	27598599	5143	98.59	883.62	V	8.021624	NA	Neutral
Alphaproteobacteria_1_27_3	53361736	9399	95.14	891.28	V	7.23887	NA	Neutral
Combined_Acidobacteria_1_4	61784581	8690	100	915.05	V	6.795471	NA	Mid
Combined_Acidobacteria_1_25_1	49072544	7397	99.22	926.58	V	5.356318	NA	Mid
Actinobacteria_1_5	83748504	15918	99.06	930.36	V	7.8018	NA	Neutral
Actinobacteria_1_11_1	51888783	9722	100	947.41	V	7.659425	NA	Neutral
Actinobacteria_1_15_5	1.25E+08	23920	97.49	1043.92	V	4.99315	NA	Acid
Alphaproteobacteria_1_35_3_1	36381632	4500	100	1060.24	V	7.039635	NA	Neutral
Actinobacteria_1_12_2_1	30244030	5830	99.53	1086.39	V	6.55743	NA	Mid
Betaproteobacteria_1	1.06E+08	21838	100	1168.7	II	8.119958	NA	Neutral
Alphaproteobacteria_1_37	1.97E+08	38477	100	1196.54	V	6.651434	NA	Mid
Verrucomicrobia_1_4_3	33758793	6906	100	1239.3	V	5.585884	NA	Mid
Actinobacteria_1_12_2	40386318	7806	100	1317.45	III	6.381564	8.12	Mid to Neutral
Verrucomicrobia_1_9	50101670	9131	100	1325.53	V	7.724459	NA	Neutral
Actinobacteria_1_11	82709383	15750	100	1353.39	V	7.391678	NA	Neutral
Verrucomicrobia_1_4	55909746	11835	100	1414.39	V	6.193391	NA	Mid
Combined_Acidobacteria_1_25_2	1.44E+08	25441	100	1460.34	V	4.960879	NA	Acid
Alphaproteobacteria_1_27	91021852	16902	97.18	1616.33	V	7.612866	NA	Neutral
Actinobacteria_1_14	1.1E+08	21867	99.84	1740.46	V	7.799426	NA	Neutral
Chloroflexi_1	1.04E+08	20677	100	1764.75	II	8.119958	NA	Neutral
Combined_Acidobacteria_1_21_3	1.54E+08	30018	100	1851.21	V	6.746639	NA	Mid
Thaumarchaeota_1	53879726	9574	100	1854.23	II	8.119958	NA	Neutral

Actinobacteria_1_15	1.9E+08	34705	99.37	1865.96	V	5.00406	NA	Acid
Alphaproteobacteria_1_35_3	1.01E+08	12713	100	1866.57	V	7.030793	NA	Neutral
Alphaproteobacteria_1_35	98699187	12457	100	1933.73	V	7.033298	NA	Neutral
Actinobacteria_1_30_3	53215900	10084	95.3	1983.55	V	7.514904	NA	Neutral
Actinobacteria_1_12	66423330	11134	100	2137.34	V	7.108308	NA	Neutral
Combined_Acidobacteria_1_21	1.69E+08	33391	100	2189.66	III	6.400336	8.12	Mid to Neutral
Combined_Acidobacteria_1_25	1.94E+08	33005	100	2446.49	V	4.962439	NA	Acid
Actinobacteria_1_30_4	95076979	19414	95.45	3015.65	III	5.979174	8.12	Mid to Neutral
Verrucomicrobia_1	1.67E+08	32855	100	3877.9	V	5.941742	NA	Mid
Actinobacteria_1_30	2.36E+08	47806	100	6635.51	III	6.2389	8.12	Mid to Neutral
Combined_Acidobacteria_1	5.06E+08	87474	100	6892.01	II	4.830042	NA	Acid
Alphaproteobacteria_1	6.91E+08	125714	100	8894.78	V	6.969583	NA	Mid
Actinobacteria_1	1.09E+09	209650	100	20503.46	II	8.119958	NA	Neutral

Table.A5.1. Statistics for metagenomic bins with a completeness of >80%. Completeness and contamination statistics were calculated with CheckM. Modelling statistics based upon HOF models on individual bin relative abundance, pH classifications assigned in reference to HOF model optima (classification described in further detail within **Chapter 5** methods section **5.2.7**).