DOCTOR OF PHILOSOPHY

Advancing knowledge of microbiallymediated lignocellulose degradation in soil using metagenomics and high-throughput in situ cultivation

Fidler, David

Award date:
2023

Advancing knowledge of microbially-mediated lignocellulose degradation in soil using metagenomics and high-throughput *in situ* cultivation

David Benjamin Fidler

March 2023

A thesis submitted to Bangor University

in candidate for the degree

Philosophiae Doctor

School of Natural Sciences

Bangor University, Deiniol Road, Bangor, LL57 2UW

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. I confirm that I am submitting this work with the agreement of my Supervisor(s).

Yr wyf drwy hyn yn datgan mai canlyniad fy ymchwil fy hun yw'r thesis hwn, ac eithrio lle nodir yn wahanol. Caiff ffynonellau eraill eu cydnabod gan droednodiadau yn rhoi cyfeiriadau eglur. Nid yw sylwedd y gwaith hwn wedi cael ei dderbyn o'r blaen ar gyfer unrhyw radd, ac nid yw'n cael ei gyflwyno ar yr un pryd mewn ymgeisiaeth am unrhyw radd oni bai ei fod, fel y cytunwyd gan y Brifysgol, am gymwysterau deuol cymeradwy. Rwy'n cadarnhau fy mod yn cyflwyno'r gwaith hwn gyda chytundeb fy Ngoruchwyliwr (Goruchwylwyr).

Summary

Soils are linked to almost half of the United Nations' Sustainable Development Goals, mostly through the ecosystem services afforded by soil organic carbon they contain (SOC). Food security, climatic stability, and maintenance of biodiversity depends upon SOC, which is largely comprised of lignocellulosic plant biomass. In combination with lignocellulose inputs to soil, the microbial degradation of lignocellulosic polymers into simple sugars is the key regulator of these life-supporting services. Despite this, the relative importance of different microbial taxa in lignocellulose degradation, or the effects of anthropogenic changes, on microbial lignocellulose decomposition are barely understood. Overarching these two challenges is the dearth of knowledge about soil microbial diversity and function. The largest commercial collection of microorganism isolates contains 20,300 species of an estimated one trillion, with significant bias in the represented groups. Our current knowledge of microbial interactions in soils is therefore severely limited. To address these knowledge gaps, this body of work has focused on (1) increasing knowledge about the relative contributions of microbial species and broad taxonomic groups to the degradative potential of lignocellulose in soils, (2) increasing knowledge about how global changes affect the genetic potential of microbial communities, and (3) increasing the diversity of cultivated microbial species to improve characterisation and prediction of microbial community dynamics.

Chapter 1 reviews the literature on microbial lignocellulose decomposition in soil. Chapter 2 combines metagenomics, metabolomics, and fibre analysis data on soil from a decade-long field experiment, to understand the impact of plant inputs on grassland microbial community composition and associated genes. The majority of lignocellulolytic genes originated from *Actinobacteria*, *Proteobacteria*, *Bacillota*, *Bacteroidota,* and *Planctomycetes*. Decade-long plant exclusion resulted in communities with high proportions of *Bacillales*, *Thermoproteota*, and *Proteobacteria*, and the composition of lignocellulolytic genes was biased towards the cellulolytic glycoside hydrolase family 5. A single year of plant-exclusion biased the composition of lignocellulolytic genes towards xylanases. Chapter 3 uses data from the UK Soil Security Programme's UGRASS experiment to understand how agricultural intensification impacts different phylogenetic and functional microbial groups, and their genes for lignocellulolytic enzymes.

Agricultural intensification decreased microbial abundance, drastically increased microbial taxonomic diversity (likely as an artefact of relic DNA), and increased relative abundance of cellulase genes. Chapter 4 used high-throughput *in situ* cultivation to isolate and cultivate lignocellulose degrading microorganisms from soil. Despite analysis of only 83 isolates, we discovered seven new species from commonly isolated soil microorganisms (predominantly *Pseudomonas*), highlighting the efficacy of high-throughput *in situ* cultivation for the isolation of new species. Genome annotation and pan-genome-wide association of accessory genes, from *Pseudomonas* isolates, identified likely causative genes of degradative phenotypes in *in vitro* tests, as well as genes which likely contributed to the rate of utilisation of different lignocellulosic polymers. Chapter 5 highlights emerging challenges for the study of lignocellulose degradation in soils, places the findings from each chapter in the wider context of global challenges, and points the way for future research to aid with societal challenges. This body of work presents advances to our knowledge of the interactions of genes and microorganisms with the major element of SOC. It addresses knowledge gaps about the identity and relative abundances of microorganisms with lignocellulolytic potential in soils, furthers our understanding of the implications of land use change on microorganisms and genes, and gives broad perspectives on the life-history strategies of *Pseudomonas* isolates which utilise different lignocellulosic polymers. Deepening our understanding of these processes will allow us to devise more effective management strategies to improve sustainable use of soil carbon for food production, climate change offsetting and biodiversity conservation. Together, these can help to reduce global inequality and allow the better formulation of environmental policy.

Acknowledgements

This thesis would not exist or would be much poorer without the guidance, insights, support, help, understanding, and banter from Prof. James McDonald, Prof. Davey Jones, and Prof. Rob Griffiths. You've been wonderful to work with throughout the project, and I would gladly do it again. Your availability for chats about experimental design, the implications of results, writing, "This thing's gone very wrong, what do I do?", and general chats about life and science have meant a lot to me throughout the course of this work. Thank you.

The Molecular Ecology Group (McDonald lab) in all its iterations has been an excellent collaborative environment to work in, which has given me some great and lasting friendships. The McDonald lab enriched the work outcomes and the enjoyability of the process—thank you all, you are diamonds. The many friends made throughout Bangor University have also been a large part of making the programme so enjoyable, cheers to all of you. Special mentions have to go Luke Hilary and Beth Pettifor, you've been great close companions throughout the whole journey, and to Maria Majka (and Tova Majka), you have helped me keep it together and have reminded me to have some fun amidst all the working. I would like to thank the Soils Training and Research Studentships (STARS) centre for doctoral training for providing funding for me to complete my studies. STARS has been great at providing training, extra funding for experimental work, interaction with talented and lovely students and supervisors and external scientists in the field of soil science and beyond, not to mention particularly fun trips in the UK and abroad. I would also like to thank the Wales Biodiversity Project Biodiversity and Ecosystem Evidence and Research Needs (BEERN) Programme for part-funding the work in Chapter 3.

Finally, thank you mum and dad for your unwavering support and chats. Life would have been much harder without you throughout the PhD and the Covid-19 pandemic, and I am eternally grateful.

Additional scientific contributions

Publications

George, P. B. L. *et al.* (2021) 'Shifts in Soil Structure, Biological, and Functional Diversity Under Long-Term Carbon Deprivation', *Frontiers in Microbiology*, 12, 2509. doi: 10.3389/fmicb.2021.735022.

Jenna L. Alexander, L.J. *et al.* (2021) 'Improving quantification of bivalve larvae in mixed plankton samples using qPCR: A case study on Mytilus edulis', *Aquaculture*, 532, 2021. doi: 10.1016/j.aquaculture.2020.736003.

Oral presentations

Fidler, D.B. 'Metagenomics and high-throughput *in situ* cultivation reveal how plant exclusion alters grassland microbial community and lignocellulolytic potential', International Society for Microbial Ecology 18, Aug. 2022, École Polytechnique Fédérale de Lausanne (EPFL), Oral presentation

Fidler, D.B. 'Using metagenomics and *in situ* cultivation to understand microbial degradation of lignocellulose in soil', Molecular Microbial Ecology Group Meeting, Dec. 2018, Swansea University, *Registration grant awarded by FEMS. Runner-up for IJSEM Microbial Diversity, Identification and Taxonomy Journal Prize*

Peer-review

I reviewed a paper for the Springer journal Waste and Biomass Valorization.

Table of contents

1

Introduction

1.1 Importance of lignocellulose in soils and its breakdown

1.1.1 Lignocellulose degradation and carbon storage in soil

Soils are Earth's largest pool of terrestrial organic carbon, storing 2500 Pg (gigatonnes) of carbon (Pg C) Worldwide in the first metre (Jobbágy and Jackson, 2000; Lal, 2008)—3.1 times more than is stored in the atmosphere, and more than is stored in the atmospheric and biotic carbon pools combined (Jobbágy and Jackson, 2000; IPCC, 2007; Lal, 2008; Scharlemann *et al.*, 2014; FAO, 2015). Soils provide a vast array of ecosystem services including supporting biodiversity, nutrient cycling, water cycling, regulation of climate and of disease, biomass production, and acting as a supporting service for recreation. The value of biomass production alone from soils has been estimated as roughly USD 230 – 22,000 per hectare per year, and nutrient cycling at 24 – 180 international dollars per hectare per year (Jónsson and Davíðsdóttir, 2016) making individual soil services worth trillions of international dollars globally. It is estimated that about 1500 Pg (two-thirds) of soil carbon is stored as soil organic carbon (SOC), which is mainly derived from dead plant material (90% lignocellulose), but also includes dead animals, bacteria, fungi, and root exudates (Lal, 2008; Scharlemann *et al.*, 2014; FAO, 2015). Soil organic carbon is the portion of soil carbon that promotes the productivity, carbon storage, and water retention capacity of soils, and reduces soil compactibility and erodibility (Carter, 2002).

The world's soils are under threat from anthropogenically driven global change. Land use change is a major factor which affects the functioning of ecosystems. Globally, land use change causes the loss of ecosystem services to the value of USD 4.3–20.2 trillion per year (Costanza *et al.*, 2014). As an example, conversion from pasture to crop-land quickly reduces soil carbon by roughly 60% within 25 years (Guo and Gifford, 2002). UK croplands typically lose 140 ± 100 kg of carbon per hectare per year, whereas pasture sequesters 110 ± 4 kg per hectare per year (Ostle *et al.*, 2009). This loss of soil carbon on large scales translates into increased human malnutrition and hidden hunger, which affects billions of people worldwide (Lal, 2009; Lowe, 2021). In addition to this,

conversion to agricultural land causes dramatic reductions to the biodiversity of macroorganisms (Dudley and Alexander, 2017; George *et al.*, 2021), as can agricultural intensification (production of more crops per unit of cropland). Atmospheric warming by 1°C has the potential to drive a decrease in global soil carbon stocks in the first 10 cm of up to 364 PgC over the next 35 years, with an increase in the rate of microbial respiration in soils being a major factor affecting this (Crowther *et al.*, 2016), although see (Crowther *et al.*, 2018; van Gestel *et al.*, 2018). The implications of the enhanced loss of soil carbon to the atmosphere include increased atmospheric warming, with a subsequent positive feedback loop (Crowther *et al.*, 2016). Preventing further loss of SOC in the face of multiple global changes is therefore critical to ensuring that the ecosystem services provided by soils will continue effectively (FAO, 2015).

The vast majority (90%) of dry plant mass, and thus SOC, is lignocellulose (Brandt *et al.*, 2013). Lignocellulose is a composite of the polymers cellulose, hemicellulose, and lignin, which together are relatively recalcitrant to degradation by most microorganisms, because of the numerous bonds between and within their constituent parts (Brandt *et al.*, 2013). The hydrolysis of these polymer chains (conversion of soil organic matter into dissolved organic matter) is the rate-limiting step in microbial decomposition of soil organic matter (Burns and Dick, 2002; A'Bear *et al.*, 2014). Changes to the rate at which lignocellulolytic soil microbes make lignocellulosic carbon available to non-lignocellulolytic members of the microbial community (in the form of simple sugars) could have global impacts on the rate of carbon emission from soils (FAO, 2015), meaning it is critical that we have a good understanding of the mechanisms and organisms involved in this process.

For example, Earth system models (ESMs) that are currently used to predict soil carbon responses to climate change, whilst providing valuable insights into how soils may respond to climatic change, are highly underparameterized (Wieder, Bonan and Allison, 2013), mal-parameterized (Kolby Smith *et al.*, 2015), and lack sufficient detail with respect to the microbial processes involved in decomposition of carbon-based compounds in soils (Wieder, Bonan and Allison, 2013; Crowther *et al.*, 2016). Current widely used ESMs rely upon first-order relationships between decomposition rate and temperature, soil moisture and clay content (Coleman and Jenkinson, 1996; Kelly *et al.*, 1997; Parton *et al.*, 1998). These equations are calculated for several pools of carbon which are more or less easily degraded (determined by *e.g.*, the ratio of lignin to nitrogen) across many time-steps, with carbon moving from slow-cycling to faster-cycling pools. Microbial models, by contrast, incorporate data about microbial biomass, splitting the microbial community

into groups with different life-histories and carbon-pool preferences. These carbon pools are then converted by the microorganisms following realistic temperature-sensitive (Michaelis-Menten) equations, with the products from this simulated degradation providing positive feedbacks to the size of the microbial community (Zhang *et al.*, 2020). Microbial models of soil carbon dynamics are gaining popularity in the academic literature, with various studies altering the model structures by including parameters for the chemical protection of SOC by sorption to clay minerals, chemical deprotection of SOC through the opposite mechanism, known pH effects on decomposition rates, density-dependent microbial carbon utilisation, utilisation of microbial carbon pools, the effects of nitrogen availability on microbial carbon use efficiency, and degradative enzyme production rates among others (Zhang *et al.*, 2020; Fan *et al.*, 2021; Huang *et al.*, 2021). Modified microbial soil carbon models have increased the accuracy of SOC stock estimates across continents relative to those which rely upon first-order relationships with temperature.

Through the use of second- and third- generation sequencing technologies, we are beginning to understand more about the spatial distribution of soil microbial biodiversity, and are gaining appreciation of the genetic functional potential of these organisms, through the genomic analyses these technologies enable (see section 1.3.1 and 1.3.3) (Griffiths *et al.*, 2011; Delgado-Baquerizo *et al.*, 2018; Wilhelm *et al.*, 2019; López-Mondéjar *et al.*, 2020; Nuccio *et al.*, 2020). As our understanding of soil microbial species composition and associated functional relationships increases, it seems logical that ESMs should begin to incorporate more spatially explicit taxonomic or functional microbial data. However, the success of this approach will rely upon accurate annotation of the nucleic acid sequences of soil microbial communities; the ability to spatially predict change; and importantly, evidence describing the mechanisms by which change in functional genetic potential translates to altered process. However, with respect to plant decomposition, we are still at the stage of needing better characterisation of the broad diversity and functional activity of lignocellulose-degradation mechanisms in soils. Novel mechanisms of lignocellulose degradation can only reliably be classified *via* isolation, cultivation and phenotypic testing of microorganisms in the laboratory. Phenotypic testing of novel species and isolates could provide quantitative information about the rates at which the organisms and enzymes in an environment degrade lignocellulose for input into ESMs which should, in combination with taxonomic information provide a greater depth of understanding of carbon dynamics in soils. However, a major barrier to this approach is the intractability of as much as 99% of soil

microorganisms to cultivation, under standard laboratory conditions (Pham and Kim, 2012). Additionally, it would not be practical to test cultures, under all biotic and abiotic conditions present in "field" conditions. Therefore, it is likely that a combination of modern molecular methods applied to assess *in situ* diversity and functional potential, must be used in combination with culture-based studies to inform our understanding of ecosystem processes.

A resurgence in interest in the cultivation of microorganisms, due to innovative techniques such as alterations to culture media (Nichols *et al.*, 2008; Vartoukian, Palmer and Wade, 2010), and high-throughput *in situ* cultivation (Nichols *et al.*, 2010), has led to the culture of up to 50% of cells from soil in a single study, and the discovery of genes for a novel antibiotic class which could have large biomedical significance (Nichols *et al.*, 2010; Ling *et al.*, 2015b). Such an approach can also be utilised to address the topic of carbon conversion in soils, improving our fundamental knowledge about how the genes from diverse soil microbes affect turnover of lignocellulose. By combining these cultivation approaches with metagenomic analyses of soil microbial communities (under "ambient" conditions and altered conditions, *e.g.*, elevated temperature, elevated CO_2, elevated ozone, land use intensification *etc.*), we can build a strong foundation for improving the accuracy of ESM projections (Crowther *et al.*, 2016; Mock *et al.*, 2016) leading to better informed management strategies for the improvement and maintenance of SOC stocks.

1.1.2 Lignocellulose degradation and biotechnology

1.1.2.1 The cellulase market: value and innovation

The global market value for industrially important enzymes in 2014 was estimated at USD 4.2 billion (Singh *et al.*, 2016). Within this market, amylases are the most important enzyme type, being worth around 25% of the market (de Souza and de Oliveira Magalhães, 2010), and cellulases are the second most important type, being responsible for 20% of the total market value (Srivastava *et al.*, 2014). Cellulases play important roles in large industries. They are used for production of second-generation biofuels, improving digestibility of livestock feed, improving the colour and aroma of alcoholic and nonalcoholic drinks, increasing the softness of fabrics, and are crucial for the action of commercially available detergents (Kuhad, Gupta and Singh, 2011; Juturu and Wu, 2014). Industries currently use a cocktail of hydrolytic and oxidative enzymes for the degradation of cellulose, utilising enzymes which are active on different physical locations in the cellulose (*e.g.*, exoglucanases, endoglucanases, β-glucosidases), or which have different modes of action (*e.g.,* glycoside hydrolases, lytic polysaccharide monooxygenases).

There is currently a significant amount of development of cellulases for commercial uses. A search of patents listed on Espacenet (https://worldwide.espacenet.com, date accessed: 2018-03-27) for "cellulase" returned 153 patents with a publication date between 2016-01-01 and 2018-01-01, so as to show where innovation for cellulases have been recently at the time of writing. The same search for "amylase" returned only 26 patents, showing significant development of cellulases, relative to the more widely used amylases. Within the titles of the 153 patents for cellulases, the words "gene" and "preparation" were present 15 times, "*Trichoderma*" was present 12 times, and "strain" appeared 11 times (Figure 1). The words "expression", "*reesei*", "*Aspergillus*", and "fermentation" occurred 9 times. The prevalence of these words is linked with the current research needs in terms of the need for novel cellulases, and for microbial strains which are effective at producing large quantities of the enzyme (Li *et al.*, 2012).

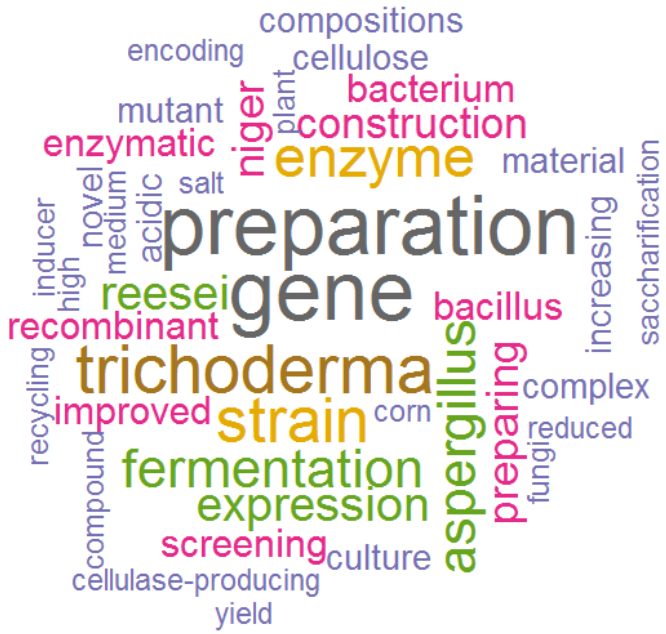

Figure 1: Frequently occurring words in the titles of the 153 patents listed on Espacenet (https://worldwide.espacenet.com) published between 2016 and 2018. Word colour and size is indicative of word frequency. The search term for the tiles was "cellulase", and the word cloud shows only words which appeared four or more times.

Isolation of microorganisms which produce cellulases that are novel, resistant to extreme conditions, or which are particularly effective at depolymerizing cellulose, is a priority for industries utilising cellulases, so that they may improve production efficiency (Juturu and Wu, 2014). The high occurrence of the genera "*Trichoderma*" and "*Aspergillus*", in the titles, after decades of study into their cellulase systems (Montenecourt and Eveleigh, 1977; Henrissat *et al.*, 1985; Tjerneld *et al.*, 1985; Divne *et al.*, 1994; Martinez *et al.*, 2008; Bischof *et al.*, 2013), shows

the potential wealth of biotechnological advancements which could come from cultivation of so far uncultivated species. Indeed, the vast majority of the cellulase industry uses enzymes derived from these genera and they make up over half of the academic literature in this area (de França Passos, Pereira and de Castro, 2018). Also of economic importance is the genus *Penicillium* (Adsul *et al.*, 2020).

1.1.2.2 Lignocellulose for biofuel production

One potentially lucrative application of lignocellulolytic enzymes is the production of renewable energy from second-generation biofuel. Whilst burning biofuels reduces carbon emissions relative to burning fossil fuels, biofuels do have their environmental limitations. Production of first-generation biofuel has adverse effects for biodiversity through increased land-use change, and also reducing the biomass of food crops being used for food production—this is a problem with increasing global food demands (Naik *et al.*, 2010). Increased first-generation biofuel production has been linked to modest increases to the cost of food in the United States of America, demonstrating the real-world consequences of the energy industry for the average person (Mueller, Anderson and Wallington, 2011). Second-generation biofuels do not generally compete with food crops, as a different portion of the plant, or more woody plants, may be used for second-generation biofuel production (Zhang and Bao, 2017). However, some crops are grown exclusively for cellulosic biofuel; *Miscanthus* is one of these, although it may be grown in marginal and contaminated soils where they can help with phytoremediation (Wang *et al.*, 2021). Second-generation biofuels are not without their environmental impacts—biodiversity reduction through monoculture, alongside large-water requirements for biomass growth and conversion to biofuels remain challenges (Evans, Kelley and Potts, 2015). To move towards an economy where the more resource efficient second-generation biofuels are used instead of their first-generation counterparts, the lignocellulolytic enzyme mixes being used need to become cheap and industrially scalable (Himmel *et al.*, 2007; Mueller, Anderson and Wallington, 2011).

Second generation biofuels rely on cellulolytic enzymes to release fermentable sugars from lignocellulosic biomass. These sugars are then converted into ethanol through fermentation. Because the carbon released by burning biofuel was first absorbed by plants, bioethanol reduces net carbon emissions significantly, relative to carbon emissions from an equivalent energy output from fossil fuels. The carbon emissions from second generation biofuels are 44 – 95% less than from gasoline (Wang, Littlewood and Murphy, 2013). Widespread usage of second-generation

biofuels, instead of gasoline, could therefore help reach COP21 targets, whilst providing energy security (Wang, Littlewood and Murphy, 2013). Globally, bioethanol supplies 3.1 exajoules (EJ) per annum (Lynd *et al.*, 2017a); first generation biofuels, produced from starch- and sugar-rich food crops give the majority of this, whilst second generation (cellulosic) biofuels are produced commercially only on a small scale (Zhang and Bao, 2017).

Production of cellulosic biofuel is a four-step process: (1) pre-treatment of the lignocellulosic material with pressurized steam and acids to depolymerise the hemicellulose, (2) hydrolysis of cellulose chains by a cocktail of enzymes releasing soluble sugars, (3) fermentation of glucose and xylose into ethanol by yeast strains, (4) distillation of the ethanol, and power production from burning the lignin-based residue (Johansen, 2016).

Major barriers to widespread, industrial-scale, conversion of lignocellulosic material to biofuel include the cost of cellulases, the relatively slow conversion of cellulose to glucose by cellulase systems, the low yield of sugars from the hemicellulosic portion of plant cell walls, and the recalcitrance of lignin and its inhibitory effect on cellulase systems (Himmel *et al.*, 2007; Olofsson *et al.*, 2017).

The current cost of cellulases is prohibitory for market-scale production of cellulosic biofuel; solutions to this costliness include (1) engineering or strain-selection for cellulases with faster kinetics, (2) cheaper production of cellulases, *via* strain selection for microbes which overexpress the gene of interest, and (3) restructuring of the production process by integrating the hydrolysis and fermentation steps in the same facility—this can be achieved by inoculating the input material with *Trichoderma reesii*, negating the need to purchase expensive enzyme mixtures (Li *et al.*, 2012; Olofsson *et al.*, 2017). Production strategies for biofuel from lignocellulose fail to make use of all substances which can be sold for biotechnology, and require transport of material between processing steps; streamlining of the industrial process could make the cellulosic biofuels much more economically profitable (Huang *et al.*, 2018). In total, these costs and inefficiencies make lignocellulosic ethanol two to three times more expensive than gasoline for an equivalent amount of energy from combustion (Carriquiry, Du and Timilsina, 2011), meaning cellulosic biofuel is produced commercially only on a small scale (Zhang and Bao, 2017).

The cost of producing second-generation biofuel is decreasing as companies invest in improving these inefficiencies. Novozymes have produced an enzyme mixture which is reported to require an

input of enzymes five times smaller than competitor enzyme mixes (Li *et al.*, 2012). This innovation could reduce the cost of cellulosic biofuel to USD 2.00 per gallon, a price which is competitive against both corn-derived ethanol and gasoline (Li *et al.*, 2012).

Lytic polysaccharide monooxygenases (LPMOs) are a recently discovered enzyme class which oxidatively catalyse the breakdown of cellulose. Since the discovery of their function (Vaaje-Kolstad *et al.*, 2010), they have been widely adopted for the production of second-generation biofuel (Johansen, 2016). Traditionally, saccharification and fermentation of the cellulosic material was simultaneous, however, because LPMOs require molecular oxygen, which is removed during fermentation, biofuel refineries now separate these two processes (Johansen, 2016). These large-scale changes to the methods employed for production of lignocellulosic biofuels, in response to the increases in yield that LPMOs produce, is indicative of the industry's ability to quickly respond to novel research (Vaaje-Kolstad *et al.*, 2010; Lo Leggio *et al.*, 2015; Johansen, 2016). Discovery of novel lignocellulolytic enzyme classes, for example from soil microorganisms, can therefore lead to increased efficiency and profit margins for biofuel production, and could lead to significant advances to the sustainability of this fuel generation process. Cultivation methods which are targeted at the discovery of novel species from diverse and abundant lineages should be a research priority for industries which rely upon these enzymes. High-throughput *in situ* cultivation represents a cultivation method which is appropriate for this task, whilst metagenomics represents an accurate screening method for the identification of environments which contain novel lignocellulase genes for industrial exploitation.

1.2 The barrier of microbial culturability on increasing knowledge of lignocellulose breakdown

One of the major barriers to a deeper understanding of how microbial communities degrade lignocellulose, is the intractability of most microorganisms to isolation and cultivation in the laboratory (Jannasch and Jones, 1959; Tabor and Neihof, 1984; Vartoukian, Palmer and Wade, 2010). The estimated number of microbial species globally reaches into the trillions (Locey and Lennon, 2016). Generally, fewer than 5% of cells isolated from soil are cultivable *via* traditional methods (Sait, Hugenholtz and Janssen, 2002). Despite this vast diversity and significant historical and current global effort to cultivate microbiota from across the tree of life, only 31% of the 181 identified bacterial phyla have cultivated representatives, with the majority of these belonging to *Proteobacteria*, *Actinobacteria*, *Bacteroidota*, and *Bacillota*, according to GTDB R214 (Vartoukian,

Palmer and Wade, 2010; Tanaka *et al.*, 2017; Parks *et al.*, 2021). Archaea remain vastly under-represented in terms of their isolation and cultivation, with only 648 species from 10 phyla having been isolated and cultured in the lab (Parks *et al.*, 2021, GTDB R214). Clearly, application of an array of techniques to improve the rate of cultivation of novel lineages from diverse branches of the tree of life will improve our understanding of ecosystem function as we begin to understand how abundant but uncultivated microorganisms interact with their environments.

Whilst microorganisms grow readily in their natural habitat, very few soil bacteria (0.3% of all cells) grow on the media which are commonly used in microbiological studies (Amann, Ludwig and Schleifer, 1995). Species which do grow readily *in vitro* are often rare, fast growing, play unclear roles within the microbial community, or may originate from dormant cells or inactive spores that are more suited to growth in lab-conditions than those experienced at the site of sampling (Jung, Aoi and Epstein, 2016).

1.3 What is lignocellulose and how is it degraded?

1.3.1 What is lignocellulose, and why is it difficult to degrade?

Lignocellulose is a composite material, primarily consisting of cellulose, hemicelluloses and lignins. It also contains pectins, inorganic compounds, proteins, waxes and lipids in smaller quantities. The relative proportions of these components depend upon the plant species and tissue in question (Brandt *et al.*, 2013).

Cellulose, the major component of lignocellulose, is an insoluble linear polysaccharide consisting of chains of up to ~5,000 repeating β-1,4-linked cellobiose units (Brandt *et al.*, 2013; Pu *et al.*, 2013). Cellobiose consists of two β-D-glucose molecules, which are rotated 180° relative to one another, and which are bound *via* a β-1,4 glycosidic bond (Leschine, 1995). Cellulose chains form microfibrils, which are clusters of parallel cellulose molecules, held together *via* hydrogen bonds and van der Waals forces; these give significant strength to the microfibril (Leschine, 1995; Pu *et al.*, 2013). These clusters of cellulose chains reduce the surface area at which enzymatic breakdown of each cellulose chain can occur. Cellulose microfibrils have crystalline and amorphous regions, with the compound being typically 90% crystalline (Leschine, 1995). Amorphous cellulose regions are degraded by microbes much more easily than are crystalline

regions, as cellulolytic enzymes can gain access to individual cellulose chains (Bornscheuer, Buchholz and Seibel, 2014).

Hemicelluloses are polysaccharide chains formed from 100 – 200 pentose and hexose monosaccharides; they can be either branching or non-branching (Brandt *et al.*, 2013; Pu *et al.*, 2013). In lignocellulose, hemicelluloses act as an amorphous matrix material, which surrounds and connects with cellulose microfibrils and other hemicellulose chains (Brandt *et al.*, 2013). In addition to providing structural integrity, hemicellulose acts as a physical barrier against cellulases, preventing degradation of cellulose microfibrils (Leschine, 1995; Álvarez, Reyes-Sosa and Díez, 2016). There is evidence that hemicellulose may pose a major barrier to lignocellulose degradation by many microbes; in a study by (Vargas-García *et al.*, 2007), only two of the 13 isolated strains from a compost-heap caused deterioration of the hemicellulosic material which was given as a growth substrate.

The third major component of lignocellulose, lignin, is a highly polymerized three-dimensional polyphenolic compound, formed primarily from ether-linked alcohols (Bourbonnais and Paice, 1990; Brandt *et al.*, 2013). These form both phenolic (10-20% of the structure), and non-phenolic (80-90% of the structure) regions within the lignin molecule (Dashtban *et al.*, 2010). Lignin is the component of lignocellulose which lends the largest degree of recalcitrance to microbial degradation, because of its numerous strong ether and carbon cross-linkages; it also provides resistance to pathogens, waterproofing, and structural integrity to plant cell walls (Pu *et al.*, 2013; de Gonzalo *et al.*, 2016). As with hemicellulose, lignin provides an additional physical barrier which prevents polysaccharide hydrolases from reaching their substrates. Further, lignin acts as a non-target substrate for cellulases, significantly reducing the rate at which they degrade cellulose (Pan *et al.*, 2005; Vermaas *et al.*, 2015). Cellulase enzyme activity on cellulose is made less efficient by ether bonds between lignin and the hydrophobic faces of cellulose microfibrils, to which cellulases preferentially bind (Vermaas *et al.*, 2015). Ester bonds between hemicellulose and lignin provide further structural integrity to the cell wall (Jin *et al.*, 2006), and increase resistance to enzymatic breakdown of lignocellulose by making access to cellulose microfibrils more difficult (Zhao, Zhang and Liu, 2012; Vermaas *et al.*, 2015).

1.3.2 Enzymatic mechanisms of lignocellulose deconstruction

1.3.2.1 Polysaccharide degradation

1.3.2.1.1 Glycoside Hydrolases

Degradation of cellulose and hemicellulose by microorganisms is mainly a hydrolytic process performed by glycoside hydrolases, although there is some degradation of these by oxidative enzymes. The full suite of cellulases degrades cellulose from both ends of the molecule (exoglucanase activity) as well as from within the cellulose chains (endoglucanase activity; Figure 2). Typically, a cellulose molecule is cleaved by endoglucanases at amorphous sites within the chain (quickly decreasing the crystallinity), whilst concurrently having cellodextrins 'trimmed' off the ends by exoglucanases (including cellobiohydrolase—CBH—which produces cellobiose as its major product). These two types of degradative activity on the cellulose molecule act synergistically, as the endoglucanases increase the number of binding sites for exoglucanases and CBH (Lynd *et al.*, 2002). The activities of CBH and endoglucanases are inhibited by cellobiose, and so for an organism to effectively degrade cellulose, they must also produce β-glucosidases which hydrolyse cellobiose to two glucose molecules (Figure 2) (Lynd *et al.*, 2002). Glucose may inhibit β-glucosidases (Figure 2), but the extent to which this is limiting for an organism's growth on cellulose depends on the β-glucosidase in question (Lynd *et al.*, 2002; Bornscheuer, Buchholz and Seibel, 2014).

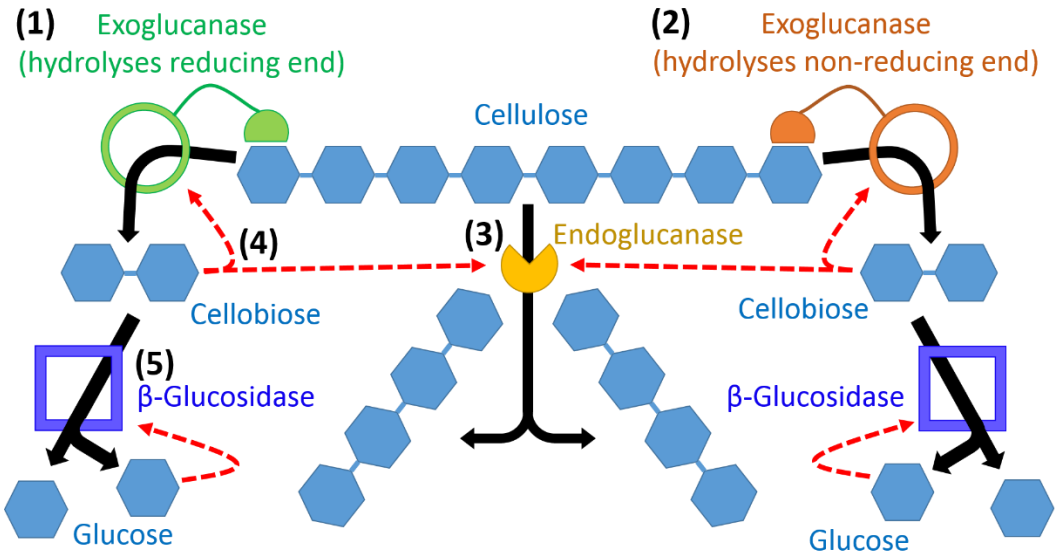

Figure 2: Overview of a typical hydrolytic free-enzyme cellulase system. Solid black arrows show enzymatic hydrolysis, red dashed arrows show enzymatic inhibition by hydrolysis products. Specific exoglucanases bind to the (1) reducing end (e.g., CBHI) and (2) to the non-reducing (e.g. CBHII) end of a cellulose chain via *their carbohydrate binding modules, where they hydrolytically cleave off cellodextrins or cellobiose units. (3) Endoglucanases hydrolyse cellulose chains within the chain (in amorphous regions) to produce*

shorter cellulose chains. (4) Cellobiose inhibits both exo- and endo-glucanase activity. (5) β-glucosidase breaks cellobiose down to glucose, although this process is self-inhibiting.

Exoglucanases typically have low substrate affinities, and thus require a carbohydrate-binding module (CBM) to localize them next to the cellulose chain to allow for maximum substrate utilisation (Lynd *et al.*, 2002). CBMs can have a preference for binding to the hydrophobic crystalline surface of a cellulose microfibril, or to amorphous regions (Boraston *et al.*, 2004). Exoglucanases, including CBH, have a preference for binding either the reducing end (*e.g.,* CBHI), or the nonreducing end (*e.g.,* CBHII), of a crystalline cellulose structure (Bornscheuer, Buchholz and Seibel, 2014); these two types of exoglucanases act synergistically to degrade cellulose crystals (Lynd *et al.*, 2002). This exo-exo synergism, as well as the exo-endo synergism between endoglucanases and exoglucanases has been demonstrated with the enzymes CBHI, CBHII, and Endoglucanse V (Lynd *et al.*, 2002). Combination of all three enzymes produced rates of crystalline cellulose degradation five times faster than the fastest single enzyme, and rates of degradation 1.5 times as fast as combination of any two of the enzymes (Lynd *et al.*, 2002). It should be noted that many cellulolytic enzymes have multiple catalytic domains which have several different functions, including CBM domains, hemicellulolytic domains, esterase domains, or polysaccharide lyase domains (Rosenberg *et al.*, 2013).

Hemicellulose is formed from combinations of many different pentose and hexose sugars; meaning a diverse suite of saccharolytic enzymes are therefore required to break each bond type between monomers. Breakdown of hemicellulose by fungi involves cleavage of bonds by endohydrolases in hemicellulose's polysaccharide backbone, releasing oligo- or mono-saccharides. Many types of enzyme then degrade cleaved oligosaccharides to their constituent monomers, with the actions of several of these having synergistic effect on the rate of hemicellulose degradation (Van Den Brink and De Vries, 2011). White- and brown-rot fungi both produce a suite of hemicellulose degrading endo- and exo-hydrolases, although the number of these in brown-rot fungi is smaller than in white-rot fungi (Valášková and Baldrian, 2006; Hori *et al.*, 2013). Bacterial hemicellulose degradation within xylanase rich families is not strongly constrained within phylogenetically restricted enzyme families—showing the functional diversity of microbial hemicellulose degradation within closely related species (Yeager *et al.*, 2017).

1.3.2.1.1.1 Lytic polysaccharide monooxygenasess

Lytic polysaccharide monooxygenases (LPMOs) are a diverse enzyme class which use an oxidative mechanism to break down crystalline cellulose, hemicellulose, or other non-lignocellulosic substrates such as chitin or peptidoglycan (Agger *et al.*, 2014; Morgenstern, Powlowski and Tsang, 2014; Hemsworth *et al.*, 2015). By oxidising crystalline cellulose, LPMOs create more sites at which processive exoglucanases can act, and can double the rate at which CBHI releases soluble sugars from crystalline cellulose (Eibinger *et al.*, 2014). They are common throughout fungal and bacterial cellulose degraders; analyses of fungal genomes suggest that genes for LPMOs are far more widespread than are genes for cellobiose dehydrogenases (Kracher *et al.*, 2016), and that some species have genes for over 40 different LPMOs (Agger *et al.*, 2014). The oxidative mechanism of action of LPMOs propagates unstable compounds which cause further oxidative cleavage (Hemsworth *et al.*, 2015; Bertini *et al.*, 2018). Different LPMOs oxidise either the C1 or C4 carbon of a glucose residue within cellulose—giving differing final products, these LPMO types act synergistically to degrade crystalline cellulose (Hemsworth *et al.*, 2015). LPMOs require molecular oxygen or hydrogen peroxide and an electron donor to degrade plant cell wall polymers (Hemsworth *et al.*, 2015; Bissaro *et al.*, 2017). There are a wide range of electron donors including cellobiose dehydrogenase, lignin, several acids found in plant material, bi-phenolics, and tri-phenolics, meaning they have great functional flexibility (Westereng *et al.*, 2015; Kracher *et al.*, 2016).

In addition to degradation of cellulose, LPMOs play an important role in deconstructing the widely abundant biopolymers hemicellulose, chitin and starch, with each LPMO appearing to act on an individual substrate (Agger *et al.*, 2014; Kracher *et al.*, 2016). It is unlikely that LPMOs degrade lignin directly (Agger *et al.*, 2014), although lignin depolymerisation is possible in combination with mediator chemicals and other enzymes through promotion of Fenton chemistry (Li, Zhang, *et al.*, 2021). Because of the improvement to the rate of solubilization of lignocellulose, LPMOs have been widely adopted for the production of biofuels, and likely play important roles in ecosystems where they release sugars to the organisms which secrete them (Müller *et al.*, 2015; Kracher *et al.*, 2016; Bissaro *et al.*, 2017; Bertini *et al.*, 2018).

1.3.2.1.1.2 Amorphogenesis inducing proteins

In addition to the hydrolytic and oxidative actions of glycoside hydrolases and LPMOs, the structure of cellulose is disrupted by amorphogenesis-inducing proteins (Din *et al.*, 1991; Gourlay

et al., 2015); these include bacterial expansin-like proteins, fungal swollenin and loosenin proteins, as well as some bacterial and fungal CBMs (Gourlay *et al.*, 2015). These proteins cause physical disruption to the cellulose microfibril through penetration and possibly expansion—weakening hydrogen bonds between cellulose chains (Din *et al.*, 1991; Gourlay *et al.*, 2015). Scanning electron micrographs of cellulose incubated with CBMs (Din *et al.*, 1991) or swollenins (Jäger *et al.*, 2011) show clear increases in the proportion of amorphous cellulose in their structures. In contrast to the CBM, incubation of cellulose with a hydrolytic enzyme which was not associated with its CBM, leads to a "polishing" of the exterior surface of the crystalline microfibril. This shows that the CBM was the primary factor leading to structural changes in the fibre (Din *et al.*, 1991). By disrupting the structure of the microfibril, amorphogenesis proteins give a much larger surface area which is accessible for hydrolysis by cellulases—increasing the overall rate of microfibril degradation by hydrolytic enzymes (Din *et al.*, 1991; Jäger *et al.*, 2011). In the case of CBMs, the effect on rate of cellulose degradation may be increased by the associated increase in hydrolytic enzyme concentration near the cellulose.

1.3.2.2 Lignin Degradation

Enzymes for degradation of lignin, from white-rot fungi, belong to the heme peroxidase and laccase classes, both of which oxidise the phenolic ring structures and non-phenolic structures within lignin (Lambertz *et al.*, 2016). These enzymes are secreted into the environment where they degrade wood exocellularly.

1.3.2.2.1 Heme peroxidases

Ligninolytic heme peroxidase are from two main evolutionary lineages: the first contains lignin-, manganese- and versatile-peroxidases (LiP, MnP, VP respectively), and the second contains the dye-decolorizing peroxidases (DyPs), which are more often associated with bacteria (Cragg *et al.*, 2015). Heme peroxidases catalyse degradation of lignin by creation of lipid peroxyl free radicals ("*Substrate*$^{\bullet}$" in the reaction in Box 1), either *via* reactions with hydrogen peroxide, or with metal ions (such as Mn^{3+}) present in the plant-tissue (Dashtban *et al.*, 2010). These free radicals oxidize phenolic and non-phenolic bonds in lignin (although lignin peroxidase mainly oxidises non-phenolic lignin units), creating more free radicals in the process; these continue to degrade the lignin structure (Kapich *et al.*, 2005; Martínez *et al.*, 2005).

Degradation of lignin by heme peroxidases starts with elevation of the iron-containing active site from its resting state by H_2O_2 (or another electron donor), resulting in a strongly reduced enzyme

active site (Fe^{4+}=O·poryphyrin$^{\bullet+}$; Box 1). Free radicals are created within the lignin molecule through oxidation by this ferric enzyme, mediating the degradation of the lignin structure. (Hofrichter, 2002; Castro *et al.*, 2016). Fungi which produce heme peroxidases give rise to white-rot symptoms as they fully degrade the lignin, leaving fibrous white cellulose and hemicellulose intact (Hofrichter, 2002).

Genes for dye-decolorizing peroxidases are most frequently found in bacterial genomes, but are also occasionally found in fungal genomes (Colpa, Fraaije and Van Bloois, 2014; Lambertz *et al.*, 2016). These enzymes may play a significant role in bacterial lignocellulose degradation in soils (Furukawa, Bello and Horsfall, 2014), yet we currently know little about the functional diversity of

Mechanisms for lignin decomposition

(a) Heme peroxidase catalytic cycle

Fe^{3+}·poryphyrin is the enzyme active site in its resting state

1) Fe^{3+}·poryphyrin + H_2O_2 → Fe^{4+}=O·poryphyrin$^{\bullet+}$ + H_2O
2) Fe^{4+}=O·poryphyrin$^{\bullet+}$ + Substrate → Fe^{4+}=O·poryphyrin + Substrate$^{\bullet}$
3) Fe^{4+}=O·poryphyrin + Substrate → Fe^{3+}·poryphyrin + Substrate$^{\bullet}$ + H_2O

(b) Laccase catalytic cycle

4 Cu+ is the enzyme active site in its resting state

1) 4 Cu^{+} + O_2 + 4 H^{+} → 4 Cu^{2+} + 2 H_2O
2) 4 Cu^{2+} + 4 Substrate → 4 Cu^{+} + 4 Substrate$^{\bullet}$

(c) Fenton Chemistry

Fe^{3+} + H_2O_2 → Fe^{2+} + $^{\bullet}$OH + $^{-}$OH

DyPs (Min *et al.*, 2015; Catucci *et al.*, 2020).

Box 1: Major pathways of degradation of lignin by (a) heme peroxidase enzymes, (b) laccase enzymes, (c) nonenzymatic Fenton chemistry. Free radicals which are formed in the above equations start free radical chain reactions within lignin molecules, and in other lignocellulosic components (Kapich et al., 2005; Riva, 2006; Dashtban et al., 2010). (a) Generalized catalytic cycle of heme peroxidases, as found in white-rot fungi (Martînez, 2002). (b) Catalytic cycle of laccase enzymes found in white-rot fungi (Riva, 2006; Dashtban et al., 2010; Furukawa, Bello and Horsfall, 2014). (c) Nonenzymatic Fenton chemistry which gives rise to free radicals within lignin structures (Cragg et al., 2015). "˙" represents a free radical.

1.3.2.2.2 Laccases

Laccases, or phenol oxidases, degrade lignin using an active site which contains four copper ions. These reduced copper ions (Cu^{+}) donate electrons to molecular oxygen, producing water. This reaction involves production of an intermediate oxygen free radical, and oxidizes Cu^{+} to Cu^{2+} (Box 1). The subsequent reduction of the Cu^{2+} ions back to Cu^{+} leads to the oxidation of a variety of substrates, including lignin (Dashtban *et al.*, 2010; Furukawa, Bello and Horsfall, 2014). Lignin's phenolic rings are oxidised directly by the copper ions at the active site, whereas non-phenolic lignin structures are oxidised indirectly *via* a redox mediator such as phenols, syringaldehyde and aniline, amongst others (Bourbonnais and Paice, 1990; Furukawa, Bello and Horsfall, 2014; Lambertz *et al.*, 2016).

Whilst mainly produced by white-rot fungi, laccase enzymes have been found to be produced by some brown-rot fungi, as well as by some bacteria (Dashtban *et al.*, 2010). Bioinformatic searches of draft genomes and assembled metagenomes have found over 1200 putative laccase genes (including several which are putatively horizontally transferred) from bacteria with highly different life strategies (anaerobes, thermophiles, autotrophs, alkaliphiles), suggesting that lignin-degrading capabilities are widespread across the bacterial tree of life (Ausec *et al.*, 2011).

In tandem with laccase enzymes, a complementary suite of enzymes is produced by many microorganisms; these provide hydrogen peroxide which reduces the products of oxidative cleavage. Cellobiose dehydrogenase plays a role in both of these functions, catalysing reduction of phenoxy radicals produced by laccases (stopping the lignin from repolymerising around the newly exposed cellulose), as well as providing H_2O_2 and Fe^{2+} to oxidation of many lignin-degrading enzymes, enabling non-enzymatic lignin breakdown pathways (Ludwig *et al.*, 2010).

1.3.2.2.3 Fenton Chemistry

The breakdown of lignocellulose through Fenton chemistry relies upon the oxidation of Fe^{2+}, which is present in all woody material, to Fe^{3+} by H_2O_2, producing a hydroxyl free radical (Cragg *et al.*, 2015). These free radicals penetrate the lignocellulosic biomass, reaching the internal structures which are unavailable for enzymatic breakdown (due to the large molecular mass of lignocellulolytic enzymes), cleaving bonds in cellulose, hemicellulose and lignin (Arantes *et al.*, 2011). In lignin, the highly reactive ·OH free radicals lead to significant demethylation and side chain oxidation, depolymerisation (cleavage of nonphenolic aryl methyl ether bonds), as well as some breakdown of aromatic structures (Arantes *et al.*, 2011; Arantes and Goodell, 2014). The

Fenton reaction does not truly degrade lignin, but instead rearranges it, creating many reactive intermediates in the process which may play a role in cellulose and hemicellulose decomposition (Arantes *et al.*, 2011; Arantes and Goodell, 2014).

In contrast with most white-rot fungi, brown-rot fungi do not produce a large suite of enzymes involved in lignocellulose decomposition. Instead, brown-rot fungi use low-molecular weight free radicals, generated by hydrogen peroxidases in the presence of iron, to oxidatively modify lignin, and to break down crystalline cellulose (Hori *et al.*, 2013; Arantes and Goodell, 2014). The hydroxyl free radicals (·OH) which are a product of the Fenton reaction are the most powerful non-specific oxidative agent produced by living organisms (Arantes and Goodell, 2014). It has recently been shown that some LPMOs and dehydrogenases are involved in supporting the Fenton reaction on lignin substrates (Li, Zhang, *et al.*, 2021; Li, Zhao, *et al.*, 2021). This greatly expands the diversity of microorganisms known to have ligninolytic potential.

1.3.2.3 Organisms and strategies involved in lignocellulose degradation

The majority of our current understanding of the organisms and mechanisms of lignocellulose degradation comes from culture-dependent techniques. Isolation of microorganisms, biochemical enzyme characterisation and annotation of genes that allow them to degrade complex carbon sources allows for identification of similar genes present in environmental samples. Knowledge of the genes involved in lignocellulose breakdown present in a community allows us to characterise the lignocellulolytic potential of the community. The total lignocellulolytic capacity of a microbial community is useful information for improving ESMs, however, for this approach to work, we must first have a basic knowledge and understanding of which genes produce lignocellulolytic phenotypes (Tian *et al.*, 2014). Glycoside hydrolases in microbial genomes are ubiquitous as they encode enzymes which are crucial for remodelling, maintenance, production, and deconstruction of the bacterial cell wall (Vermassen *et al*., 2019). While use of glycoside hydrolase families is by no means a perfect method for prediction of substrate type, with families often containing broad substrate specificites, analysis of particular subfamilies with well characterised substrate specificities (*e.g.,* GH5 subfamily 1 (GH5_1) as endoglucanases, GH5_5 as endo-β-1,4-glucanases, (Aspeborg *et al*., 2012)) can provide some level of insight into the genetic potential for broadly distributed biogeochemical processes such as lignocellulose degradation. A recently developed bioinformatics pipeline, dbCAN3 uses carbohydrate active enzyme (CAZy) subfamily, and gene

cluster composition overall predicted substrate specificities of each gene to predict the function of genes and gene clusters (Zheng *et al*., 2023).

The composition of active lignocellulose degraders in soil depends on the oxygen availability in their environment (Schellenberger, Kolb and Drake, 2010). Bacteria in aerobic systems tend to produce extracellular free enzymes, whereas membrane-bound extracellular enzymes and multi-enzyme cell-associated cellulosomes are typically employed in anoxic environments (Lynd *et al.*, 2002). Cellulosomes are attached to the cell wall *via* scaffoldin proteins (Arntzen *et al.*, 2017). This is an oversimplification of the true complexity and diversity of the cellulose-degrading mechanisms utilised by fungi, protists, bacteria, and archaea, which is biased by our current knowledge of CAZyme systems from cultivated microorganisms. Recent studies have elucidated alternative modes of cellulose utilisation, including polysaccharide-utilisation loci (PUL), gliding bacteria with cell membrane bound individual enzymes, and vesicle-bound multi-CAZyme complexes (Arntzen *et al.*, 2017). Further cultivation of novel isolates from diverse phyla, combined with *in vitro* tests and functional annotation will give a more representative overview of the different methods of microbial lignocellulose degradation.

The advent of high-throughput DNA and RNA sequencing, and comparative metagenomics, have transformed our understanding of the diversity of genes and organisms involved in lignocellulose degradation. These studies have investigated degradation of lignocellulose by microbes in a variety of habitats, including aerobic and anaerobic soils and sediments at a variety of latitudes, and under a variety of extreme conditions (Kanokratana *et al.*, 2011; Xia *et al.*, 2013; Tveit, Urich and Svenning, 2014; Jiménez, Chaves-Moreno and Van Elsas, 2015; Yeager *et al.*, 2017). Such studies, whilst useful, suffer from the CAZymes identified having multiple activities within CAZy family, with some CAZymes within the same family synthesising a substrate, and others degrading it. This makes interpretation of activities from genetic data difficult. Other environments explored through metagenomics for lignocellulase genes include compost (Martins *et al.*, 2013; Simmons *et al.*, 2014; Wang *et al.*, 2016), and the digestive tracts of vertebrates and invertebrates (Brulc *et al.*, 2009; Nimchua *et al.*, 2012; Lopes *et al.*, 2015).

The CAZy database (CAZy.org) is a valuable resource of enzymatic and genomic information, taken from organisms with the capability to degrade carbohydrates (Lombard *et al.*, 2014). To understand which organisms have been most thoroughly researched in terms of CAZyme

production, I took the names of the publicly available bacterial and archaeal genomes (2018-05-16 and 2018-05-24, respectively) on CAZy, and assigned the species to phyla and families (Chamberlain and Szöcs, 2013). The phyla with the highest number of genomes on the CAZy database were *Proteobacteria* (5589 genomes), *Bacillota* (2112 genomes), *Actinobacteria* (1038 genomes), and *Bacteroidota* (352 genomes) (Figure 5a), reflecting the number of genomes in these groups on GenBank. *Enterobacteriaceae* (*Gammaproteobacteria*) was the family with the most genomes (1847) available on the CAZy database (Figure 5b). Of the genomes in *Enterobacteriaceae*, 1470 (79.5%) were from the medically important species *Klebsiella pneumonia*, *Escherichia coli*, and *Salmonella enterica*. It is unlikely that these species are the dominant degraders of lignocellulose in soils due to the few CAZyme genes that their genomes contain and the multiple alternative functions of cellulases (Medie *et al.*, 2012), especially considering that they are not typically dominant soil bacteria. Whilst the collection of genomes on the CAZy database will be biased by species which are medically or are biotechnologically important, this analysis shows the depth of study into each of these groups. Other bacterial families with many genomes available on the CAZy database include *Bacillaceae* (480 genomes), *Streptococcaceae* (393 genomes), *Pseudomonadaceae* (390 genomes), *Staphylococcaceae* (316 genomes), *Burkholderiaceae* (274 genomes), *Lactobacillaceae* (252), *Mycobacteriaceae* (235), *Campylobacteraceae* (215), *Corynebacteriaceae* (181), and *Flavobacteriaceae* (180).

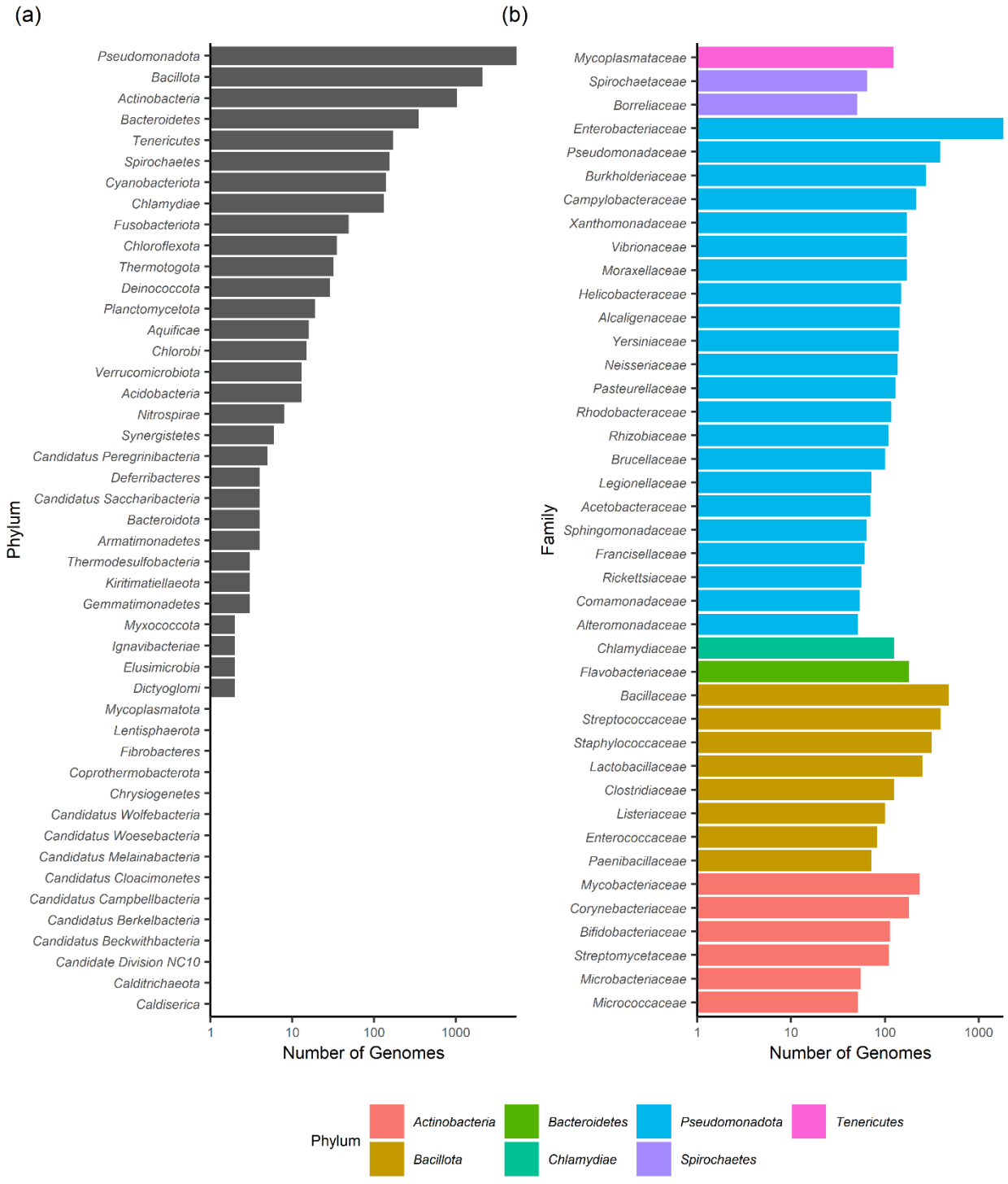

Figure 3: Number of bacterial genomes publicly available on the CAZy database (www.CAZy.org) in different (a) phyla, and (b) families for which there were more than 50 genomes available.

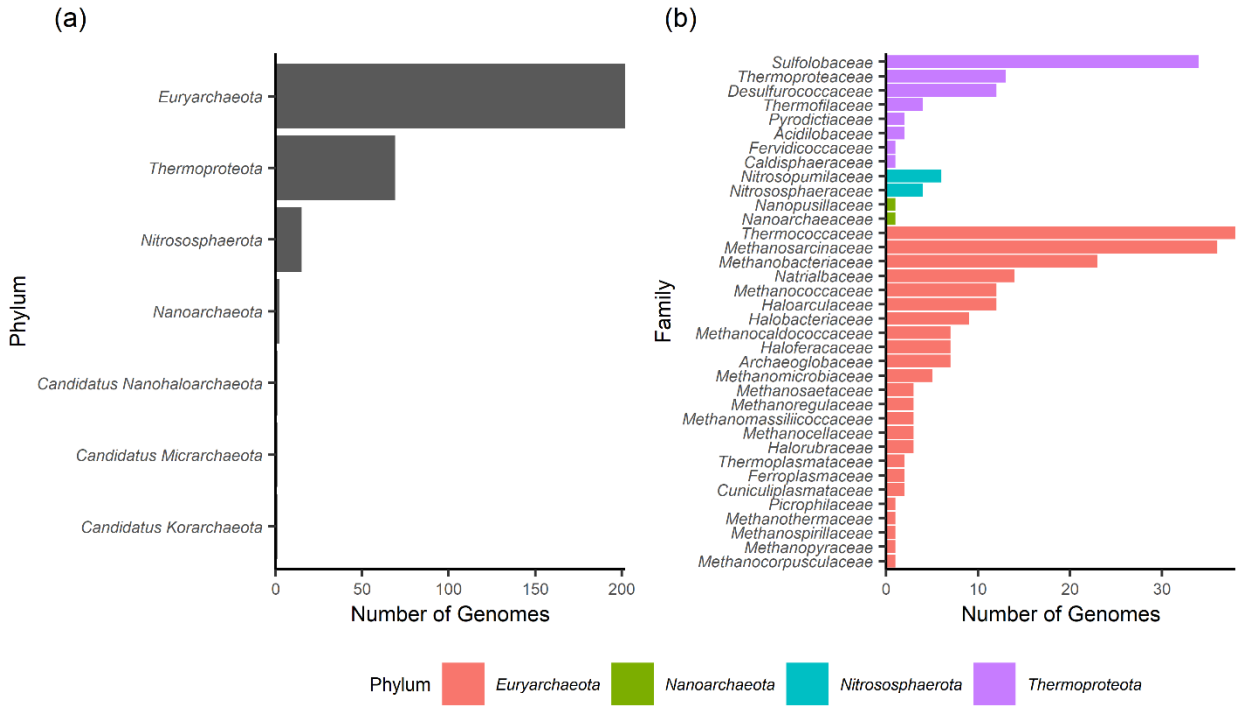

Figure 4: Number of archaeal genomes publicly available on the CAZy database (www.CAZy.org) in different (a) phyla, and (b) families for which there were more than 50 genomes available.

Whilst about 40% of bacterial genomes found on the CAZy database (a repository of carbohydrate active DNA sequences) contain at least one cellulase gene (multiple families and activities), bacteria with more than three cellulase genes (as the authors to note that cooccurrence of functionally related genes suggests relatedness to life history traits—in this case cellulolytic activity) make up fewer than 15% of species in all phyla (with the exception of *Actinobacteria*, *Thermotogota*, *Bacteriodetes*, and Class *Clostridia*) (Medie *et al.*, 2012). The majority of bacteria with cellulase genes are therefore likely ineffective at degrading crystalline cellulose (Koeck *et al.*, 2014), suggesting that cellulolysis is not a major part of their life-history strategy, and that these species will likely have little impact on the rate of conversion of SOM into DOM in soils—if they are even present, abundant, or active. This information does not give insights about the rate of transcription or translation of these genes, or activity of the expressed proteins however, meaning that knowledge about these species is limited only to potential activity. Medie *et al.* (2012) identified that *Actinobacteria* was the bacterial phylum with most hydrolytic species, having more than three cellulase genes (31% of genomes analysed). Other taxa which contained many species more than three cellulase genes were *Clostridia* (25% of genomes), *Bacteroidota* (22% of genomes), *Thermotogota* (18% of genomes), and *Chloroflexota* (14% of genomes). Additionally,

2% of *Betaproteobacteria*, 5% of *Alphaproteobacteria*, and 7% of *Gammaproteobacteria* genomes had more than three cellulase genes (Medie *et al.*, 2012). These percentages should be considered with the context of the disproportionate number of but medically relevant genomes available. Analyses of publicly available genomes are inherently biased by predominant foci of research activity and ease of cultivability of each group (as only complete genomes are assessed by CAZy), with results giving a skewed perspective of the importance of certain taxa if questions about diversity are being addressed. The global focus on genes from cultivable and medically relevant microorganisms provides a starting point for understanding microbially driven processes in complex environmental ecosystems. To accurately predict changes to ecosystem functioning (and thus ecosystem services) with global change, and to provide step changes in the efficiency of lignocellulase-mediated industrial reactions, research must focus on obtaining undiscovered enzymes and the organisms that produce them. Knowledge of which community members are the most important for the degradation of lignocellulose will be an important facet in this. Furthering knowledge about the diversity of enzymatic mechanisms of bacteria for lignocellulose degradation will stimulate research into novel, potentially biotechnologically important, cellulase systems. Metagenomics and culture-based studies are now beginning to explore the diversity of other bacterial lignocellulose degraders (Yeager *et al.*, 2017; López-Mondéjar, Algora and Baldrian, 2019). This again points to the need to directly observe the phenotypes of novel individual species. Cultivation of microorganisms from soils represents a potential gold-mine for the discovery and *in vitro* characterisation of species and novel enzyme classes. Here we highlight well known lignocellulose degraders and knowledge about the enzyme systems from a range of taxa.

1.3.2.3.1 Fungal lignocellulose decomposition

The most well-studied lignocellulose degrading group are the white-rot fungi that have the ability to completely mineralize lignocellulose (Pandey and Pitman, 2003). White-rot fungi include the biotechnologically important species *Trichoderma reesei* (synonym *Hypocrea jecorina*), which has been extensively selected and engineered for cellulase production for a variety of industries (Lynd *et al.*, 2002).

White-rot fungi degrade lignocellulose *via* complex cocktails of extracellular enzymes, which act to disrupt the surface of the lignocellulose. For example, *Trichoderma reesei* produces many functionally redundant cellulases, producing two exoglucanases, five endoglucanases, two β-glucosidases, and three lytic polysaccharide monooxygenases (LPMOs) (Lynd *et al.*, 2002; Müller *et*

al., 2015). Some of these enzymes act synergistically, whilst others do not (Lynd *et al.*, 2002). Brown-rot fungi, in contrast, employ a non-enzymatic mechanism of lignocellulose degradation whereby small free radicals penetrate into the interior of the lignocellulosic material, modifying the lignin structures within, and allowing access to oxidised cellodextrins (Floudas *et al.*, 2012; Riley *et al.*, 2014). Brown-rot fungi are a polyphyletic group, which have evolved from white-rot ancestors multiple times (Floudas *et al.*, 2012), suggesting that the production of fewer CAZymes may be a beneficial overall life-history strategy for wood-rotting fungi.

Whilst it is generally accepted that ascomycete fungi are major degraders of lignocellulose under aerobic conditions, analysis of decomposing rice-straw metatranscriptomes identified only a minimal contribution of ascomycete transcripts to the number of glycoside hydrolases produced under thermophilic conditions. Further, they found that there was no fungal expression of glycoside hydrolases under mesophilic conditions (Simmons *et al.*, 2014). In contrast, a metatranscriptome study which identified CAZymes from woodland soils found the majority of CAZymes to be of fungal origin (Damon *et al.*, 2012). Such variation is likely the result of the different systems being studied, highlighting how variable microbial polysaccharide utilisation can be.

1.3.2.3.2 Bacterial lignocellulose decomposition

1.3.2.3.2.1 *Actinobacteria*

Actinobacteria are the third best represented bacterial phylum on the CAZy database, with 1038 genomes listed (Figure 5). Medie *et al.* (2012) found that 31% of genomes in *Actinobacteria* contain over three cellulase genes, 28% have genes for two or three cellulases, and 15% contain genes for one cellulase. Their analysis identified *Actinobacteria* as the phylum with the highest cellulolytic potential. However, the coding density of these cellulases may not be so high due to the large genome sizes of species in this group (7.7 – 9.7 Mbp; Ventura *et al.*, 2007)—suggesting that they may not be critical for the carbon acquisition strategies of many of these organisms. The enzyme systems of *Actinobacteria* tend to be extracellular (Rosenberg *et al.*, 2013). In decomposing rice-straw under thermophilic conditions, *Actinobacteria* have been shown to produce the majority of glycoside hydrolase transcripts, with the genus *Micromonospora* playing a particularly large role (Simmons *et al.*, 2014).

Within the order *Actinomycetales* (phylum *Actinobacteria*), the members of family *Micromonosporaceae* have the highest average number of genes for degradation of cellulose, chitin, xylan and pectin (Yeager *et al.*, 2017). An analysis of genomes from *Micromonosporaceae* found more than 10 cellulose degradation genes in every genome analysed, as well as more than 15 genes for the degradation of xylan (Yeager *et al.*, 2017). Whilst micromonosporacids had a higher number of CAZymes present in their genomes than do the well-known lignocellulolytic genera *Streptomyces* (*Streptomcyetaceae*) and *Saccothrix* (*Pseudonocardiaceae*), activity of micromonosporacid CAZymes was lower than for *Streptomyces* and *Saccothrix* strains (Yeager *et al.*, 2017). These results show the importance of relating data on enzymatic expression and activity to gene abundance when using genomic data to model processes at larger scales.

Recent genomic, metagenomic, and physiological studies have found species within the actinomycete family *Pseudonocardiaceae* to contain a high numbers of CAZyme genes (Anderson *et al.*, 2012; Berlemont *et al.*, 2014; Koeck *et al.*, 2014; Větrovský, Steffen and Baldrian, 2014; Yeager *et al.*, 2017), and these have been associated with high levels of cellulase activity (Koeck *et al.*, 2014).

As well as CAZymes for degradation of cellulose, hemicellulose, chitin and pectin, Actinobacteria produce strongly oxidative peroxidase and laccase enzymes which are involved in the degradation of lignin. *Streptomyces* is the best studied genus of *Actinobacteria*, and is an effective degrader of lignin (Tian *et al.*, 2014). *Thermobifda fusca* is a thermophilic actinomycete which produces thermo- and alkali- tolerant laccases (Chen *et al.*, 2013) which have industrial applications in biobleaching, bioremediation, and in the food industry (Sondhi *et al.*, 2014). There are several other cultivated genera of *Actinobacteria* (*Mycobacterium*, *Microbacterium*, *Micrococcus*) which have been identified as ligninolytic; these have diverse laccase and peroxidase enzymes (Tian *et al.*, 2014).

1.3.2.3.2.2 *Bacillota_A*

The recently defined phylum *Bacillota_A* (Parks *et al*., 2021) degrade cellulose in natural environments with the class *Clostridia* being particularly well noted for this (Wang *et al.*, 2011; Ransom-Jones *et al.*, 2017). An analysis of CAZymes in landfill leachate metagenomes identified *Firmicutes* (*Bacillota and Bacillota_A - H*) as the phylum with the second most CAZymes, after *Bacteroidota* (Ransom-Jones *et al.*, 2017). The *CelA* multi-enzyme of *Caldicellulosiruptor bescii*

(phylum *Bacillota_A*) has a distinct action on cellulose microfibrils from the free enzyme systems of other cellulases, degrading cavities in the microfibril rather than "polishing" the fibres as is the case with fungal CBHI and CBHII (Din *et al.*, 1991; Brunecky *et al.*, 2013).

One quarter of curated genomes on the CAZy database from class *Clostridia* have more than three cellulase genes (Medie *et al.*, 2012). Several species of *Clostridium* and *Ruminococcus* have a particularly high number of cellulase (20 – 32) and hemicellulase (12 – 30) genes (Medie *et al.*, 2012), which may account for their reputation as prolific degraders of lignocellulose.

Clostridium thermocellum is an obligately anaerobic microorganism which degrades cellulose using a cellulosome structure—a complex of CAZymes and carbohydrate binding modules, held together by a scaffoldin protein, and dockerin and cohesion modules. The scaffoldin protein is attached to the bacterial cell, keeping the cell in close contact with the released sugars (Zverlov and Schwarz, 2008). The cellulosome of *C. thermocellum* is regulated by a minimum of 79 genes (Hirano *et al.*, 2016). Two isolates of *C. thermocellum* (ATCC 27405 (CP000568) and DSM 1313 (CP002416)) both had 39 genes specifically for cellulases, and 19 and 20 hemicellulase genes, respectively (Medie *et al.*, 2012).

In addition to the cellulosome, there is evidence that *C. thermocellum* may produce synergistically-acting free-enzymes for the degradation of cellulose and hemicellulose (Berger *et al.*, 2007).

The anaerobic *Clostridia* species *Caldicellulosiruptor bescii* can convert untreated lignocellulose into ethanol, meaning that interest in this species for industry has spiked. It has over 50 genes for glycosyl hydrolases, including *CelA*, which functions effectively on highly-crystalline cellulose at high temperatures. *CelA* is a free-enzyme with multiple catalytic domains, and multiple CBHs, and it gives rates of cellulose degradation faster than commercial enzyme mixtures (Brunecky *et al.*, 2013; Kim *et al.*, 2018). There are no genes for extracellular β-glucosidases in *C. bescii*; this may result from a scarcity of ATP under anaerobic conditions, meaning that importation of many glucose molecules into the cell is energetically unfavourable. In contrast, importation of an oligosaccharide (which requires only a single ATP molecule), followed by internal cleavage by cellobiose/cellodextrin phosphorylases may be a more efficient strategy utilised by bacteria living under anaerobic conditions (Kim *et al.*, 2018).

Sequencing of DNA-SIP studies (where amendment of an environment with isotopically labelled substrates is followed by sequencing of the DNA which has incorporated the heavy isotopes)

revealed that *Clostridium spp.* from aerobic soils were more closely associated with communities enriched with ^{13}C-glucose than they were with communities enriched with ^{13}C-cellulose (Pinnell *et al.*, 2014). This may suggest that, at least in aerobic soils, *Clostridium* species may play only a minor role in cellulose degradation.

1.3.2.3.2.3 *Bacillota*

Bacillota can be either aerobic or facultatively anaerobic, and have lignocellulose degrading systems which are expressed both extracellularly, and on the cell-surface (Jones, van Dyk and Pletschke, 2012; Amore *et al.*, 2013; Grondin *et al.*, 2017). Their cell-bound enzymes are encoded by an operon for a transcriptional regulator gene, a carbohydrate transport system, and a CAZyme (Grondin *et al.*, 2017).

The majority of *Bacillota* which have been cultured belong to the class *Bacilli*. This class contains multiple genera (*e.g.*, *Bacillus*, *Paenibacillus*) which possess laccase and peroxidase genes which are used for degradation of lignin (Tian *et al.*, 2014). Representatives of the genera *Bacillus* and *Paenibacillus* are highly competent at degrading hemicellulose, and produce multi-enzyme complexes for this purpose (Jones, van Dyk and Pletschke, 2012). For degradation of cellulose, *Bacillus subtilis* is one of the few microorganisms known to produce multifunctional enzymes, and produce an endoglucanase which also has significant β-glucosidase activity (Ko *et al.*, 2013). Culture of more representatives of the *Bacillota* may lead to the discovery of more highly effective lignocellulose degrading organisms and novel enzymatic mechanisms for its degradation.

1.3.2.3.2.4 *Bacteroidota*

The *Bacteroidota* are an interesting phylum from the perspective of cellulose utilisation. They are efficient cellulose degraders which are found almost ubiquitously (soils, oceans, freshwater, landfills, hydrothermal vents, the angiosperm microbiome, the gastro-intestinal-tract microbiome of mammals, birds, and echinoderms, and the termite hind-gut) (Thomas *et al.*, 2011). In the gut microbiome, *Bacteroidota* are the phylum with the largest number of CAZymes in their genomes, and represent a large proportion of the total cell count (Thomas *et al.*, 2011). Metagenome annotation of bacteria from landfill identified *Bacteroidota* as the phylum with the most CAZymes (Ransom-Jones *et al.*, 2017). An analysis of the metatranscriptome of decomposing rice-straw under mesophilic and thermophilic conditions identified that *Bacteroidota* were the phylum which produced the second most glycoside hydrolase transcripts under both conditions, with the genera *Niabella* and *Niastella* expressing the majority of transcripts involved in the degradation of

hemicellulose (Simmons *et al.*, 2014). Part of the reason that degradative abilities are so widespread in the *Bacteroidota* is because gene clusters for the detection, breakdown and importation of carbohydrates are transferred between *Bacteroidota* individuals both horizontally and vertically, meaning that many species have dozens of these gene clusters in their genomes (Terrapon *et al.*, 2015).

Members of *Bacteroidota* have thus far been found to utilise two atypical mechanisms for the breakdown of lignocellulose: gliding-dependent membrane bound cellulases and PULs (polysaccharide utilisation loci).

The genus *Cytophaga* within the phylum *Bacteroidota* degrades crystalline cellulose by gliding over the substrate, allowing its cell-associated endoglucanases to deconstruct cellulose polymers. Most of the research into this genus has so far been focussed on the soil bacterium *Cytophaga hutchinsonii*.

The gliding cellulose-utilisation mechanism relies on (1) motility genes (Zhu and McBride, 2014), (2) cell-bound enzymes present on the outer cell membrane, and in the periplasm (Zhu and McBride, 2017), and (3) putative type II secretion systems, which allow for efficient adhesion to the crystalline substrate (Wang *et al.*, 2017).

The presence of periplasmic endoglucanase genes in the genome of *C. hutchinsonii* suggests that cellodextrins are imported into the periplasm as they are in PUL systems—the genes responsible for this in *C. hutchinsonii* have not yet been identified (Thomas *et al.*, 2011). Genomic analysis of *C. hutchinsonii* identified nine endoglucanase genes, as well as four β-glucosidase genes with differing rates of hydrolysis and differing tolerances to inhibition by glucose (Zhu and McBride, 2017). Locomotion plays a crucial role in the degradation of cellulose by *C. hutchinsonii*; deletion of the *spR* gene which controls cell motility was shown to vastly reduce motility, concurrently halting digestion of crystalline cellulose (Zhu and McBride, 2014).

Crystalline cellulose is the carbon source which gives fastest growth of *C. hutchinsonii*. Unusually for a truly cellulose degrading organism, *C. hutchinsonii* seems to lack processive exoglucanases, which would act synergistically on crystalline cellulose with their endoglucanases. The lack of exoglucanases in the *C. hutchinsonii* genome suggests that these organisms may move to amorphous regions of the cellulose crystal and utilise these parts preferentially. Movement of the cells may potentially aid disruption of the cellulose crystal structure (Zhu and McBride, 2017).

C. hutchinsonii has genes for xylanases and pectinases, but is unable to grow on xylan or pectin as sole carbon sources; these enzymes may therefore be used to promote access to the cellulose in lignocellulosic substrates (Zhu and McBride, 2017).

The other mechanism of lignocellulose degradation used by members of the *Bacteroidota* is the polysaccharide utilisation locus (PUL). These genomic loci encode a physically linked and complete system for the detection of, adhesion to, and degradation of polysaccharides, as well as for importation of the breakdown products into the cell (Thomas *et al.*, 2011).

PULs are specialised for different carbohydrate types, with PULs for starch (Anderson and Salyers, 1989), hemicellulose (Rosewarne *et al.*, 2014), and chitin (Larsbrink *et al.*, 2016) having been thus far been identified. There is as yet no firm evidence of cellulose-degrading PULs, as cellulose-specific PULs have only been predicted from uncultured representatives of *Bacteroidota via* metagenomic assembly and annotation (Naas *et al.*, 2014; Larsbrink *et al.*, 2016).

In addition to PULs, a gene (*CelEx-BR12*) from uncultured representative of *Bacteroidota* (85% similarity to *Prevotella ruminicola 23*), expressed in *Escherichia coli*, had exoglucanase, endoglucanase, and xylanase activity (Ko *et al.*, 2013).

Members of the *Bacteroidota* are extremely efficient degraders of lignin (Taylor *et al.*, 2012). However, there are very few cultured representatives of ligninolytic *Bacteroidota*, and so our understanding of their role in the degradation of organic carbon in natural ecosystems is limited (Ten *et al.*, 2006; Suihko and Skyttä, 2009; Taylor *et al.*, 2012).

1.3.2.3.2.5 *Thermotogota*

The hyperthermophillic *Thermotogota* are particularly well represented on the CAZy database (Figure 3), most likely because of the use of *Thermotoga* in industrial conversion of biofuel crops into hydrogen (Liebl, 2001; Pollo, Zhaxybayeva and Nesbø, 2015). Isolated from hot springs, *Thermotoga maratima* has been adopted as a source of industrial enzymes because its cellulases are maximally active at 80°C (Liebl, 2001). The *Thermotogota* are anaerobes which can be hyperthermophilic, thermophilic, or mesophilic. Mesophilic *Thermotogota* are distributed in largely-thermophilic clades, suggesting that species of *Thermotogota* may be widely distributed in mesothermic environments (Pollo, Zhaxybayeva and Nesbø, 2015). *Thermotogota* generally have small genomes which has been suggested to be linked to their degree of thermophily (Akram *et al.*, 2022). As such, their genomes are simple to reconstruct, and many complete genomes are

available on GenBank—meaning that CAZy annotates these genomes, possibly leading to high representation.

The enzyme systems of *Thermotoga maratima* contain many endoglucanases, a β-glucodisase, a laminarinase, and a cellobiose phosphorylase (Liebl, 2001). Other members of the *Thermotoga* have been identified as having thermostable, glucose- and toxic inhibitor-tolerant β-glucosidases (Akram *et al.*, 2016). Additionally, members of *Thermotogota* express thermostable xylanases (Chen *et al.*, 1997; Pandit *et al.*, 2016).

1.3.2.3.2.6 *Chloroflexota*

Profiling of ^{13}C-enriched DNA (DNA-SIP) and ^{13}C-enriched RNA (RNA-SIP) from ^{13}C-cellulose amended soils detected that *Chloroflexota* (along with *Bacteroidota* and *Planctomycetes*) were one of the major phyla involved in degradation of cellulose under oxic conditions (Schellenberger, Kolb and Drake, 2010; Pinnell *et al.*, 2014; Pepe-Ranney *et al.*, 2016). The same SIP study also found that *Chloroflexota* are unlikely important in the utilisation of hemicellulose (Pepe-Ranney *et al.*, 2016).

The anaerobic class *Anaerolineae* has been identified as one of the major classes (by cell-number) present in anaerobic digesters (Xia *et al.*, 2016). The *Anaerolinae* have been shown to express few CAZymes in anaerobic digesters (Xia *et al.*, 2014), and instead, their dominance in anaerobic digesters may be due to their ability to adhere to the cellulosic substrate (Xia *et al.*, 2016).

The *Chloroflexota* with completed genomes appear to have extracellular free-enzyme cellulase systems, as their genomes contain no cohesin, dockerin, or CBM genes (Xia *et al.*, 2016). The aerobic *Chloroflexota*, *Kallotenue papyrolyticum* (family *Kallotenuaceae*, although genome-based taxonomy does not give this a species designation) has a genome containing 55 glycoside hydrolases, and it is able to degrade cellulose and hemicellulose; (Hedlund *et al.*, 2015).

1.3.2.3.2.7 *Proteobacteria*

Proteobacteria are hugely diverse, widely distributed, and well-studied, phylum of aerobic and anaerobic bacteria which include many medically and agriculturally important species. Metagenomics on ^{13}C-enriched DNA from DNA-SIP experiments in soils has identified members of the Alpha-, Beta-, Gamma-, and Delta- *Proteobacteria* as being active cellulose degraders (Haichar *et al.*, 2007; Pinnell *et al.*, 2014). Metatranscriptomic analysis of decomposing rice-straw identified

that the majority of transcripts (as well as glycoside hydrolase transcripts) were of proteobacterial origin (Simmons *et al.*, 2014).

Myxococcales are a largely aerobic order of *Deltaroteobacteria* which are involved in the degradation of complex polysaccharides. Their genomes contain many glycoside hydrolases, however, not all members of this order are capable of degrading cellulose (Sharma, Khatri and Subramanian, 2016). The aerobic *Deltaproteobacteria* genus *Sorangium* (Family *Myxococcales*) has been shown to be effective in the degradation of cellulose and hemicellulose. It uses gliding motility and cell-surface bound enzyme complexes, similar to the cellulosomes of anaerobic species such as *Clostridium thermocellum*, for the degradation of cellulose (Hou *et al.*, 2006).

Gammaproteobacteria have been shown to be active degraders of cellulose in cellulose enrichment and DNA-SIP studies (Edwards *et al.*, 2010; Lee *et al.*, 2011). Within the *Gammaproteobacteria* the cellulase system seems to depend on whether the organism is primarily aerobic or anaerobic. For example, the facultatively anaerobic *Cellvibrio gilvius* and *Cellulomonas firmi* secrete extracellular multi-domain glycoside hydrolases which degrade cellulose and hemicellulose (Christopherson *et al.*, 2013). In contrast, the aerobic species *Cellvibrio japonicus* (synonym *Pseudomonas cellulosa*), *Saccharophagus degradans,* and *Cellvibrio japonicus* produce single-domain cell surface-bound glycoside hydrolases (Beylot *et al.*, 2001).

The *Alphaproteobacteria* have been associated with breakdown of cellulose, hemicellulose and lignin in soils through DNA-SIP techniques (Pinnell *et al.*, 2014; Verastegui *et al.*, 2014). In another DNA-SIP study, *Alphaproteobacteria* which incorporated most ^{13}C into their DNA from ^{13}C-cellulose were the *Rhizobiales*, *Caulobacterales*, *Rhodospirillales*, and *Sphingomonadales* (Eichorst and Kuske, 2012). Some of the organisms which incorporated ^{13}C into their DNA may have done so through the assimilation of released oligosaccharides from the labelled substrate, and so these taxa may not all be primary lignocellulose degraders (Pinnell *et al.*, 2014). Some of the taxa which were identified by DNA-SIP likely were primary utilisers of the labelled substrate, as certain members of *Alphaproteobacteria* have (ligno-)cellulase genes in their genomes, and some strains have been proven to exhibit hemicellulose, exo-and endo-glucanase activity (Hottes *et al.*, 2004; Soares Júnior *et al.*, 2013; Premalatha *et al.*, 2015). Caution should be advised, however, when attributing functional information to microorganisms based off a single marker gene which is not directly related to the function of interest.

Proteobacteria contains many species with pathways for catabolism of aromatic compounds, which are crucial for the degradation of lignin. These are mainly found in the *Alphaproteobacteria* (genera *Brucella*, *Ochrobactrum*, *Sphingomonas*, *Sphingobium*, *Sagittula*) and *Gammaproteobacteria* (genera *Pseudomonas*, *Enterobacter*, *Citrobacter*, *Klebsiella, Pandoraea*, *Burkholderia*), however, Deltaproteobacteria (genus *Geobacter*) also contain lignin degraders (Tian *et al.*, 2014).

1.3.2.3.2.8 *Fibrobacterota*

The *Fibrobacterota* is a phylum which contains only seven cultivated (GTDB R214), obligately anaerobic, species. Within the genus *Fibrobacter*, are the species *Fibrobacter succinogenes* and *Fibrobacter intestinalis*. Both of these are found within herbivore guts and have a high rate of cellulose decomposition (Arntzen *et al.*, 2017). Investigation into how *F. succinogenes* gains and utilises carbon so efficiently has recently elucidated a novel mechanisms of cellulolysis; attachment of the cell to the cellulose *via* fibro-slime proteins is followed by production of extracellular cellulases—these release cellodextrins which are then transported into the bacterial periplasm for hydrolysis by β-glucanases or other classes of cellulase (Burnet *et al.*, 2015). Additionally, *F. succinogenes* produces vesicles which contain many CAZymes for the degradation of hemicellulose and pectin; this is particularly interesting as *F. succinogenes* is unable to utilise the pentose sugars released by hydrolysis of hemicellulose and pectin (Arntzen *et al.*, 2017). Pretreatment of lignocellulosic substrates with these vesicles increased the efficiency of cellulase mixtures 2.4-fold, showing the importance of this strategy for a bacterium which is only able to utilise hexose sugars (Arntzen *et al.*, 2017).

Annotated metagenomes from landfill leachate identified very few classes of CAZyme belonging to *Fibrobacterota*. These annotations identified only four different types of CAZyme (carbohydrate binding modules, glycoside hydrolases, glycosyl transferases, and polysaccharide lyases) assigned to *Fibrobacterota*, as opposed to the nine assigned to *Bacteroidota* and *Bacillota*, seven assigned to *Proteobacteria*, and six assigned to *Spirochaetes* (Ransom-Jones *et al.*, 2017).

Interestingly, high numbers of diverse genes for lignocellulolytic enzymes and carbohydrate binding modules were found in the genomes of all metagenome-assembled genomes in a recent study (López-Mondéjar *et al*., 2022), however, there are no representatives of the *Fibrobacterota*

phylum except *F. succinogenes* on the CAZy database, despite high quality genomes being available.

1.3.2.3.2.9 *Planctomycetes*

The *Planctomcyetes* have been identified as being strongly associated with breakdown of lignocellulosic material through DNA-SIP studies (Schellenberger, Kolb and Drake, 2010; Wang *et al.*, 2015).

Metatranscriptomic analysis of *Sphagnum* peat amended separately with cellulose, xylan, pectin, and chitin identified plactomycete OTUs (operational taxonomic units) which responded positively to each of the substrates (Ivanova, Wegner, *et al.*, 2017). In addition to this, genomic analysis identified 44 glycoside hydrolase genes, 83 glycosyltransferase genes, and 12 genes for carbohydrate esterases in the genome of *Paludisphaera borealis*, excluding many putative CAZymes (Ivanova, Naumoff, *et al.*, 2017). Most of the CAZymes in the genome of *P. borealis*, and some other *Plancotmycetes*, are arranged clusters which suggests that they may be co-expressed, although this is not true for all species (Ivanova, Naumoff, *et al.*, 2017).

Cultured representatives of the *Planctomycetes* are able to grow on several complex polysaccharides, however only a single cultured representative has been shown to be able to degrade crystalline cellulose (Kulichevskaya *et al.*, 2012; Ravin *et al.*, 2018).

1.3.2.3.3 Archaea

Archaea are unlikely major utilisers of cellulose in *Bacteria*- and *Fungi*-dominated ecosystems due to their slow growth and small genome sizes. With respect to carbon cycling, archaeal species are mostly known for their role in methanogenesis, however, species with the metabolic capability for glycolysis and beta oxidation of fatty acids are present in phyla within all known archaeal superphyla (Baker *et al.*, 2020). Analysis of metagenomes from cellulose enriched anaerobic sludge identified 12 glycoside hydrolases from *Archaea*, relative to the 236 from *Bacteria*. In the same study, polysaccharide and oligosaccharide metabolism were almost entirely associated with *Bacteria* (Xia *et al.*, 2013). Very few *Archaea* have been isolated (648 species with cultivated representatives (Parks *et al*., 2021, GTDB R214)), and very few have been proven to use elements of lignocellulose as energy sources (Gavrilov *et al.*, 2016), with many resources suggesting that they are more specialised at utilisation of lower-energy compounds which are the end-products of bacterial metabolism. Recent papers on thermophilic archaea have found species with

(hemi)cellulases, some of which have several different catalytic domains and little DNA sequence homology to bacterial CAZymes (Graham *et al.*, 2011; Leis *et al*., 2015; Gavrilov *et al.*, 2016; Lewin *et al.*, 2017). Hypersaline environments have also recently yielded novel archaeal species (*Halorhabdus tiamatea* and *Halorhabdus utahensis*) with a high density of xylanolytic and cellulolytic genes (Zhang *et al.*, 2011; Werner *et al.*, 2014). It is therefore likely that there is much more diversity in archaeal (ligno)cellulases than we have thus far discovered.

Five archaeal genes from oil reservoirs gave *in vitro* cellulase activity when expressed in *Escherichia coli* (Lewin *et al.*, 2017). These enzymes had endo-1,4-β-glucanase and possible exoglucanase activity, with three of the enzymes being thermostable. The most active cellulase from this study was especially active on CMC, giving higher rates of degradation than traditional enzyme mixtures (endo-1,4-β-glucanase, cellobiohydrolase I, β-glucosidase), and had a carbohydrate binding module and three endoglucanase domains with low sequence homology with previously annotated genes (Lewin *et al.*, 2017). Furthermore, this enzyme had action on crystalline cellulose and 4-methylumbelliferyl-β-d-cellobioside (MUC), producing cellobiose and some glucose as products (Lewin *et al.*, 2017). A strain of the hyperthermophilic euryarchaeon *Thermococcus* has been isolated which is able to degrade CMC, amorphous cellulose, xyloglucan, and chitin (Gavrilov *et al.*, 2016). The genome of this isolate contains 18 glycoside hydrolases and carbohydrate esterases, five of which were predicted to be involved in the degradation of β-glycosides, and four of which were predicted to be extracellular. One of the enzymes has three glycoside hydrolase domains and two CBMs, and is active on CMC, microcrystalline cellulose, and xylan. The activity on microcrystalline cellulose suggests that the enzyme may act as an endoglucanase, an exoglucanase and a β-glucosidase (Gavrilov *et al.*, 2016). Other studies have reported archaeal thermo- and halo-tolerant β-glucosidases and endoglucanases (Sinha and Datta, 2016), glycohydrolases (Susanti *et al.*, 2012), and novel, uncharacterised cellulose utilisation mechanisms (Mardanov *et al.*, 2012). In addition to the degradation of cellulose and hemicellulose, *Archaea* which have genes for the degradation of lignin have been isolated. Laccase and manganese peroxidase genes have been found in the genomes of the *Halobacteriales* (*Euryarchaeota*) and *Thermoproteales* (*Crenarcheota*) (Tian *et al.*, 2014). Further study into the diversity of non-eukaryotic archaeal (Eme *et al.*, 2017) life is an exciting prospect which may reveal many novel genes and functions.

1.4 Methods for study of microbial lignocellulose degradation

1.4.1 Metagenomics and metatranscriptomics

Direct sequencing of environmental samples which contain DNA from the whole community of organisms present, or metagenomic sequencing, is a common technique to study communities of organisms which are difficult to observe directly. Through reconstruction of genomes *via* assembly and binning, and homology-based functional and taxonomic annotation of genes and sequences, much can be learned about the taxonomic profile and functional potential of a community (Stewart *et al.*, 2019). Metagenomics also circumvents the difficulty associated with cultivating diverse microorganisms, and quickly gives a community-scale view of the system under study, instead of detailed information about one or two species and their functions. Community-level overviews of the functional potential of ecological systems are useful tools for comparison of ecosystems or treatment effects on the composition of organisms and genes, although the status of well-informed prediction of function from metagenomes is still firmly in the elemental phase (Martinez, 2023). Metagenomic sequencing may not be suitable for detecting short-term changes in community composition however, as DNA can persist in soil for thousands of years (Rawlence *et al.*, 2014). Additionally, metagenomic annotation relies upon homology, providing results which are biased towards annotating well characterised groups and genes (Hugenholtz and Tyson, 2008)

As DNA sequencing has become cheaper and computational techniques and hardware have improved, recovery of thousands of complete or near-complete microbial genomes from environmental samples containing thousands of species has become possible (Stewart *et al.*, 2018, 2019). It is important to note that assembly from complex samples may integrate sequences from multiple closely related strains, and so assembled metagenome-assembled genomes (MAGs) give an “average genome” for that species, although methods exist which reduce the likelihood of chimeric sequences being included in the MAG contig bins; these include searching for differential coverage, and checking the composition of genome-wide tetranucleotide frequency (Sangwan, Xia and Gilbert, 2016). Further limitations to metagenomics for understanding community function include the recovery of genomic DNA from inactive community members which are either dormant, senescing, or dead (Buerger *et al.*, 2012; Carini *et al.*, 2016).

The related technique of metatranscriptomics (*i.e.,* sequencing of the total RNA from an environmental sample) ensures that only active cells and species are detected which is a major advantage when interpreting how microbial communities differ in composition. Producing

metatranscriptomic datasets presents several technical challenges such as preventing the rapid degradation of the nucleotides, increased difficulty of extraction of RNA relative to DNA, removal of high abundances of rRNA which dominate sequencing libraries through depletion or oversequencing, the need for de-novo assembly of the resulting sequences which depends upon input RNA integrity, or a high-quality metagenomic assembly to map sequences to, and high sequencing costs. Metatranscriptome analyses often provide lists of differentially expressed genes across samples and treatments which are involved in the cellular processes occurring in the microbial community. These lists of genes provide good targets for further investigation *via* cultivation or heterologous expression and translation, and characterisation of enzymatic activities.

Whilst both techniques allow identification of genes which may be involved in the deconstruction of lignocellulose, characterisation of the functional activities that they encode is not possible through this approach. Functional metagenomics however, *i.e.*, heterologous expression of genes from metagenomic samples in libraries of common laboratory organisms (often *E. coli*), provides a method for understanding the functional potential of genes from uncultivated organisms. Such approaches have been used to elucidate the lignocellulolytic potential of species of *Archaea* from deep-sea vents (Leis *et al*., 2015) and from deep-subsurface petroleum reservoirs (Lewin *et al*., 2017). Due to the likely difficulty in recreating suitable conditions for laboratory growth of these organisms, functional metagenomic approaches and functional screening of metagenomic DNA allows for great expansion of the environments that can be studied *in vitro*.

1.4.2 Stable isotope probing and profiling of enrichments

Stable isotope profiling (SIP) can help to elucidate an organism's role in the environment. Experimentally adding a stable-isotope-labelled substrate (usually ^{13}C or ^{15}N) allows identification of the organisms which have incorporated isotopes from the substrate into their cellular structures. Ultracentrifugation of the labelled biomarker (*e.g.*, DNA, RNA, PLFA) in caesium chloride can be used to separate the heavy (labelling-isotope incorporated) and light (labelling-isotope not incorporated) fractions of biomarker (Neufeld, Dumont, *et al.*, 2007; Neufeld, Vohra, *et al.*, 2007; Leung *et al.*, 2016). Biomarkers used in stable isotope profiling include phospholipid fatty acids (PLFA), DNA, and RNA (Neufeld, Dumont, *et al.*, 2007). These markers each have different advantages and drawbacks which must be considered when choosing a method. PLFA is the most sensitive of these techniques, requiring only 0.1% incorporation of the stable isotope,

whereas DNA and RNA require 25-30% incorporation for detection of the differences (Jehmlich *et al.*, 2010). The taxonomic resolution attainable with PLFA-SIP is much lower than is attainable with DNA- or RNA-SIP. However, because of the fast rate of incorporation of the heavy isotope, PLFA-SIP allows study of processes in oligotrophic systems, where slow growing bacteria or archaea are the focus, or where the organisms of interest are known to utilise multiple carbon sources (Bull *et al.*, 2000; Neufeld, Dumont, *et al.*, 2007). Using DNA as the biomarker has the advantage of allowing for powerful downstream high-resolution analyses (including metagenomics) of genes belonging to organisms which are actively involved in the breakdown of compounds (Neufeld, Dumont, *et al.*, 2007), with the possibility of reconstruction of complete genomes from isotopically-labelled DNA. However, DNA-SIP has the disadvantage that the substrate must be left *in situ* long enough for cell division to ensure the incorporation of ^{13}C; this time dependency increases the chance of predation or 'cheaters' incorporating the ^{13}C into their cellular components (Neufeld, Dumont, *et al.*, 2007).

When combined with sequencing, SIP allow for identification of microorganisms actively involved in C assimilation (those with DNA which has incorporated more of the heavy isotope), whilst using less expensive sequencing options such as 16S rRNA gene and ITS gene profiling. By comparison with the barcodes in the heavy fractions, and with the equivalent fractions from controls, tentative identification of the organisms involved in the metabolism of the amended substrate is possible. This technique does not preclude detected microorganisms from having incorporated the radiolabelled carbon *via* predation of the organism which initially broke down the substrate, nor does it stop detection of microorganisms which opportunistically utilised isotopically labelled monosaccharides released by other organisms (Leung *et al.*, 2016). Additionally, slow-growing microorganisms which are active degraders of the amended substrate may not be detected, as they incorporate radiolabelled carbon more slowly into their cellular components (Leung *et al.*, 2016).

These approaches highlight which oranisms or genes may be involved in the processes, but offer no functional data about the actual activities of the novel genes that they discover. Stable isotope probing can also be used for the construction of fosmid or cosmid libraries, which allow for screening and characterisation of the functions of the metagenomic sequences expressed through a laboratory-cultivated host, which has been used for the discovery of lignocellulosic genes from uncultivated microorganisms (Neufeld *et al*., 2008; Verastegui *et al*., 2014).

1.4.3 Isolation and cultivation of lignocellulolytic microorganisms

1.4.3.1 Standard media

Isolation of lignocellulolytic microorganisms from environmental samples using traditional techniques involves detachment of microorganisms from the environmental sample using a liquid medium (*e.g.*, phosphate buffer saline), and shaking, serial dilution, and plating onto a solid medium before observing the growth of individual microbial colonies. These colonies are then picked and streaked onto new agar plates where further subculturing can occur, until isolate purity is achieved. These isolates may then be screened for lignocellulose decomposition activity using a range of phenotypic screens. These include screening *via* the Congo red assay which binds to carboxymethycellulose (CMC), looking for zones of clearing, assays for reducing sugar content when polysaccharides such as xylan are given as a sole carbon source (Malgas and Pletschke, 2019), and spectrophotometry of fluorescently labelled breakdown products (Ahmad *et al.*, 2010).

1.4.3.2 Choice of media

Alterations to or choice of artificial media which make the medium more similar to the environment the sample was taken from can increase the number of microorganisms which can be isolated and grown in pure culture. Nutrient-poor media has been shown to increase the number of colony forming units (CFU) obtained from soil, relative to the number of CFU obtained using the same broth at a higher concentration (Janssen *et al.*, 2002; Vartoukian, Palmer and Wade, 2010). The isolates obtained by cultivation on minimal media have been shown to be significantly divergent from those cultivated on nutrient-rich media—and a large proportion of these isolates have been shown to be previously uncultured (Janssen *et al.*, 2002). Cultivation on minimal media is thought to allow oligotrophic microorganisms time to grow, whilst inhibiting growth of the fast-growing species which are traditionally recovered through culture work. Some bacteria rely on metabolites produced by other bacteria for their survival, either for nutrition, or as signalling molecules which suggest that a habitat is stable, rather than ephemeral. These organisms are therefore only found on standard growth media in coculture, although study into the coculture can yield strategies which can be used for their axenic culture (Nichols *et al.*, 2008). For example, addition of signalling molecules to liquid media has been shown to vastly increase the numbers of cultivable microorganisms (Vartoukian, Palmer and Wade, 2010). Even closer to the original environment is a sterilised sample of the environment (Taylor, 1951). Sterilisation techniques such as autoclaving can strongly alter the chemistry of a sample, and preclude growth of many

microorganisms, however sterilisation through gamma irradiation may be viable techniques for increasing cultivability of microorganisms (Salonius *et al.*, 1967).

Alterations to the gelling agent used for solid media help to cultivate different microorganisms as each gelling agent allows different pH ranges, has different setting properties with regards to temperature, and may have inhibitory effects on the growth of particular species allowing the growth of organisms with different life histories (Das *et al.*, 2015; Rygaard *et al.*, 2017).

The method of preparation of media can strongly affect the rate of discovery of novel species. This has been shown to increase by 0.4-fold for fast-growing species, and by 8.4-fold for slow growing species (Kato *et al*., 2018).

1.4.3.3 Increased incubation time

Long-term incubation of microorganisms has led to the cultivation of novel microorganisms, including the first representatives of *Pelagibacteriales* and *Verrucomicrobia*, although a study on the effect of cultivation time on the novelty of growing species in marine sediment and soils showed that this is likely due to the random activation time of different inactive but viable cells (Buerger *et al.*, 2012; Kurm *et al*., 2019). Therefore, higher sampling and cultivation effort leads to increased novelty of microorganisms, whether this be through isolation of more cells, or through longer incubation times. However, the authors note that the techniques used in long-term cultivation experiments may be effort-efficient methods of cultivating novel microorganisms (Buerger *et al.*, 2012).

1.4.3.4 High-throughput dilution to extinction and physical isolation of individual cells

Because traditional cultivation methods bias the isolates in favour of abundant and competitive species on the medium used for cultivation—excluding reproducing but less competitive species—alternative techniques are required to increase the diversity of cultivated isolates. Dilution of the inoculation liquid can be used to reduce the diversity of the sample to the point where only single cells, or no cells are growing on the medium. This technique is effective and has been utilised since the 1930s (Kim *et al.*, 2020). To combat the laborious nature of microbial isolation by traditional methods, and to allow upscaling of isolation of diverse microbial community members, high-throughput isolation and cultivation methods have been developed which utilise the dilution to extinction method. In one approach a series of 96-well microplates containing serially diluted inoculum are prepared, and plates in which fewer than a defined percentage of wells (typically

30%) show microbial growth are kept. This cut-off increases the likelihood that visible colonies originate from single cells (Zhang *et al.*, 2021). If a small number of cells were present in each microwell at inoculation, it is often possible to purify the microorganisms further through streaking or other methods. Using this approach, it is relatively simple, space-, and time-efficient to obtain hundreds to thousands of isolates from a single sample.

As many microbial cells adhere to solid substrates (*e.g.*, soil, lignocellulose), dilution alone may still provide diverse communities of microorganism if a small piece of substrate is included in the dilute sample. Cell-sorting methods resolve this issue and allow cultivation of pure stains highly efficiently, although these rely upon expensive machines which are not available in many laboratories. Single-cell isolation techniques such as flow cytometry, can increase the phenotypic relevance and phylogenetic diversity of cultivated microorganisms. This is especially true when coupled with techniques to identify cells in particular metabolic states or with particular phenotpyes (*e.g.*, fluorescent labelling dependent on membrane permeability or detection of incorporation of alkyne-labelled amino acids intoproteins using bioorthogonal non-canonical amino acid tagging (BONCAT), sorted using fluorescence activated cell-sorting (FACS)) (Espina, 2020; Couradeau *et al*., 2019)

1.4.3.5 In situ cultivation

To ensure that the community of organisms which are cultivated is as similar as possible to the community in the soil, *in situ* cultivation provides the conditions, growth factors and chemicals which occur in the environment naturally.

Early work in the 1920s and 1930s, by pioneers of microbiological research in the environment, studied microorganisms grown *in situ*, or in conditions similar to those found in the microorganism's native environment.

Conn famously developed cell staining methods with dyes produced in the USA after the end of the first world war to separate living and dead cells in environmental samples, test for presence of organelles or prokaryotic cellular components such as flagella, as well as for many medical purposes (Conn, 1960). Winogradsky pioneered the field of functional community ecology for microorganisms, and created columnar mesocosms consisting of sand, water and multiple carbon and nitrogen sources with which he enriched different functional groups of microorganisms in spatially separated areas of the columns. Using this approach, Winogradsky cultivated

microorganisms from particular areas of the mesocosms to discover the functional roles that particular microorganisms play in the environment—through cultivation of these communities and their members, he elucidated several organisms involved in the terrestrial nitrogen cycle (bacteria involved in the two steps in nitrification, and anaerobic nitrogen fixers), and was the first to identify chemoautotropic microorganisms (Dworkin and Gutnick, 2012). Cholodny also studied microorganisms in their natural habitat—in 1930 he characterised the community of microbial colonies which grew on glass slides placed into the soil, and in 1934 he set up "soil chambers" which allowed for microscopy which showed the soil microbial community dynamics of colonization over time *via* microscopy (Cholodny, 1934). Microbial community dynamics over time is a major factor which should be considered when assessing the impact of environmental perturbances on a microbial community. Such approaches were independently used by Rossi (Rossi *et al*., 1936). ZoBell adopted the idea of studying microorganisms *in situ* by using a device which held glas slides in place while they were incubated in the sea. This work allowed ZoBell to cultivate diverse bacteria, and lay the foundations for understanding biofilm formation by studying individual species and their population sizes as they attached to the slides (ZoBell, 1946).

More recent approaches to *in situ* cultivation rely upon diffusion chambers, which are chambers containing solid cultivation medium, covered by membranes with small pores which allow the transfer of small molecules, but not of microorganisms. By trapping microorganisms in a diffusion chamber (Kaeberlein, Lewis and Epstein, 2002; Bollmann, Lewis and Epstein, 2007; Remenár *et al.*, 2015; Jung, Aoi and Epstein, 2016) or iChip (Nichols *et al.*, 2010), and then placing them back into the site of sampling, the microorganisms can interact with the chemicals they require for growth, whilst adapting to an agarose environment. This adaptation or "domestication" process allows a larger number of species to grow *in vitro* (Bollmann, Lewis and Epstein, 2007; Remenár *et al.*, 2015). Additionally, cultivation *in situ* ensures that microorganisms which can grow effectively in their realized niches are the ones which are studied further (Jung, Aoi and Epstein, 2016).

Early work on *in situ* cultivation noted that the number of cells which formed microcolonies were variable (mean of 22% maximum of 40%, minimum of 2%) as dependent on sampling month (Kaeberlein, Lewis and Epstein, 2002). Additionally, the authors noted that only 14% of these microcolonies would grow when passaged on to Petri-dishes (*i.e.,* 3% of inoculated cells form colonies), and that a substantial number of these were mixed-cultures which required several passages to become axenic (Kaeberlein, Lewis and Epstein, 2002).

In situ cultivation in diffusion chambers, followed by passage onto Petri-dishes, has been shown to increase the diversity and genetic novelty of the isolates attained relative to that attained *via* standard cultivation techniques (Bollmann, Lewis and Epstein, 2007). Further, *in* situ cultivation promotes growth of different species within the same phyla as attained by isolation directly on Petri-dishes containing the same media (Bollmann, Lewis and Epstein, 2007; Remenár *et al.*, 2015). Very few studies to date have used *in situ* cultivation, meaning that there is a wealth of novel genes and enzymes which are as yet untapped.

The community of isolates from a single round of *in situ* cultivation, from several studies, belong mostly to the phyla *Alphaproteobacteria*, *Betaproteobacteria* and *Gammaproteobacteria* (Bollmann, Lewis and Epstein, 2007; Nichols *et al.*, 2010; Jung, Aoi and Epstein, 2016), but isolates from *Bacillota* and *Bacteriodetes* have also commonly been found in the first round of cultivation in diffusion chambers (Bollmann, Lewis and Epstein, 2007; Remenár *et al.*, 2015). Subsequent rounds of *in situ* cultivation have been shown to lead to increased richness within and between phyla, and also give substantial increases in the recovery rate of microcolonies and colonies which will grow on agar (Bollmann, Lewis and Epstein, 2007; Remenár *et al.*, 2015). This increase in the number of isolates from multiple rounds of *in situ* cultivation may be due to adaptation by the bacteria to the agarose media, making them less reliant on their natural environment for reproduction (Bollmann, Lewis and Epstein, 2007), or perhaps as with long-term incubation studies, may be the result of increased sampling effort with extra time (Buerger *et al.*, 2012). Bollmann et al. (2007) showed that four rounds of *in situ* cultivation, could retrieve up to 70% of the inoculated cells.

1.5 Current challenges and opportunities

Study of the microorganisms that degrade soil carbon stocks provides several challenges, not least because of the diversity of microorganisms involved. Genome analyses have identified the potential for lignocellulolytic activity in all domains of life, across many phyla. However, the relative importance of different species depends upon the genes present, microbial life-history strategy (*e.g.*, growers *versus* upregulators (Nuccio *et al.*, 2020)), interactions between species, and the ways in which the environment and lignocellulose composition affects these microbial traits and interactions. Generalisation of results is therefore difficult because of the diversity of species, genes and strains involved.

The initial barrier to a generalisable understanding of microbial community functionality is the vast sea of uncultivated or undetected microbial species, which could be two million times greater than the number of cultivated species (Locey and Lennon, 2016; Parks *et al.*, 2021). The recent increase in the rate of production of MAGs and single-cell amplified genomes (SAGs) is beginning to provide a roadmap to the previously undetected and uncharacterised microbial genomic diversity, however our understanding of this uncultivated diversity presently relies wholly on the functioning of genes of the cultivated minority. With respect to lignocellulose degradation, this gives the possibility that there are thousands of unidentified degradative mechanisms that we cannot detect, because there is no cultivated species with proteins which perform these functions. Significant effort in the cultivation of lignocellulosic microorganisms will be needed to meet this challenge. The recent development of high-throughput techniques such as *in situ* cultivation of enrichments with the iChip (Nichols *et al.*, 2010), next-generation physiology coupled with single-cell isolation methods (Hatzenpichler *et al.*, 2020), or deep-learning guided colony picking, phenotyping and genotyping platforms which output pure isolates and genomes at a total cost of USD 6.82 per sample, at a rate of 2,000 isolates per hour (Huang *et al.*, 2023) will allow a step change in the rate of discovery of linked genotype and phenotypic data. Alongside increasing cultivation efforts, significant advancement in the prediction of function from gene sequences is now possible *in silico* using deep-learning approaches (Jumper *et al.*, 2021), however, the outputs of these models still require validation as the model relies upon the small fraction of proteins that we have characterised *in vitro*.

Understanding and cataloguing the life history strategies of different microorganisms is another major barrier to our understanding of community-level microbial functions. For genome-resolved microorganisms, the combination of genes and their predicted functions, as well as other genomic characteristics provide measurable insights into the fundamental and realised niches that a microorganism may occupy (Wood, Tang and Franks, 2018). Microbial genome-wide association (GWA) studies allow us to relate natural variation in the genotypes of closely related microorganisms to their phenotypes. Unlike its human counterpart, microbial genome-wide association (GWA) has only been developed recently (Earle *et al.*, 2016) and has been little used because laboratory strains allow causative mechanisms to be found through genetic knockouts, and because bacterial clonal replication invalidates some of the methods developed for human GWA (Falush, 2016). The 10-year gap between successful human GWA and microbial GWA studies

allowed for significant advancement in the tools used, and development of GWA tools for both humans and microorganisms continues. There are three main types of microbial GWA study: those which search for phenotype correlated single nucleotide polymorphisms (SNPs), genes, or copy number variations. Microbial GWA studies are becoming more common in medical science, relating variation in microbial genomes to clinical outcomes. This can be used to inform decisions about drug repositioning and improve the design of vaccines and antimicrobials (San *et al.*, 2020). As high-throughput cultivation and screening (both genotypic and phenotypic) become cheaper and more commonplace, microbial GWA will become an effective technique for accurately identifying metabolic pathways involved in biological processes for natural populations, and for functionally classifying the genomes of microorganisms.

Genome-scale metabolic models (GSMMs) are *in silico* models which utilise information from genomes, proteomes, reactomes (metabolic reactions encoded by the genome), and any other lab-based data available, to simulate the metabolic processes of an organism under different environmental conditions. GSMMs are quickly becoming effective tools for the simulation of microbial interactions for increasingly complex communities, and have strong promise for understanding how variation in microbiomes can lead to disease symptoms (Stolyar *et al.*, 2007; Rosario *et al.*, 2018). There is however, a significant difficulty in constructing GSMMs that there are extremely high levels of misannotation, or uncertainty around the function that proteins perform when produced by diverse microorganisms *in situ* (Gerlt, 2017).

As the robustness of the results of GSMM simulations improve, our ability to generalise community-level microbial functional responses to environmental change across ecosystems and biomes will increase. Combination of GSMMs will become an invaluable technique when metabolic models of complex communities from metagenomic inputs become reliable and robust (Zorrilla *et al.*, 2021). Common use of such GSMMs would transform microbial ecology from a primarily descriptive discipline into a predictive discipline, which sets out to test specific hypotheses about the response of microbial species and communities to environmental perturbations. Improvement of these models, and changing the focus from medically and industrially relevant species towards globally abundant and important species in ecosystems, will allow rapid development for many aspects of microbial ecology.

The correct interpretation of data relating to microbial communities is challenging. For instance, to infer effects of an experimental treatment on the microbial community, a microbial ecologist will indirectly measure the relative abundances of different microorganisms, using marker molecules (DNA, RNA, proteins, metabolites) because microbial communities *in situ* are difficult to observe or measure directly. Because the returned data are compositional in nature, their interpretation must be considered through the lens of absolute abundance of the microbial community or marker molecule. Such a simple concept is often missed, and is not reflected by commonly used summary statistics, such as the Shannon diversity of a community, unless compositional data are scaled by a metric of absolute abundance. Further, analysis of the microbial community at an appropriate taxonomic resolution, and a relevant *a priori* choice of genes to study—especially when thousands of genes with possibly related function correlate with differences in experimental treatment—make the study of a complex system where very little is known about the mechanistic drivers of community assembly, particularly challenging.

Consequently, the major remaining challenges and opportunities for understanding and capitalising lignocellulose degradation by soil microbial communities are:

1) Knowledge on the relative contributions of species and broad taxonomic groups to the degradative potential and actual turnover of lignocellulose in soils.
2) Knowledge about how the genetic potential and realised degradation by microorganisms is affected by global changes.
3) Cultivation of diverse and uncultivated species so that we can begin to characterise phenotypes and predict community dynamics with fewer unknowns.
4) Production of (meta-)genome informed carbon cycling models to inform policy makers about land-use decisions in the face of further global changes.
5) Study and selection of microorganisms using high-throughput automated single-cell isolation, cultivation, and characterisation techniques, to aid in the search for novel lignocellulolytic gene classes for sustainable biotechnology.

1.6 Aims of the thesis

The overall aim of this thesis is to further our understanding of the microorganisms involved in the degradation of lignocellulose in soils, and to understand how this may be affected by global

challenges. Because of the strengths and limitations of different approaches for study of lignocellulose degrading microorganisms and communities, the following body of work will use a combination of techniques, including metagenomic sequencing of soil microbial communities (chapters 2 and 3), enrichment culture, high-throughput cultivation, *in situ* cultivation, and genome-wide association between accessory genes and lignocellulolytic phenotypes (chapter 4). The challenges for understanding lignocellulose degradation by soil microbial communities that this thesis hopes to contribute towards are:

1) Knowledge about the relative contributions of species and broad groups to the degradative potential of lignocellulose in soils.
2) Knowledge about how the genetic potential of the community of soil microorganisms is affected by global changes.
3) Cultivation of diverse and uncultivated species for further diverse phenotypic characterisation so that we can begin to characterise and predict community dynamics with fewer unknowns.

Chapter 2 utilises a field experiment in which plants were removed and soil was covered to manipulate carbon inputs to the soil, which has been established for over a decade, to understand the effect of plant exclusion on microbial community composition and effects on soil carbon composition. Chapter 3 utilises the Soil Security Programme's landscape-scale UGRASS experiment to addresses the impact of agricultural land-use intensification on the soil microbial community, and the associated degradative potential. Chapter 4 utilises enrichment culture and high-throughput *in situ* cultivation of microorganisms, phenotypic screens, and genome-wide association to better understand the functional and genomic characteristics of lignocellulolytic soil microorganisms. Finally, chapter 5 summarises and synthesises the findings of this thesis, placing these results into the wider context, and points at future research directions.

2

Impacts of plant exclusion on lignocellulolytic microbial community composition and function

2.1 Abstract

Plant cell wall polysaccharides are the most abundant form of organic carbon in soils and their degradation by microorganisms represents a major link in the global carbon cycle. Soil carbon storage is a valuable ecosystem service, buffering against increasingly rapid climatic change, and underpinning services such as food production and flood prevention. Despite the importance of this link, little is known about the relative contributions of different microorganisms and their genes to lignocellulose degradation in soils. Here, we used a 10-year plant-exclusion experiment on grasslands to study how reduced plant carbon inputs affects the microbial community composition, and genes which are putatively associated with lignocellulolysis. We show that 10-year plant exclusion consistently favours genera in *Bacillales*, *Thermoproteota*, and diverse lineages of *Proteobacteria*, alters the repertoire of lignocellulolytic genes present, and that taxonomic and genetic changes are not clearly linked. A single year of plant-exclusion was linked to increased hemicellulose breakdown product abundance and increased xylanase gene diversity and altered lignocellulase gene composition. Additionally, this study investigates the fundamental and realised niches of specific soil microorganisms, and shows the temporal scale at which the microbial community responds to changing carbon inputs in grasslands.

2.2 Introduction

Soils are Earth's largest store of terrestrial carbon, containing 1,500 gigatonnes of organic carbon (Lal, 2008), predominantly in the form of decaying lignocellulosic plant material (Wu *et al.*, 2021). The microbially-mediated decomposition of lignocellulose in soils is therefore a key feature of the global carbon cycle. Critically, intensive agriculture and climate change have both been shown to negatively affect terrestrial carbon source-sink dynamics, accelerating carbon losses from soil organic matter (SOM) reserves (Crowther *et al.*, 2016; Malik *et al.*, 2018; Chen *et al.*, 2020). An in-depth understanding of the fundamental mechanisms which underpin lignocellulose turnover in soils is therefore required to inform management options that may help to mitigate against these carbon losses. Additionally, the microbial lignocellulolytic enzymes which drive the flux of carbon in soils have vast biotechnological potential; finding and exploiting enzymes with high reaction rates can vastly improve product quality and production rates (Lynd *et al.*, 2017b).

Globally, grasslands store 34% of terrestrial carbon, cover 26% of land area, and account for 80% of agriculturally productive land (Diamond *et al.*, 2019). Soil organic carbon (SOC) accumulates at a rate of 150 g C m^{-2} y^{-1} in alpine grasslands, and microbial respiration in temperate grassland ecosystems releases 390 g C m^{-2} y^{-1} (Wang and Fang, 2009) (Martinez *et al.*, 2016). Microorganisms have evolved diverse genomic strategies for the decomposition of insoluble lignocellulosic plant biomass, which are widely distributed across bacterial and fungal taxa, and underpin major global nutrient cycles (Finzi *et al.*, 2011). However relatively few studies have linked the oxidative and hydrolytic enzymes to the organisms which produce them. Fewer still have studied the link between genomic strategies for the processing of lignocellulosic substrates and taxonomic identity. Linking organisms to the functions they perform is a basic requirement for understanding how environmental perturbations may affect the functioning of ecosystems, for

targeted analysis of enzymes or microbiota with potential biotechnological applications, and for improving the next generation of carbon cycling models.

Metagenomic approaches have been used to link taxonomy to lignocellulolytic function in forest and peat soils (Tveit *et al.*, 2013; Tveit, Urich and Svenning, 2014; Pold *et al.*, 2016; Žifčáková *et al.*, 2017; Abdallah, Wegner and Liesack, 2019; Diamond *et al.*, 2019; Wilhelm *et al.*, 2019; López-Mondéjar *et al.*, 2020), however these relationships remain unstudied in grasslands, which represents a key knowledge gap given the importance of grasslands for carbon storage and agriculture, and as a key component of the terrestrial biosphere. The effect of environmental perturbations on the carbon processing potential of soil microorganisms across grassland ecosystems therefore remains understudied, with potential impacts on future C-cycling models that rely on limited data.

Bare-fallow is an agricultural management practice where soils are kept unvegetated for a limited amount of time in order to promote increased crop yields the following season. It is practiced in semi-arid regions, or regions with highly variable rainfall patterns, and was widely used in response to the great Dust Bowl in the 1930s (Nielsen *et al.*, 2011). Previous studies have shown that bare-fallow soils have an increased proportion of *Actinomycetes* and *Fungi*, reduced SOM content, microbial biomass, and mineralization rates of insoluble carbon substrates, relative to grassland soils (Paterson *et al.*, 2011). Active removal of plants from soils reduces inputs of labile carbon and provides a useful study system to identify members of the microbial community and their genomic strategies for carbon acquisition from complex sources such as lignocellulose.

Here, we wanted to address the paucity of information on the microorganisms and genomic strategies responsible for lignocellulosic plant biomass decomposition in grassland soils, due to the fundamental importance of these microorganisms and their degradative processes in upholding

vital ecosystem services, such as food production. To do this we used a long-term carbon deprivation experiment (grassland vs. bare plots) and a combination of fibre analysis, elemental analysis, metabolomics, and metagenomics to achieve two goals. First, we explored the effects of grassland *vs* bare treatments on soil chemical characteristics such as the abundance of lignocellulose and its breakdown products. Second, we identified the effect of grassland and bare treatments on microbial community composition and gene families associated with lignocellulose turnover in soils (carbohydrate-active enzymes, CAZymes, lignocellulolytic genes, cellulase genes, hemicellulose genes, xylanase genes, auxiliary activities), and how these relate to soil properties. We hypothesised that microbial community function in bare plots would transition towards microbiota with many genes for plant biomass degradation, when compared to grassland plots with high labile carbon inputs. This study therefore reveals the taxonomic and functional properties of key microbiota associated with lignocellulose decomposition and SOC loss under bare conditions.

2.3 Methods

2.3.1 Experimental design

Soil for the experiment was taken from plots that were established at Bangor University's Henfaes Research Centre, Abergwyngregyn, UK (53.24°N, 4.02°W; EL: 12 m). Six 9 m^2 plots were established in 2005 (henceforth "10-year"), demarcated by plastic frames reaching 25 cm into the soil, with 5 – 8 cm protruding above ground. In 2015, a further eight plots (henceforth "1-year") were established adjacent to the 2005 plots to increase replication, and explore temporal effects. Two layers of black gas and water permeable fabric (henceforth "bare") covered half of the plots in each of the age categories, to prevent plant growth. The remaining plots were left as controls (henceforth "grassland"), and were mown annually. Grass outside of established plots was mown frequently. The bare treatment was designed to reduce carbon inputs (root exudates, sloughed

cells, and dead plant matter) from plants to the soil, which we refer to as carbon deprivation. The site has a mean annual soil temperature at 10 cm of 10.2^{o}C, a mean annual rainfall of 1060 mm and has a temperate oceanic climate regime. The soil is classified as a Eutric Cambisol and the vegetation consists largely of *Lolium perenne* L. interspersed with *Holcus lanatus* L. and *Festuca ovina* L.

Ten subsamples of soil were collected from each plot (n=14 plots) in spring 2015 and 2016 using a 1 cm diameter stainless steel soil corer (0 – 10 cm depth), for the 2005 and 2015 plots, respectively. Subsamples from each plot were homogenised and pooled. Each sample was subsampled and either air-dried for physiochemical analysis, immediately frozen (-80°C) and freeze-dried for metabolomic analyses, or transferred to a -80°C freezer for DNA extraction, phospholipid fatty acid (PLFA) profiling and fibre analysis (George *et al.*, 2021). Soil chemistry, respiration and PLFA data were previously reported by George *et al.* (2021).

2.3.2 Fibre analysis and total carbon

To measure total soil carbon, frozen soil was oven dried at 105°C, ground to pass a 2 mm sieve, and was accurately weighed into tin crucibles for analysis on a TruSpec CN Analyser (Leco Corp, St Joseph, MI). For fibre analysis, frozen soils were oven-dried at 40°C for 24 h, and approximately 0.5 g dried soil from each sample was accurately weighed into an Ankom F57 fibre filter bag (ANKOM Technology, Macedon NY, USA) which was then heat-sealed. Neutral detergent fibre (NDF) and acid detergent fibre (ADF) procedures were performed sequentially on an Ankom 2000 following the manufacturer's instructions. After the NDF and ADF cycles, the filter bags were washed in acetone for 5 minutes and were oven dried at 105°C for 4 h before being weighed. Lignin content was measured (sequentially) *via* acid detergent lignin (ADL) in a daisy^{II} incubator (Ankom) for 3 h, following the manufacturer's instructions. Samples were left to air dry, before

being oven dried at 105°C for 4 h; each sample was then weighed. Ash content of the samples was determined through combustion in a Carbolite CWF 1200 muffle furnace (Carbolite, Hope Valley, UK) at 525°C for 3 h. Samples were then weighed. The proportion of cellulose, hemicellulose, and lignin in each sample was calculated as in other studies (Baker, Charlton and Hale, 2019).

2.3.3 Metabolomics

Untargeted metabolomics was performed as described elsewhere (Withers *et al.*, 2020) by the West Coast Metabolomics Center, and the relative abundances of lignocellulose breakdown products was quantified. Briefly, lyophilised soil with plant litter removed were finely ground, and primary metabolites were extracted by shaking 1:0.025 (w/v) soil-to-3:3:2 (v/v/v) MeCN/IPA/H_2O solution for 5 min at 4 °C and centrifuging to recover the supernatant. Untargeted metabolomics for was conducted using ALEX-CIS GCTOF MS (automated liner exchange cold injection system gas chromatography time of flight mass spectrometry) and CSH-ESI QTOF MS/MS (complex lipid analysis by charged surface hybrid column electrospray ionization quadrupole time of flight tandem mass spectrometry) by the West Coast Metabolomics Center (UC Davis Genome Center, Davis, CA, USA). Data preprocessing was performed using ChromaTOF vs. 2.32, before validation, alignment and filtering using the using the BinBase algorithm (rtx 5). Standards were included only for quality control purposes, meaning that the data presented are qualitative and compounds are tentatively identified. This is common practice for untargeted metabolomics analysis.

Glucose was used as a proxy for cellulose breakdown. Hemicellulose breakdown products were xylose, fucose, and 3,6-anhydro-D-galactose (Van Den Brink and De Vries, 2011), and lignin breakdown products analysed were vanillic acid, 4-hydroxybenzoic acid, and benzoic acid (Bugg *et al.*, 2011; Zhu *et al.*, 2017). The compound 3,6-anhydro-D-galactose may better represent the breakdown of algal cell walls than breakdown of higher-plant cell walls (Christiansen *et al*., 2020), which are likely prevalent in these samples due to the proximity of the site to the sea.

2.3.4 Metagenomics

2.3.4.1 DNA extraction and sequencing

DNA was extracted from the frozen soil samples, following the CTAB/Phenol Chloroform-based extraction method of Griffiths *et al.* (2000), but with an additional RNAse A treatment (RNAse A added to 700 µL of sample to give a final concentration of 100 $\mu g \cdot mL^{-1}$ RNAse A) prior to the PEG precipitation step. The samples with the RNAse A were then incubated at 37°C for 30 mins, followed by a wash step with chloroform/isoamyl alcohol, as in Griffiths *et al.* (2000)). A blank sample was included to act as a negative process control.

All samples and the negative control were sent for library preparation (TruSeq Nano kit 350 bp inserts; Illumina, Cambridge, UK) and paired-end sequencing (single lane of an Illumina HiSeq 4000, 2x150 bp) at the Centre for Genomic Research, Liverpool University.

2.3.4.2 Bioinformatics

2.3.4.2.1 Sequence quality control

Sequence reads underwent quality control as follows: adapter sequences were trimmed using Cutadapt 1.2.1 (Martin, 2011) with -O 3. Sickle 1.200 was used to quality-trim the files, using a minimum window phred score of 20 (Joshi and Fass, 2011). Reads shorter than 20 bp were removed. Sequence quality was checked using fastq-stats from EAUtils (Aronesty, 2011).

2.3.4.2.2 Sequence assembly

For the assembly of all sequences in the metagenome, each library was dereplicated using prinseq-lite 0.20.4 with '-derep 1' (Schmieder and Edwards, 2011). Dereplicated reads were then co-assembled using MEGAHIT 1.1.3 using default settings (Li *et al.*, 2016). Basic assembly statistics were checked using Metaquast-5.0.0 (Mikheenko, Saveliev and Gurevich, 2016).

2.3.4.2.3 CAZyme prediction

Open reading frames and translated protein sequences were predicted from the co-assembly using Prodigal 2.6.3 using "-p meta" and "-a" options (Hyatt *et al.*, 2010). The reads for each sample were mapped back to the assembly using bowtie2 2.3.4.3 using a seed of 1 (Langmead and Salzberg, 2013). The resulting SAM files were converted to sorted BAM files using SAMtools 1.9 (Li *et al.*, 2009). Reads mapping to predicted gene sequences were counted for each sample using featureCounts 1.6.3 (Liao, Smyth and Shi, 2014), with options "-P", "-f", "-B", and "-C". Feature type counted was "CDS", and the gene identifier column was "ID". CAZy sequences in the assembly were identified using the dbCAN2 pipeline (default setting for CAZyme identification were kept, these are: HMMER coverage = 0.35, HMMER E-value = 1×10^{-15}, DIAMOND E-value = 1×10^{-102}, Hotpep Hit value = 6, Hotpep Frequency value = 2.6) where only genes which were identified by two or more of the tools are deemed to be CAZymes, and only sequences genes with a signal peptide (identified using 53ignal 6.0g) were kept for further analysis (Zhang *et al.*, 2018; Teufel *et al.*, 2022). The rule for assigning CAZy domain identity was HMMER > DIAMOND > Hotpep. Only glycoside hydrolases (GH), carbohydrate binding modules (CBM) and auxiliary activities (AA) were analysed, as these gene types should be the most involved in the decomposition of lignocellulose (henceforth 'CAZy genes') (Lynd *et al.*, 2002).

2.3.4.2.4 Taxonomic annotation of contigs

Kraken2 (Wood, Lu and Langmead, 2019), Kaiju 1.6.3 (Menzel, Ng and Krogh, 2016) and CLARK v1.2.6 (Ounit *et al.*, 2015) were used to assign taxonomy to contigs, aiming to maximise classification, searching against genomes from *Bacteria*, *Archaea*, *Protozoa*, and *Fungi*, on RefSeq release 93 (National Center for Biotechnology Information, 2018). Final taxonomic assignment followed the rule Kraken2 > Kaiju > CLARK (Piro, Matschkowski and Renard, 2017; Wood, Lu and Langmead, 2019). SAMtools faidx and BEDtools genomeCoverageBed (Quinlan and Hall, 2010)

were used to count reads mapping to each contig from each sample. Contigs found in the negative control library were removed from all analyses. Contig level counts per million (CPM), analogous to the commonly used transcriptomics metric transcripts per million—which is proportional abundance of length-scaled contigs according to average read depth for that contig, were calculated from fragments per kilobase million values from pileup.sh from bbTools (Bushnell, 2014). Contigs that could not be assigned to a phylum were not included in analyses of community composition or CAZyme origins.

2.3.5 Data analysis

R 3.5.0 (R Core Team, 2017) was used for all subsequent analyses. Significance of all models was tested using the drop1 function with either an F test for continuous or binomial data, or a χ^2 test for count data, unless otherwise specified. The effect of experimental treatment on all tested response variables was assessed using ANOVA (lm function), using the p-values from the model t-tests (summary function) to check for differences between treatments, unless otherwise specified.

2.3.5.1 Fibre analysis and total carbon

Analysis of percentage abundances of cellulose, hemicellulose and lignin between treatments was performed using Kruskal-Wallis tests followed by Dunn's test (Dunn, 1964) without p-value adjustment because of small sample size in each group.

2.3.5.2 Metabolomics

The abundance of each breakdown product was standardized to allow concurrent analysis of multiple metabolites. A general linear model (GLM) was used to estimate the relative abundance of glucose in each treatment. Hemicellulose and lignin breakdown product relative abundance in each of the treatments was estimated using generalised linear mixed effects models (GLMMs) (Brooks *et al.*, 2017), because there were multiple metabolites from each sample which could have different responses to treatment. These models used a sample-level random intercept, and the

interaction of chemical × treatment as fixed predictors. Stepwise deletion was used to determine significance of model terms.

2.3.5.3 Taxonomic and CAZyme community composition

The diversity, richness, and composition of microbial species in soils have marked effects on the flux of elements through ecosystems (Wagg *et al.*, 2019). To capture genes involved in lignocelluolysis we *a priori* chose to focus on CAZy families with high proportions of genes that have been shown to cause, or be involved in, the breakdown of specific lignocellulosic polymers; these were: (i) cellulases: GH5, GH6, GH7, GH8, GH9, GH12, GH44, GH45, GH48; (ii) xylanases: GH10, GH11, GH8, GH30; (iii) LPMOs: AA9, AA10, AA11, AA13, AA14, AA15, (iv) all CBM families, and (v) all other AA families (Nguyen *et al.*, 2018; Oates *et al.*, 2021). All subfamilies of these CAZy families were included in the analysis.

We tested for differences in the richness, diversity (Simpson's D, Shannon's H´) and dominance of species and lignocellulolytic genes using Kruskal-Wallis and Dunn's tests with unadjusted p-values due to the small sample size (Thiese, Ronna and Ott, 2016). The relative abundance of dominant species was assessed as the sum of the CPM belonging to the 500 most abundant species.

Differences in the composition of species and generally lignocellulolytic CAZy families between treatments were assessed using nonmetric multidimensional scaling (NMDS; metaMDS function from vegan) and Permutational MANOVA with 60 000 permutations (adonis2 function from vegan, after determining that between-group dispersion was not significantly different using the betadisper function with type = "median") (Oksanen *et al.*, 2008); this was also used to find relationships between CAZy gene composition and the abundance of lignocellulose breakdown products or lignocellulosic polymers. Permutational MANOVA (*Perm*MANOVA) and stepwise deletion was also used to understand the effect of soil properties (carbon, nitrogen, and

phosphorous content, N:C, P:C, total cations) on the community composition of microorganisms and CAZy genes. The envfit function was used to visualise how community composition related to predictors (significance assessed using stepwise deletion).

The effect of treatment on the abundances of individual taxonomic groups and CAZy families was assessed using GLMs on logit transformed CPM data for each genus or CAZy domain combination in a single gene. Genera and CAZymes with a significant $\log_2$ fold change greater than 1 or less than -1, relative to in the 10-year grassland treatment, were deemed to have changed in abundance.

2.4 Results and discussion

2.4.1 Transition from grassland to bare soil reduces total soil C, cellulose content, and increases the presence of lignocellulose breakdown products

Removal of plants from the soil plots altered the abundance of carbon and lignocellulosic polymers in the samples, meaning we could compare characteristics of the associated microbial community. Total carbon differed between grassland and bare treatments (ANOVA: $F_{3, 10} = 5.666$, $p = 0.016$), with the 10-year bare plots having reduced total carbon (2.54%, SD = 0.58%) relative to the two grassland treatments (10-year grassland: 3.73%, SD = 0.3%, 1-year grassland: 3.57%, SD = 0.37%), but not the 1-year bare soil (3.11%, SD = 0.32%). Cellulosic biomass was similarly affected, with the 10-year bare treatment having a reduced percentage of cellulosic biomass (median = 15%) compared to all other treatments (Kruskal-Wallis test: $\chi^2_3 = 6.843$, $p = 0.077$; Figure 1). In addition, glucose was less abundant in the 10-year bare soils than in the 1-year bare soils (GLM: $t = -2.618$, d.f. = 1, $p = 0.026$). Together these results support the hypothesis that there is an active cellulolytic microbial community, and a high rate of cellulose degradation in the 1-year bare soils, as has been shown in similar systems (Cheng *et al.*, 2007). The 10-year bare soils in contrast appear to have lost the active component of the microbial community which causes a conversion

of cellulose to glucose, possibly due to inaccessibility of this resource to a large portion of the microbial community as it becomes less abundant and more associated with complex organic molecules or minerals (Hemingway *et al.*, 2019; Lehmann *et al.*, 2020).

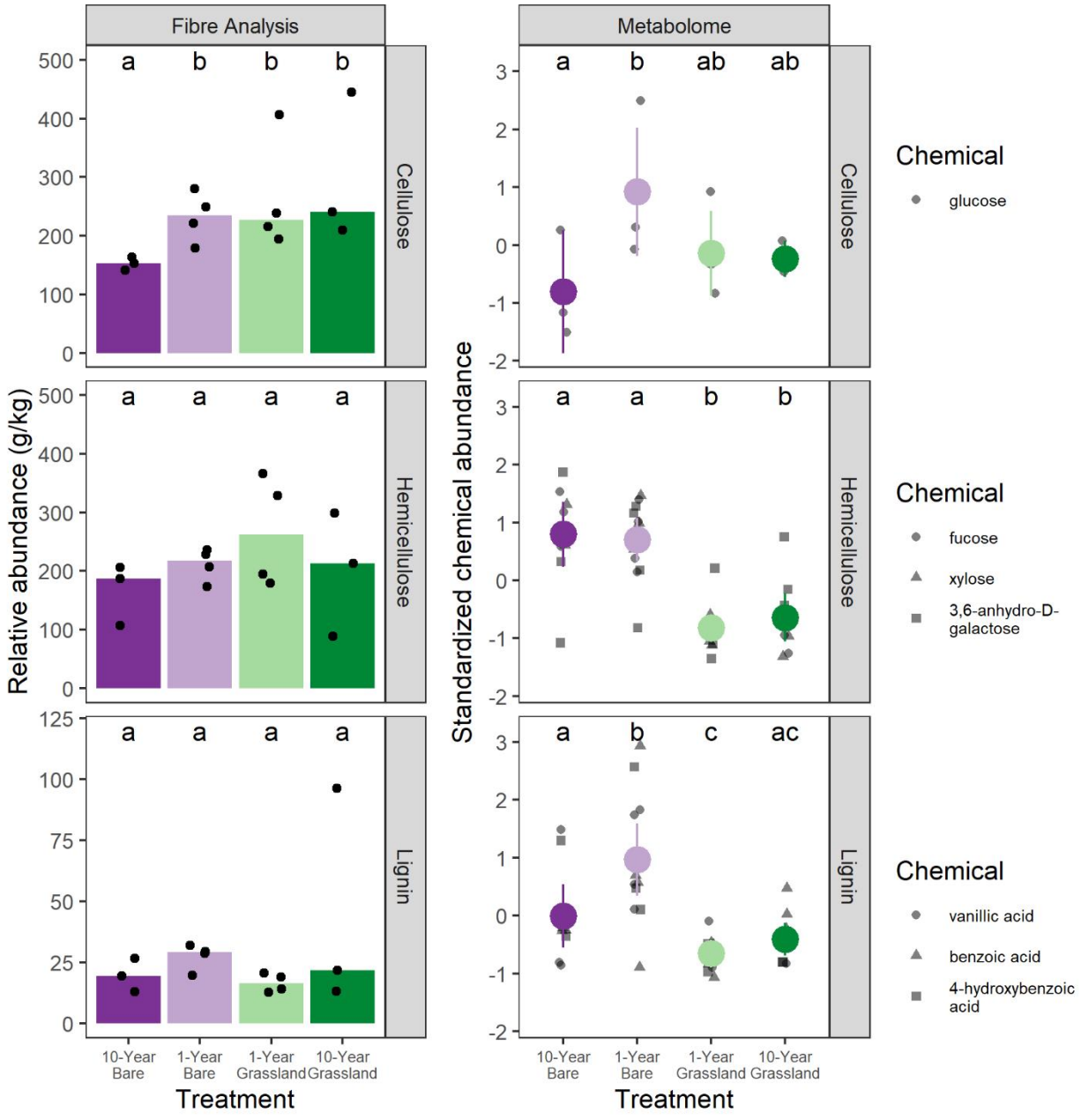

Figure 1: Abundance of lignocellulose polymers and breakdown products. Panels on the left show results from the fibre analysis procedures. Coloured bars show median values for each lignocellulosic polymer in each experimental treatment. Letters denote groupings according to the model used for testing and are relevant only within a panel. The abundance of each chemical was standardised to allow for simple comparison. Large points represent the treatment means, and the error bars represent 95% confidence intervals.

The proportion of hemicellulose and lignin was unaffected by experimental treatment (hemicellulose: Kruskal-Wallis test: $\chi^2_3 = 2.138$, $p = 0.542$; lignin: Kruskal-Wallis test: $\chi^2_3 = 5.091$, $p = 0.165$), however the abundance of lignocellulosic breakdown products shows shorter-term effects of carbon deprivation and microbial activity on the rate of degradation of hemicellulose and lignin (Figure 1). The hemicellulose breakdown products 3,6-anhydro-D-galactose, fucose, and xylose were more abundant in both bare soil treatments than in the grassland treatments (GLMM: $\chi^2_3 = 27.859$, $p < 0.001$, although 3,6-anhydro-D-galactose may show a less strong response: treatment × chemical term GLMM: $\chi^2_6 = 10.686$, $p = 0.099$), whilst the lignin breakdown products 4-hydroxybenzoic acid, benzoic acid and vanillic acid were more abundant in both bare soil treatments than in the 1-year grassland plots (quasibinomial GLMM: $\chi^2_3 = 13.404$, $p = 0.004$; Figure 1). We hypothesised that reductions in the total carbon, cellulose, and glucose content in the 10-year bare plots would substantially alter the microbial community, favouring microorganisms with the ability to utilise lignocellulosic polymers. This may be evident in the 1-year bare soil plots which had a high abundance of lignin and hemicellulose breakdown products (Figure 1), and likely represents removal of small amounts of lignin and hemicellulose as specialist cellulolytic soil microorganisms gain physical access to cellulose (Suen *et al.*, 2011). Alternatively, these results may be explained by the utilisation of polysaccharides by generalist microorganisms in response to decreased oligosaccharide availability (Brandt *et al.*, 2004; Gänzle and Follador, 2012).

2.4.2 Carbon deprivation for 10 years consistently favours *Bacillales, Thermoproteota*, and diverse *Proteobacteria*

Since the microbial community drives decomposition of lignocellulose in soils, we wanted to understand how removal of plants, and subsequent changes to the amount and types of carbon, from grassland ecosystems impacts the taxa that are present. Sequencing library statistics are given in Tables S3-4, Plant removal impacted microbial respiration and PLFA biomass as reported

in George *et al.*, (2021), with bare soils having reduced microbial respiration, and reduced biomass within each age class, although the biomass of the 1-year bare treatment was greater than in the 10-year grassland treatment. Biomass was particularly reduced in the 10-year bare plots. A significantly higher proportion of the microbial community was occupied by dominant species in each of the 10-year treatments (mean = 249000, IQR = 2300 and median = 250000, IQR = 600 CPM for bare and grassland, respectively) relative to in the 1-year treatments (median = 188000, IQR = 3200 and median = 189000, IQR = 3900 for bare and grassland, respectively; Kruskal-Wallis: χ^2_3 = 9.95, p = 0.019), potentially resulting from increased ecosystem stability enabling the growth of competitive or high-yield species which best thrive in each of the ecosystems. The reduction to microbial abundance and the high levels of dominance in the 10-year bare treatment suggest that, while plant-exclusion reduces microbial populations, 10 years of plant exclusion is a long enough period of time to reach a climax community for bare soil.

Plant removal did not affect species richness (Kruskal-Wallis: χ^2_3 = 7.75, p = 0.051) or Shannon's H´ (Kruskal-Wallis: χ^2_3 = 4.73, p = 0.192). However, plant exclusion impacted species diversity as measured by Simspon's D (Kruskal-Wallis χ^2_3 = 9.68, p = 0.022) with a marginally higher alpha diversity in the 10-year bare treatment (median = 0.988, IQR < 0.001), relative to the 1-year bare treatment (median = 0.987, IQR < 0.001) and 1-year grassland treatment (median = 0.986, IQR < 0.001). This is surprising given the higher levels of dominance in the 10-year bare soils, and perhaps this should not be over-interpreted since richness and H´ were not affected. The community composition at the species level was significantly impacted by experimental treatment (*Perm*MANOVA: $F_{3,10}$ = 2.75, p < 0.001; Figure 2). Relative to in the 10-year grassland, 97 genera (or sequences classified at a lower level) belonging to 19 phyla doubled or halved in relative abundance in one of the other treatments (Figure S1). Our analysis suggests that long-term

exclusion of plants substantially alters microbial community composition, increasing the proportion of *Thermoproteota* members at the phylum level, *Bacillales*, and *Clostridiales*, *Rhodobacterales*, *Rhizobiales, Enterobacterales* and *Alteromonadales* (Figure 3). Genera from *Thermoproteota* showed mostly consistent responses to experimental treatment, with 9 of the 10 detected genera increasing in the 10-year bare plots, reflecting abundance changes seen for this phylum in response to tillage (Nelkner *et al.*, 2019) and agricultural management in general (Zhalnina *et al.*, 2013). Whilst genera in *Euryarchaeota* responded similarly to genera in *Thermoproteota*, only a small percentage (<10%) had significant fold changes in mean abundance relative to in 10-year grassland plots. Taxa from *Thermoproteota* also showed consistent decreases in the 1-year grassland plots, suggesting that members from this entire phylum flourish in the absence of living plants due to their carbon fixing and nitrification abilities. *Nitrososphaerales*, the only group of *Thermoproteota* in which genera did not increase in the 10-year bare treatment, are atypical for *Thermoproteota*, possessing notably more CAZy genes per genome than their sister clades, including sequences with predicted hemicellulolytic capabilities which they may use to gain monosaccharides for incorporation into their extracellular polysaccharides (Könneke *et al.*, 2014; Sheridan *et al.*, 2020). The increase in *Thermoproteota* generally, coupled with the lack of increase in *Nitrososphaerales* in the bare plots (and one significant fold reduction), may be a result of reduced available carbon and nitrogen favouring ammonia oxidizers and efficient fixers of carbon, over those with genomes more targeted to biopolymer depolymerisation, or may be an effect of increased detection of rare taxa due to reduced bacterial biomass in the 10-year bare treatment. The only archaeal CAZy gene retrieved in this dataset was a GH135 gene from *Candidatus Nitrosotenuis cloacae*. GH135 has been shown by a single study to have activity against the fungal cell wall and biofilm polysaccharide galactosaminogalactan (GAG) which is used in human pathogenesis by *Aspergillus* (Speth *et al.*,

2019); reads from all 10-year bare libraries and one 1-year bare library mapped to this contig suggesting that perhaps *Archaea* in bare soils are degraders of dead fungal biomass. *Aspergillus spp*. were relatively abundant in this study with a mean of 88 CPM (SD = 10.4) across all treatments.

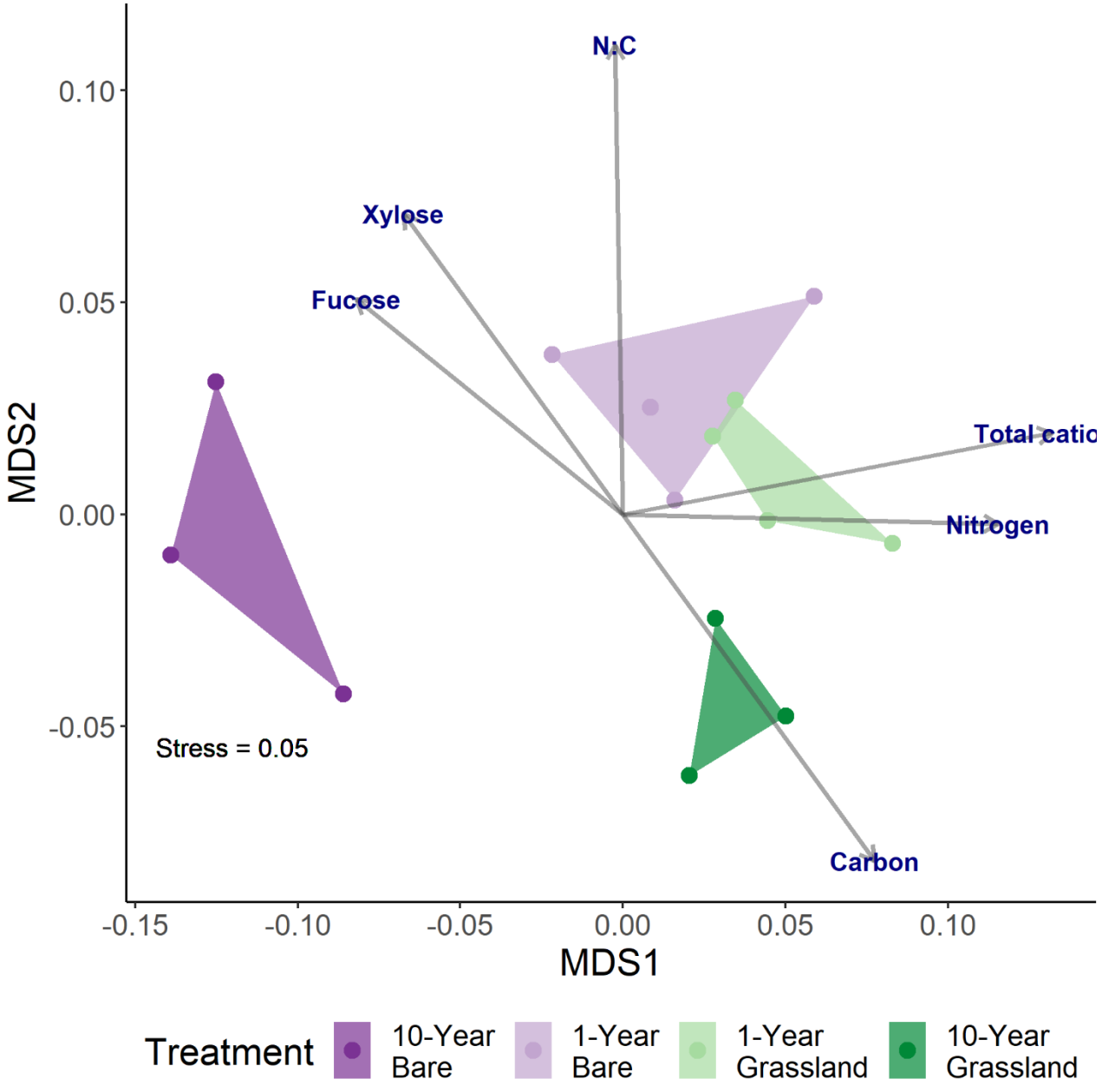

Figure 2: Effect of plant exclusion on microbial community composition measured by metagenomics, and correlations with soil chemistry. Plot shows a nonmetric multidimensional scaling of the microbial community across experimental treatments. Arrows show the direction for different soil chemicals which is maximally correlated with environmental parameters.

Bacillota (in particular *Clostridia* and the *Bacilli* families *Bacilliaceae* and *Planococcoceae*) were also consistently (15% of detected *Bacillota* genera) significantly increased in abundance in the 10-year bare plots. CAZy families detected from *Bacilliaceae* were GH18, GH3, GH13_31 and GH81, which have substrate specificities including plant, fungal and bacterial cell wall polymers, oligosaccharides, and polymers with α-glucoside linkages (Consortium, 2017). There were no CAZy genes detected in the 570 contigs (1 kbp minimum) assigned as belonging to *Planococcaceae*, and nor were there any CAZy genes detected in the 30 contigs assigned as *Clostridiales*.

Responsive genera (those with a significant $\log_2$(fold change) with magnitude larger than 1, relative to in the 10-year grassland plots) within *Proteobacteria* also mostly increased in the 10-year bare treatment, although it should be noted that only 4% of *Proteobacteria* genera showed a significant fold change. These increases were seen across *Alphaproteobacteria*, and *Enterobacteriaceae* and *Alteromonadales* within *Gammaproteobacteria*. The only CAZy gene belonging to responsive genera in *Alphaproteobacteria* was a GH23 from an unclassified member of *Acetobacteraceae*. For responsive members of *Gammaproteobacteria* only a GH27 from a contig assigned as *Phytobacter sp.* SCO41, and a GH103 from *Citrobacter sp.* were detected. These GH families all act on peptidoglycan, which may reflect that degradation of bacterial cell walls is the most profitable source of nutrients for these groups in plant-excluded soils.

The observed changes to microbial community composition between treatments suggests that broadly distributed members of *Bacillota*, *Proteobacteria*, and *Thermoproteota* are the most able taxa to capitalize on available carbon and nutrients after a long absence of plant inputs, which they may achieve directly by degradation of bacterial, fungal, or plant cell wall polymers, autotrophy, or *via* synergistic interactions with other microbiota. Unfortunately, the potential for high-resolution analysis of the different roles of specific taxa in this study is limited by sequencing

depth and assembly quality, as high quality metagenome-assembled genomes could not be retrieved from this dataset.


Genus or highest classification
Candidatus Nitrosocaldus
Nitrososphaera
Candidatus Nitrososcosmicus
Unclassified Nitrosopumilaceae
Nitrosopumilus
Candidatus Nitrosomarinus
Unclassified Thaumarchaeota
Candidatus Nitrosotenuis
Candidatus Nitrosopelagicus
Carnobacterium
Tetragenococcus
Unclassified Lactobacillales
Unclassified Alicyclobacillaceae
Virgibacillus
Oceanobacillus
Lysinibacillus
Bacillus
Aeribacillus
Unclassified Planococcaceae
Ureibacillus
Sporosarcina
Solibacillus
Rummeliibacillus
Kurthia
Jeotgalicoccus
Auricoccus
Halocella
Halanaerobium
Anoxybacter
Herbinix
Anaerobutyricum
Dehalobacterium
Paeniclostridium
Acetoanaerobium
Turicibacter
Faecalitalea
Unclassified Tissierellaceae
Sporanaerobacter
Unclassified Campylobacteraceae
Unclassified Desulfuromonadaceae
Unclassified Acetobacteraceae
Parasaccharibacter
Commensalibacter
Unclassified Brucellaceae
Unclassified Methylocystaceae
Terasakiella
Ehrlichia
Orientia
Unclassified Sinobacteraceae
Candidatus Nitrosoglobus
Fluoribacter
Candidatus Portiera
Candidatus Evansia
Reinekea
Glaciecola
Psychromonas
Parashewanella
Enterovibrio
Phytobacter
Citrobacter
Candidatus Ishikawaella
No rank
Candidatus Nitrosocaldales
Nitrososphaeraceae
Nitrososphaerales
Nitrososphaeria
Nitrosopumilaceae
Nitrosopumilales
Nitrososphaerota
No rank
Carnobacteriaceae
Enterococcaceae
Lactobacillales
Alicyclobacillaceae
Bacillaceae
Bacilli
Bacillales
Planococcaceae
Staphylococcaceae
Bacillota
Halanaerobiaceae
Halanaerobiales
Lachnospiraceae
Clostridia
Peptococcaceae
Clostridiales
Peptostreptococcaceae
Erysipelotrichaceae
Erysipelotrichales
Erysipelotrichia
Tissierellaceae
Tissierellales
Tissierellia
Campylobacteraceae
Campylobacterales
ε-Proteobacteria
Desulfuromonadaceae
Desulfuromonadales
δ-Proteobacteria
Acetobacteraceae
Rhodospirillales
Brucellaceae
Rhizobiales
α-Proteobacteria
Methylocystaceae
Anaplasmataceae
Rickettsiales
Rickettsiaceae
Sinobacteraceae
Nevskiales
Chromatiaceae
Chromatiales
Pseudomonadota
Legionellaceae
Legionellales
Halomonadaceae
Oceanospirillales
Saccharospirillaceae
Alteromonadaceae
γ-Proteobacteria
Psychromonadaceae
Alteromonadales
Shewanellaceae
Vibrionaceae
Vibrionales
Enterobacteriaceae
Enterobacterales
-5
0
5
10
15
Log2(Fold change) Relative to 10-year grassland
Treatment
10-Year Bare
1-Year Bare
1-Year Grassland

Figure 3: Significant log_2 fold changes in CPM of reads mapping to genera and unclassified sequences at the genus level, for Thermoproteota, Bacillota, and Proteobacteria. Genera are positioned based on taxonomy in published phylogeny and phylogenomic studies to aid in broader interpretation.

The richness of genes containing CAZy domains within phyla closely followed the richness of species within those phyla, with *Proteobacteria* and *Actinobacteria* having by far the highest number of genes from highly lignocellulolytic CAZy families (Figure 4), likely representing the abundance and functional diversity of these taxa, as well as database bias. *Acidobacteria* had a high richness of these genes for the number of species present, echoing results from other studies which report the ability of members of this phylum to utilise a range of polysaccharides due to many diverse CAZy genes (Ward *et al.*, 2009; Berlemont and Martiny, 2013; Kalam *et al.*, 2020).

Table 1: Marginal correlations between soil chemical properties and the composition of different sets of lignocellulolytic genes. Table shows marginal PermMANOVA results for each set of genes.

Response	Predictor(s)	F	D.F.	p
Lignocellulases	**Treatment**	3.84	3,10	**< 0.001**
Lignocellulases	**Total cations**	5.21	1,10	**<0.002**
	N:C	3.12	1,10	**0.014**
	Total carbon	2.39	1,10	**0.048**
	% Phosphorous	1.16	1,9	0.313
	% Nitrogen	0.81	1,8	0.547
	P:C	0.693	1,7	0.653
Lignocellulases	Glucose	0.76	1, 12	0.589
Lignocellulases	**Hemicellulose breakdown products**	2.5	1, 12	**0.04**
Lignocellulases	Lignin breakdown products	0.76	1, 12	0.6
Lignocellulases	Cellulose	1.4	1, 12	0.195
Lignocellulases	Hemicellulose	0.92	1, 12	0.459
Lignocellulases	Lignin	0.59	1, 12	0.697
Cellulases	**Treatment**	2.89	3,10	**< 0.001**

Cellulases	**Total cations**	5.61	1,10	**< 0.001**
	N:C	4.39	1,10	**< 0.001**
	Total carbon	3.69	1,10	**0.002**
	Cellulose	1.4	1,9	0.203
	P:C	1.41	1,8	0.213
	% Phosphorous	0.95	1,7	0.459
	% Nitrogen	1.4	1,6	0.221
Cellulases	Cellulose	1.61	1, 12	0.14
Cellulases	Glucose	0.99	1, 12	0.409
Xylanases	**Treatment**	2.3	3, 10	**0.036**
Xylanases	**Total cations**	7.12	1, 9	**< 0.001**
	N:C	4.47	1, 9	**0.008**
	Total carbon	3.61	1, 9	**0.021**
	% Nitrogen	2.86	1, 9	**0.049**
	Hemicellulose	0.88	1, 8	0.482
	P:C	-0.28	1, 7	0.998
	% Phosphorous	-0.04	1, 6	0.99
Xylanases	Hemicellulose	0.919	1, 12	0.455
Xylanases	**Hemicellulose breakdown products**	3.78	1, 12	**0.014**
Xylanases	**Fucose**	3.68	1, 12	**0.016**
Xylanases	Xylose	2.79	1, 12	0.05
Xylanases	3,6-anhydro-D-Galactose	1.83	1, 12	0.152
Auxiliary Activities	**Treatment**	7.47	3,10	**< 0.001**
Auxiliary Activities	**Total cations**	15.16	1, 9	**< 0.001**
	N:C	2.9	1, 9	0.075
	P:C	0.625	1, 9	0.545
	% Nitrogen	0.75	1, 9	0.462
	Phosphorous	0.65	1, 8	0.524
	Lignin	0.11	1, 7	0.972
	Total carbon	0.27	1, 6	0.857
Auxiliary Activities	Lignin	0.37	1, 12	0.787

Auxiliary Activities	Lignin breakdown products	0.52	1,12	0.62
Auxiliary Activities	Vanillic acid	0.68	1, 12	0.506
Auxiliary Activities	4-Hydroxybenzoic acid	0.8	1, 12	0.445
Auxiliary Activities	Benzoic acid	0.07	1, 12	0.987

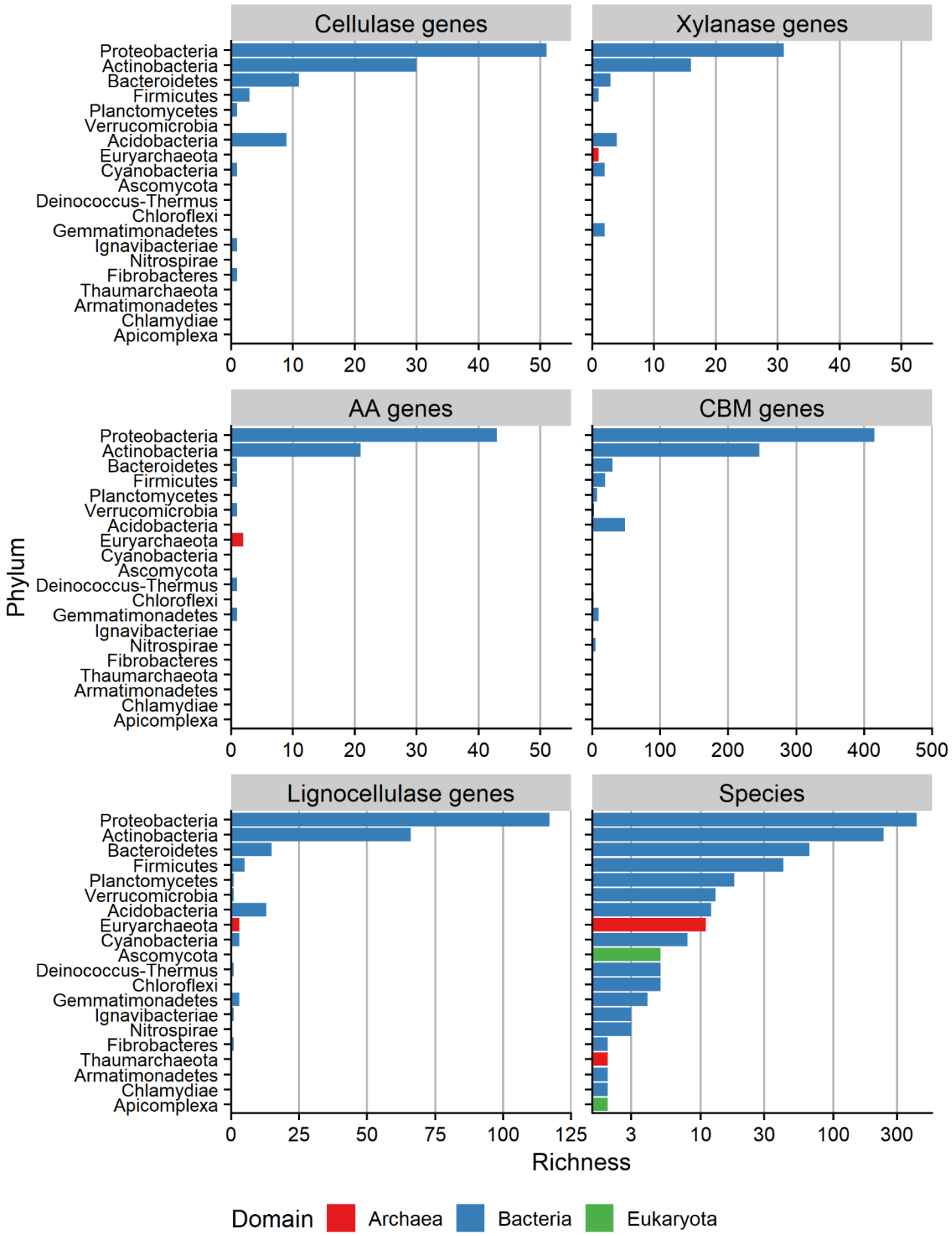

Figure 4: Richness of species within phyla and genes from extracellular CAZy families with high proportions of plant cell wall degrading activities, within phyla in the soil samples. Species richness values had 1 added to them so that the x axis could be placed onto a log10 scale to better show differences in species richness.

2.4.3 Plant exclusion substantially alters the composition of genes with lignocellulolytic potential

The different plant inputs due to experimental treatment affected the abundance, richness, diversity and composition of lignocellulolytic genes in the soil (Figure 2, Table 1), which may lead to functional differences in the breakdown of lignocellulose between treatments. Assuming the activity profiles of these gene families are well known, differences in composition of these genes should give insights into the carbon acquisition strategies which are beneficial to microorganisms depending on the quality and quantity of carbon inputs in each experimental treatment.

There were no differences in the abundances of lignocellulase genes overall (Kruskal-Wallis: χ^2_3 = 3.80, p = 0.284), cellulase genes (Kruskal-Wallis: χ^2_3 = 0.50, p = 0.919) or xylanase genes (Kruskal-Wallis: χ^2_3 = 0.72, p = 0.868), however, auxiliary activity genes were more abundant in the 10-year bare treatment than in both 1-year treatments (Dunn's test p = 0.015, 0.015 respectively). There was a smaller diversity and richness of cellulase genes in the 10-year bare soils, relative to in the 1-year grassland soils (Dunn's test on richness: p = 0.022, Dunn's test on Shannon's H′: p = 0.037, Dunn's test on Simpson's D: p = 0.054). As the more enzymatically accessible (soluble) amorphous cellulose from fresh plant inputs is utilised by less non-specialist cellulose degraders in the 10-year bare treatment, there may be a selection pressure for species which utilise the less enzymatically accessible crystalline cellulose; the 10-year bare soils were more associated with the largely cellulolytic GH5 subfamilies 1 (linked to CBM2), 5, 25, 46 (linked and not linked to CBM6 domains), as well as subfamilies with more diverse enzyme activities which were subfamilies 36 (EC 3.2.1.78), 13 (EC 3.2.1.146 and 3.2.1.55), 19 (EC 3.2.1.100 and 3.2.1.25), relative to the other treatments (Figure 2, Table 1), supporting this hypothesis.

By contrast, xylanase genes had increased richness and Shannon diversity in the soil after 1 year of plant exclusion (Dunn's test p = 0.005), but after 10 years there were no detectable differences

from the 10-year grassland plots (Dunn's test p = 0.414). Additionally, the Shannon diversity of xylanase genes was greater in the 1-year bare plots (Dunn's test p = 0.005) than in the 10-year bare plots, although Simpson's diversity was unaffected by treatment (Kruskal-Wallis χ^2_3 = 0.369). Similar patterns were observed for AA genes with the 10-year bare treatment being less rich and diverse (D) than the 1-year bare (Dunn's test for richness: p = 0.021; Dunn's test for Simpson's D: p = 0.019) and 1-year grassland treatments (Dunn's test for richness: p = 0.033; Dunn's test for Simpson's D: p = 0.023).

Increased diversity and richness of xylanase and AA genes after 1-year of plant exclusion suggests that the lack of plant inputs favours microorganisms with diverse hemicellulolytic and ligninolytic capabilities, but that after 10 years this increased diversity is no longer as beneficial a life-history strategy. Similar results have been found in forest soils following afforestation, although with differing timescales for the increase in lignocellulolytic gene abundance. In the present system, plant exclusion for 1 year increased the diversity of xylanases, and abundance of auxiliary activities increased after 10, whereas afforested soils had a peak in GH and AA gene abundance at 20 years (Ren *et al.*, 2021); these differences may be attributable to the differing qualities of the plant inputs, with the ratio of lignin to polysaccharide based polymers being the controlling factor.

The composition of xylanases and auxiliary activities were also affected by treatment (Figure 6, Table 1); the 10-year bare plots were associated with GH30 and GH30 subfamilies 1 and 3 which have known activities on xylan and as endoglucanases, and AA10|CBM73 which are lytic polysaccharide monooxygenases with associated chitin-binding activities. The other treatments were more associated with GH8, GH11, GH10, and GH30_2 which are gene families with many recorded xylanolytic activities, and AA3, AA3_2, AA5, AA6, AA7, AA10, and AA12 which have each

been shown to play a role in the degradation of lignin, reflecting the diversity of gene families in the other treatments.

To add further insights, we analysed the taxonomic origins of the lignocellulase genes that were associated with different treatments (Figure 6). The 21 lignocellulolytic CAZy (sub)families associated with the 10-year bare plots came from 168 contigs, belonging to 162 different taxonomic classifications at different ranks, highlighting the diversity of microbial species which are likely involved in lignocellulose degradation in soils. The majority of these genes were annotated as members of the *Proteobacteria*, *Actinobacteria* and *Bacteroidota*, and *Acidobacteria* which is not unexpected as multiple soil metagenome studies have found these to be the dominant lignocellulose degraders (Figure 6, Table 1) (Wilhelm *et al.*, 2019). Cellulase CAZy families associated with the 10-year bare treatment had taxonomic origins from *Acidobacteria* (five species), *Actinobacteria* (four species, and a sequence which could only be assigned taxonomy at phylum level), *Bacteroidota* (five species), *Planctomycetes* (*Gemmata sp.* SH-PL17), and *Proteobacteria* (mostly from species within *Rhizobiales*). Xylanase gene families associated with the 10-year bare treatment came from species of *Acidobacteria* (*Candidatus Koribacter versatilis*), *Actinobacteria* (*Actinoplanes* and *Streptomyces*), *Bacteroidota* (*three species*), *Bacillota* (*Clostridium sp.*), and diverse *Proteobacteria* (*Alpha-*, *Beta-*, *Gamma-*, and *Deltaproteobacteria* classes). The single auxiliary activity distinguishing the 10-year bare treatment from the other treatments (AA10|CBM73) could not be associated to a taxon. Species with 10-year bare associated CAZy families generally did not respond to experimental treatment (Figure S1, Table S2), demonstrating that changes to genetic composition, and so functionality, of the soil cannot be easily estimated from taxonomy-based shifts. Two species of *Acidobacteria*, *Candidatus Koribacter versatilis* and *Luteitalea pratensis*, were both significantly responsive in abundance to

experimental treatment; both had GH5_13 genes, *Candidatus Koribacter versatilis* had GH30 genes, and *Luteitalea pratensis* had GH10 genes, suggesting that *Acidobacteria* play a large role in shaping the composition of lignocellulolytic genes in soil.

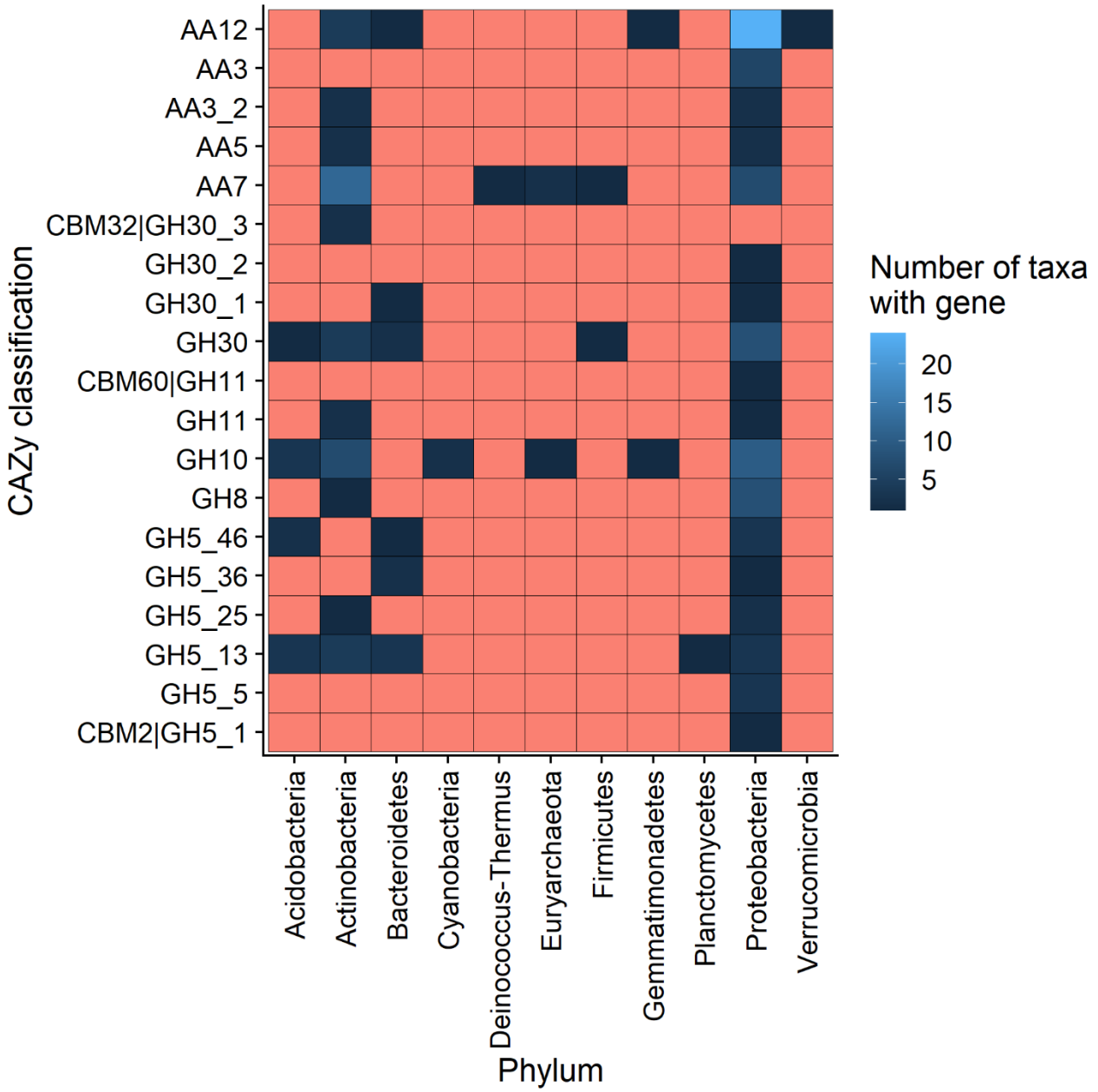

Figure 5: Phylum-level taxonomic origins of lignocellulolytic CAZy gene (sub)families which were associated with the 10-year bare treatment. Number of taxa refers to number of contigs with a taxonomic classification at any rank. Salmon coloured squares indicate that no gene was found for that taxon.

2.4.4 Drivers of lignocellulolytic gene composition

To explore the strength of the relationships between genetic potential of the microbial community and observable function, we checked for associations between lignocellulolytic gene composition and the abundance of cations, carbon, nitrogen, phosphorus, lignocellulosic polymers that the gene-products have known activities on, as well as products from the breakdown of these polymers (Figure 6, Table 1).

The composition of lignocellulolytic genes was associated with changes in total carbon and total cations, both of which are well known to determine microbial community composition (Q. Zheng *et al.*, 2019), suggesting bottom-up control over the lignocellulolytic genes present. Increased concentrations of monovalent cations can increase SOM solubility, which may be another mechanism by which cations influence the composition of lignocellulase genes in this study (Curtin, Peterson and Anderson, 2016). Variation within sites was best predicted by NC ratio (Figure 6, Table 1) and may reflect how the enzymatic toolkit of the microbial community responds to local availability of nitrogen, determining finer scale changes in microbial community composition. Electrical conductivity and C:N have been shown to have major influences on CAZy gene composition in saline soils (Chao Yang *et al.*, 2021), adding further support to these findings. Similar patterns were observed for the sub-groups of cellulase, xylanase, and auxiliary activity genes, although xylanases were also related to the nitrogen content of the soil, and total cations was the only significant predictor of auxiliary activity gene composition.

Lignocellulolytic gene composition (*i.e.*, the relationship between gene relative abundances) was correlated with the abundance of hemicellulose breakdown products, as was the composition of xylanase genes (Figure 6). Xylanase gene composition was related to fucose, but not percentage hemicellulose, xylose, or 3,6-anhydro-D-galactose content of the soil. Relationships with relevant

polymers and breakdown products were not found in relation to the relative abundances of cellulase, or auxiliary activity, gene families. The strength of the relationship between both lignocellulases generally and xylanases with hemicellulose breakdown products may be a result of relatively fast degradation rates of hemicellulose and relatively little metabolism of their breakdown products by the soil microbial community. However, xylose has been shown to be rapidly utilised by *Bacteroidota*, *Bacillota*, and *Proteobacteria* (Pepe-Ranney *et al.*, 2016), all of which had multiple members significantly responding in abundance to experimental treatment in this study. The lack of relationships between lignocellulolytic gene composition and lignocellulosic polymer content of the soil was surprising and may reflect the difference in turnover rates between the genetic content of the bare soil microbial communities, and the processes that they mediate, or otherwise may be a product of the diverse metabolic capabilities of the soil microbial community and the population dynamics of species from different functional groups.

Further research is needed to better understand the relationships between the content of lignocellulolytic genes and soil carbon turnover. Our finding of a relationship between the lignocellulolytic gene community and hemicellulose breakdown product abundance could lead the way to predictive modelling of hemicellulose breakdown across soil ecosystems. The lack of other relationships of this sort, however, are both unexpected and interesting; perhaps a similar experiment with increased statistical power could find relationships between these variables, which could lead to the production of exciting gene-based carbon turnover models.

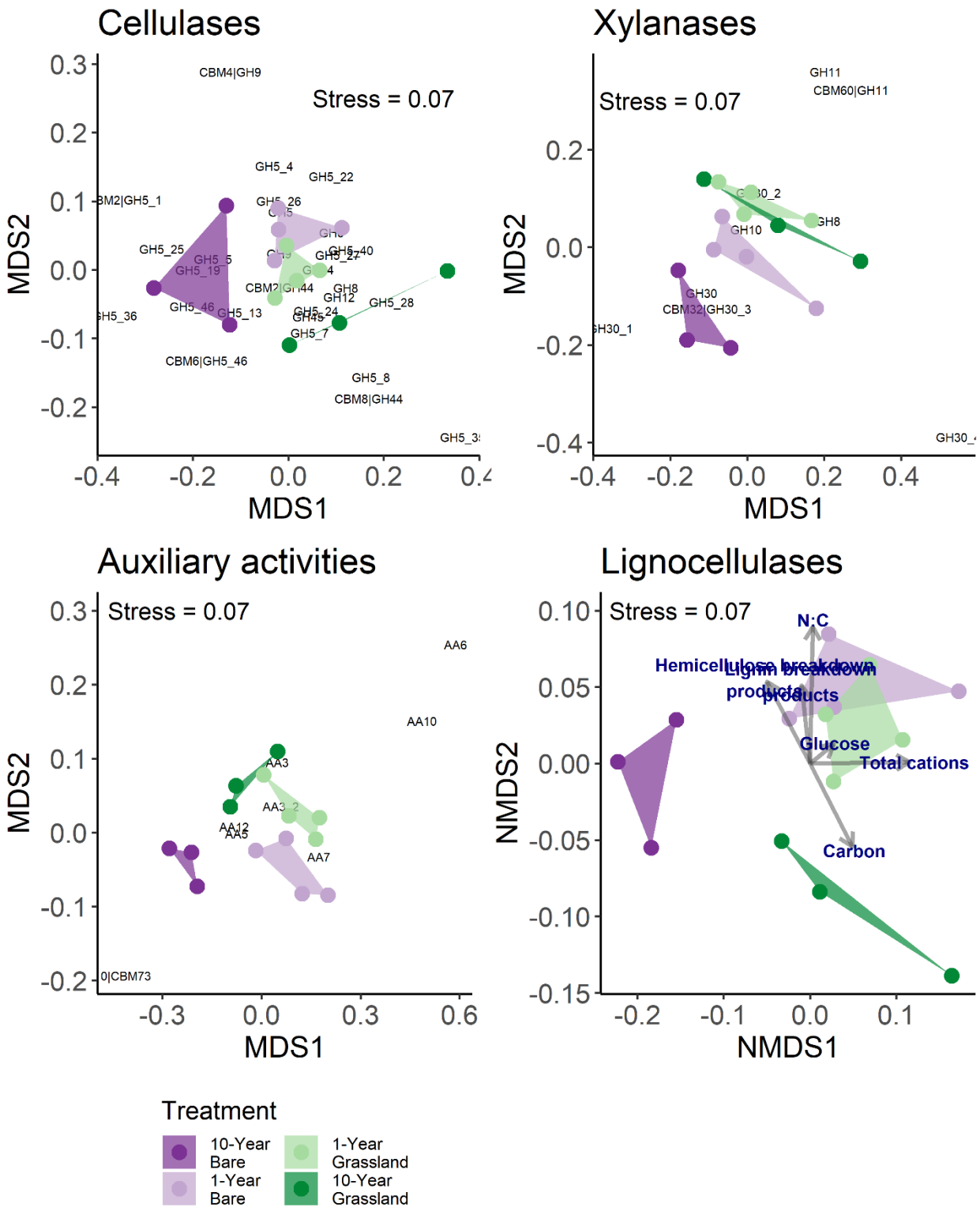

Figure 6: Nonmetric multidimensional scaling of the community of CAZy families with a high proportion of lignocellulolytic activities across treatments. Arrows show the direction which is maximally correlated with environmental parameters.

2.4.5 Conclusions

Long-term reductions to carbon substrates due to experimental treatment (Figure 1) decreased soil microbial biomass (George *et al.*, 2021), consistently increasing the abundance of genera within *Bacillales*, *Thermoproteota*, and diverse lineages of *Proteobacteria*, alongside idiosyncratic taxon-specific changes (Figures 2, 3, S1). The reductions to microbial biomass may be partly responsible for the consistent broadly taxonomically distributed changes we have measured, however, the equivalent levels of microbial dominance between bare and vegetated treatments in each of the age classes suggests that this is not a major reason for the differences that we detected. One year of plant-exclusion may not be enough time to see changes to the abundance of carbon, and lignocellulosic polymers, giving few changes to the composition of specific taxa, low microbial dominance, and no change to the composition of the microbial community. Similarly, plant-exclusion reduced the diversity of cellulase genes, and impacted the composition of lignocellulolytic genes, possibly leading to fewer genes with xylanolytic activities (Figure 5, Table 1). We found increased xylanase diversity after a single year of plant exclusion which may represent subtle community shifts towards species with the ability to utilise hemicelluloses, reflecting the increased relative profitability of dead plant material as a source of carbon and other nutrients. This concurs with the increased abundance of hemicellulose breakdown products in the 1-year bare treatment. We found carbon content, N:C, cation content, and nitrogen to be significant predictors of soil lignocellulolytic gene composition. Additionally, we found correlations between gene composition and the abundance of hemicellulose breakdown products, further suggesting active breakdown resulting from the shift in hemicellulase gene composition (Figure 5). In general, lignocellulolytic gene families associated with long-term plant exclusion were mostly not associated with taxa which were significantly responsive to treatment, although there were

two exceptions from *Acidobacteria* which may be a result of their abundance in soils and their high genomic glycoside hydrolase gene content (Figure S1, Table S2).

In this study, we fill knowledge gaps about the fundamental and realised niches of broad phylogenetic groups of soil microorganisms, gain insight into the temporal scale at which the microbial community responds to changing carbon inputs and how this relates to changes in the carbon content of the soil, describe and analyse how changes to microbial community composition relates to changes in the composition of genes in soil and highlight specific taxa which may be influential in shaping changes to the functioning of one of the largest fluxes of carbon on Earth.

2.5 Supplementary information

Figure S1

https://github.com/fidlerdb/Plant_exclusion_experiment_lignocellulase_genes/blob/ea8594cc44f949e27fca76aa5c06bda22af9e68e/Species_changes_2022-05-10_supplementary.pdf

Table S2

https://github.com/fidlerdb/Plant_exclusion_experiment_lignocellulase_genes/blob/15bdf28211dd3763e49f0fed9c896f18b1c56d69/CAZy_gene_PermMANOVA_results.csv

Table S3

Sample	Total number of reads (pre-quality control)	Toal number of reads (post-quality control)	Number of unpaired reads	Mean phred score ± SD	
				Forward read	Reverse read
c1	25599822	25026515	98591	39.0 ± 4.2	38.9 ± 4.6
c2	40423536	40058096	308146	39.0 ± 4.3	38.3 ± 5.5
c3	39824734	39355430	356814	39.0 ± 4.3	38.1 ± 5.7
c4	44174930	43824259	306967	39.1 ± 4.2	38.3 ± 5.4
c5	41863660	41533356	260384	39.0 ± 4.3	38.4 ± 5.3
c6	37902940	37646088	213910	39.0 ± 4.2	38.5 ± 5.1
c7	34485136	33998556	290584	39.0 ± 4.3	38.0 ± 5.8

b1	34443578	34151212	248010	39.1 ± 4.2	38.3 ± 5.4
b2	39414296	39109814	280080	39.0 ± 4.3	38.2 ± 5.5
b3	45395230	44911778	462252	39.0 ± 4.3	38.0 ± 5.8
b4	39524656	39133275	331003	39.0 ± 4.2	38.2 ± 5.6
b5	38993776	38673647	274305	39.0 ± 5.2	38.3 ± 5.4
b6	43317956	42870965	303901	39.0 ± 4.3	38.3 ± 5.4
b7	37423404	37124607	239113	39.0 ± 4.2	38.4 ± 5.3
n1	338388	22095	409	39.2 ± 4.0	39.3 ± 4.1

Table S4

Assembly	Statistic
# contigs (>= 0 bp)	6770073
# contigs (>= 1000 bp)	1036068
# contigs (>= 5000 bp)	34606
# contigs (>= 10000 bp)	6876
# contigs (>= 25000 bp)	718
# contigs (>= 50000 bp)	148
Total length (>= 0 bp)	4913859402
Total length (>= 1000 bp)	1977906684
Total length (>= 5000 bp)	299313231
Total length (>= 10000 bp)	116803116
Total length (>= 25000 bp)	30940139
Total length (>= 50000 bp)	12094433
# contigs	3526896
Largest contig	375786
Total length	3681985028
GC (%)	63.41
N50	1073
N75	710
L50	903767
L75	1977586
# N's per 100 kbp	0

3

Agricultural intensification alters grassland soil microbial community structure and increases lignocellulase gene relative abundance, but does not benefit lignocellulolytic microorganisms

3.1 Abstract

The balance of carbon accrual and mineralization in soils globally is a critical regulator of ecosystem services such as climatic stability and food production potential. Agricultural land-use intensification is widespread and reduces soil organic matter content, damaging the long-term sustainability of the services that soils provide. Little is known about how agricultural intensification alters the utilisation of plant cell wall polymers by soil microorganisms, and thus how anthropogenic activities interfere with degradative pathways which regulate important global C pools. We utilised soil metagenomes from six replicated land use contrasts to investigate how agricultural intensification affects the soil microbiome and specific degrader communities. Extensive land use was associated with increased dominance of phylogenetically constrained bacterial taxa, and reduction of this dominance in intensively managed soils was associated with apparent increased relative abundance of over 50% of species from diverse phyla. Land use altered the composition of the lignocellulase gene pool, and gave an apparent 20% increase in cellulase gene relative abundance in arable soils. However, few species which significantly increased in relative abundance in intensive soils possessed (ligno)cellulase genes. Additionally, high within-species metagenomic lignocellulase gene abundance was correlated with extensive grassland-associated taxa. We therefore suggest that the decreased abundance of dominant non-

lignocellulose degrading taxa and a resilient microbial lignocellulolytic community are responsible for the apparent increased lignocellulase content in arable soils, but that lignocellulolytic species may be competitive in soils with large native organic carbon pools. In conclusion, this study demonstrates how an improved functional and taxonomic understanding of the soil microbiome can enhance our mechanistic understanding of soil organic matter dynamics, which is crucial for improving management of Earth's ecosystem services with a growing population.

3.2 Introduction

The land area dedicated to agriculture globally has increased by almost 2 million km^2 since 1960. Human usage of the planet's soil resources is unsustainable; farmland occupies 38% of the global land area, and nearly 30% of Earth's net primary productivity is utilised by humans, yet the demand for agricultural products is predicted to double by 2050. Intensive agricultural practices are well known to threaten biodiversity, contribute significantly to climatic changes, and cause loss of multiple ecosystem services (Tilman *et al.*, 2011; Zabel *et al.*, 2019; Winkler *et al.*, 2021). One mechanism by which agriculture causes global changes is through loss of soil organic carbon (SOC), due to factors relating to plant cover and residue removal and soil disturbance (tillage) (Wuaden *et al.*, 2020). There is now a widespread need to better understand soil processes and the influence of agriculture since soil carbon underpins many critical ecosystem services, including climatic stability, food security and productivity, and terrestrial biodiversity (FAO, 2020; FAO and ITPS, 2021).

Changes in land use from pasture to crop reduce soil carbon by roughly 60% (Guo and Gifford, 2002), and in the UK, croplands lose carbon at a rate of 140 ± 100 kg of carbon per hectare per year, relative to the rate of sequestration of 240 ± 200 kg of carbon per hectare per year by grasslands (Ostle *et al.*, 2009). An outcome of these soil carbon fluxes is that arable soils have the lowest SOC content of any broad habitat type (Ostle *et al.*, 2009). In contrast, extensive management (lower productivity agricultural land which can include low-density grazing) of grasslands can promote SOC accumulation, and support high diversities of plants and animals, whereas intensive management of grasslands (characterised by high tillage frequency, nutrient addition, cropping in multiple seasons), such as in arable systems, reduces meso- and macrofaunal and plant diversity, network complexity, and ecosystem functionality (Tsiafouli *et al.*, 2015).

Intensive agriculture strongly reduces microbial biomass (de Vries *et al.*, 2013; Sun *et al.*, 2016; Malik *et al.*, 2018), however, the effects on microbial diversity noted in the literature are variable, with studies showing decreased (Guo *et al.*, 2020), unchanged (de Graaff *et al.*, 2019; van Rijssel *et al.*, 2022), and increased (Delgado-Baquerizo, Maestre, *et al.*, 2016; George *et al.*, 2019; Romdhane *et al.*, 2022) soil microbial diversity. Increases to diversity may be related partly to increases in soil pH due to management practices as demonstrated by other national- and continent-scale studies (Griffiths *et al.*, 2011, 2016; Karimi *et al.*, 2018). Perturbances to soil ecosystems can increase the relative abundance of particular microbial taxa, decreasing diversity and evenness; this has been a common theme in soil microbial community research (Qiu *et al.*, 2021). A generalisable understanding of the response of soil microbial diversity to agricultural intensification, and the reasons for each response type, is required to understand the functioning of natural and agroecosystems.

Microbial composition is also strongly affected by land use, with broad groups of microorganisms showing large differences in their relative abundances. As well as SOC content, the ratio of carbon to nitrogen (C:N) of soils and pH represent the major environmental controls over soil microbial community composition (Griffiths *et al.*, 2011; George *et al.*, 2019). Intensive agriculture has only a small negative effect of C:N (associated with poor soil quality) (Kopittke *et al.*, 2017), while pH is often increased in agroecosystems due to liming practices. Farming practices may benefit species with copiotrophic and stress tolerant life-histories (Malik *et al.*, 2018), possibly because of increased nutrient inputs and increased disturbance. Archaea, by contrast, have greater species richness and abundance in less productive ecosystems, possibly as they are mostly adapted to stable environments and generally follow oligotrophic life-history strategies (George *et al.*, 2019). Arable soils, however, are associated with high abundances of *Nitrososphaeria*, and cropped or bare land may promote the abundance of particular groups of *Archaea*; the exact mechanisms which control archaeal diversity and abundance require further investigation (Karimi *et al.*, 2018; Korzhenkov *et al.*, 2019; Trivedi *et al.*, 2019; Armbruster *et al.*, 2021; Saghaï *et al.*, 2022). The reported increases to diversity and evenness in agricultural systems may well be an artefact of measuring diversity using compositional data. If dominant taxa are strongly negatively affected by agricultural land use intensification, then compositional data (such as DNA sequencing data) should show increased species richness and community evenness in agricultural soils as more rare taxa and more relic DNA strands are measured per unit effort of sampling—especially when

microbial biomass overall is strongly reduced by intensive management practices (Carini *et al.*, 2016; Griffiths *et al.*, 2016).

High-throughput sequencing of metagenomic DNA allows characterisation of the functional potential of soil communities, moving the capability of scientists beyond prediction of the taxonomic composition of soil. This, and other techniques have shown that the genetic composition of grasslands can be altered by land use. Agriculture did not affect the relative abundance of clusters of orthologous genes (COG) involved in carbohydrate metabolism in Argentinean Pampas soil, although the genes in COG categories related to intracellular trafficking and secretion, amino acid transport and metabolism, and energy production and conversion increase in conventionally tilled soils, suggesting again that copiotrophs are adapted to these disturbed ecosystems with additional nutrient inputs (Carbonetto *et al.*, 2014). The relative abundance of different gene classes between grasslands and wheat farms in Sweden found decreased abundance of the GH5 and GH7 (associated with cellulose hydrolysis), and AA9 lytic polysaccharide monooxygenase (LPMO) genes in arable (Manoharan *et al.*, 2017), suggesting bias in utilisation of compounds which are more energetically profitable than lignocellulosic polymers in wheat fields. In contrast, conventional tillage and crop rotation management practices can increase the abundance of carbohydrate metabolism-related gene fragments (Souza *et al.*, 2015), suggesting the opposite. Once again, the compositional nature of these results and the relationship with actual abundance must be taken into consideration when drawing ecological conclusions about the genetic basis for ecosystem functioning.

Direct measurements show that land-use alters soil functionality: more intensive agricultural practices reduce soil respiration, give reduced community-level enzymatic function, and reduce litter decomposition rates (Bielińska and Mocek-Płóciniak, 2012; Lienhard *et al.*, 2013; de Graaff *et al.*, 2019). Interestingly, the priming effect from addition of ground wheat stubble was significantly stronger in improved soil systems than in pastures in soils in Laos, possibly reflecting the selection pressure for copiotrophic microorganisms, which may include lignocellulolytic species, in agricultural systems (Lienhard *et al.*, 2013).

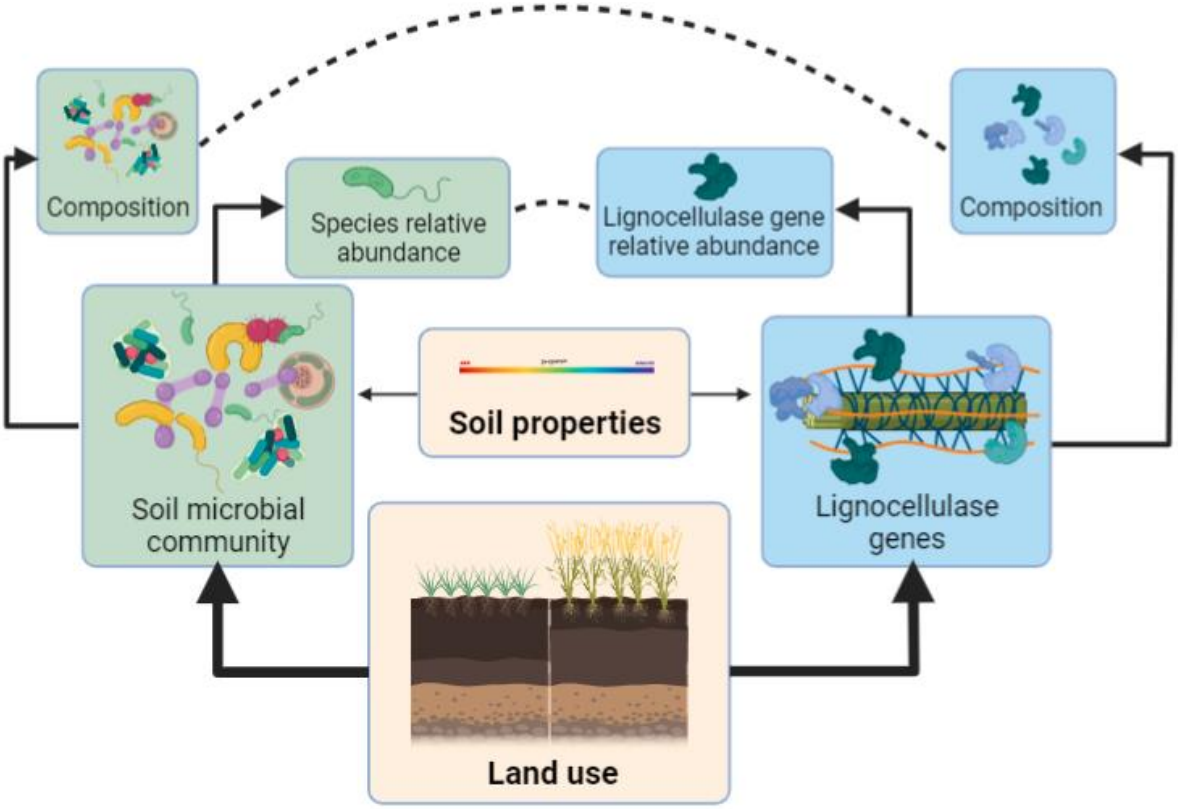

Figure 1: Conceptual overview of the relationships observed in this study. Here, we investigate land use impacts on the composition of species and lignocellulase genes, focus on changes to particular species and gene families, and investigate how land use affects relationships between functional groupings of lignocellulase genes and soil properties. N.B. The arrows do not show causal pathways in all cases, but rather the relationships investigated. Additionally, we show the enzymes produced by the genes for lignocellulases, but this does not mean we are measuring relative abundances of these proteins in the soils.

Here we investigate how the soil microbial community and associated lignocellulolytic genes (defined here as belonging to one of several gene families, with a signal peptide denoting exocellular nature of the produced protein) are related to land use change and to soil properties (Figure 1). We compare the microbial and genetic composition of the soil from extensive grasslands and intensive arable fields across six sites in England and Scotland. Based on the findings of other studies, we expect to show that land use intensification will increase soil microbial diversity, and will thus affect the origins of lignocellulolytic genes. We hypothesise that the genetic potential for lignocellulose utilisation will be reduced by intensive agricultural practices which should allow microorganisms to utilise the more freely available N and C from organic fertilizers and native forms of these elements which are released due to tillage (loss of physical protection), instead of investing in complex degradative enzyme systems with high energetic production costs. We expect that this will also be expressed as metagenomic DNA in extensive grasslands having a higher proportion of species with lignocellulase genes, and when these genes are present, higher average lignocellulase gene content than in arable soils. Testing

these hypotheses will begin to fill a key knowledge gap about the functional capacity and life-histories of microorganisms associated with different land uses. Finally, we expect that land use intensification will alter the relationships between lignocellulolytic gene relative abundance and environmental parameters, owing to the interaction between hypothesised life-history strategies (lignocellulolysis for carbon and nutrient gain in extensive grasslands *versus* copiotrophic utilisation of simple compounds in arable soils) and niche spaces of different soil microorganisms.

3.3 Methods

3.3.1 Experimental design and soil sampling

This study utilised existing assembled metagenomic DNA sequences obtained from the Soil Security programme's UGRASS project, which was first reported by Malik *et al.* (2018). The focus of UGRASS is to understand how land-use intensification affects the microbial community and ecosystem functioning of grasslands, using ten sites with paired grassland land uses (extensive grassland, intensive grassland, arable grassland, bare fallow) distributed across England and Scotland. Each site has a 'pristine' extensively managed grassland field adjacent to at least one field from the other land-use categories. In this study we focussed only on the extensive grassland and arable samples, limiting this study to data from six sites. Extensively managed grasslands in this study had a mix of managements which were established at the end of the 1940s, including pristine grassland, no-till and no-cut, low density sheep and horse grazing, cutting twice annually, and ridge and furrow, and reseeded grassland. Arable fields in this study were similarly long-standing, being established in between 1949 and 1959 with managements including cultivation of arable rotation (wheat and oats) with fallow, and continuous arable on rotation with and without cover crop with legumes. All arable fields had tillage, liming, NPK fertiliser addition, or a combination of these.

The dataset consists of ninety-six samples (32 for the land use contrast), with three to four soil samples per site per land use type. Each sample (5 cm diameter soil cores, with 2 m between samples) is associated with elemental, physiochemical, and biological (e.g., total carbon, organic carbon, nitrogen, phosphorus, pH, soil moisture, bulk density, bacterial cells per gram dry soil) data, as well as metagenomic DNA from the soils; samples were collected between April and August 2015.

Metagenomic DNA was extracted from 2 g of soil using the MoBio power max soil DNA isolation soil kit, and was subsequently purified using a Millipore amplicon ultra buffer exchange. The purified DNA was sequenced on an Illumina HiSeq 4000 using Illumina TruSeq libraries (insert size < 500 – 600 bp). Paired-end sequencing (2 x 150 bp) on 96 indexed libraries were multiplexed across 8 lanes and generated over 280 M clusters per lane.

3.3.2 Bioinformatic processing

Metagenomic reads underwent bioinformatic pre-processing, Illumina adaptor sequences were removed using Cutadapt 1.2.1, reads were then trimmed with Sickle 1.200 with a minimum window quality score of 20. Reads shorter than 20 bp after trimming were discarded. Metagenomic reads were then co-assembled in blocks of sites with similar soil characteristics using MEGAHIT with a minimum contig length of 1000 bp to maximise the contiguity of the DNA. The taxonomic classification of contigs from the assemblies was determined using the unique k-mer based kraken2 v2.1.2 with default settings. The reference database for kraken2 was constructed from RefSeq genomes (downloaded 2021-06-02) from bacterial, archaeal, and fungal clades This taxonomic assignment method will likely give ecologically nonsensical species-level classifications because the RefSeq database is relatively small and does not capture the true diversity of soil microorganisms. Species assignments where this is the case likely show closely related microorganisms from similar clades (and the likelihood of misasignment will depend on how well categorised each clade is), however, there is little guarantee of this where “unique” to the RefSeq database sections of DNA for a species are shared by wider members of the microbial community. Further, the assembly may include chimeric contigs. However, we believe that imperfect classification of contigs is a useful tool for helping to understand the vast complexity of microbial communities. Gene and contig abundance in each sample was calculated following the equation for transcripts per million (TPM) which corrects for number of mapped reads and contig length (Wagner, Kin and Lynch, 2012), giving the relative abundance metric counts-per million (CPM). This allows for the comparison of the abundance of nucleotides mapping to different contigs between samples.

The relative abundance of nucleotides mapping to each gene and contig was calculated using the following pipeline: bowtie2 version 2.3.4.3 with settings -q –sensitive, on 8 threads with a random seed of 1 was used to count mapped reads in each sample. The SAM files produced were converted to sorted BAM files using samtools (version 1.9) view -S with 4 threads. Exact duplicate

reads were removed with picard 2.20.2 MarkDuplicates with settings AS = TRUE, VALIDATION_STRINGENCY = LENIENT, MAX_FILE_HANDLES_FOR_READ_ENDS_MAP = 1000, REMOVE_DUPLICATES = TRUE. The number of mapped reads per sample was obtained using samtools view with settings -c, -F 260. For the calculation of contig relative abundance, deduplicated picard outputs were passed to the pileup.sh program from bbTools (Bushnell, 2014) to obtain fragments per kilobase million (FPKM) values for each contig, with these being used to calculate CPM values using a custom R script. For the calculation of gene relative abundance, deduplicated picard outputs and prodigal 2.6.3 GFF3 outputs (using setting -gff and -p meta) were passed to FeatureCounts with settings -p, -T 32, -t CDS, -f, and -g ID. Read counts were converted to FPKM and then CPM values using custom R scripts.

3.3.3 Statistics

3.3.3.1 Analysis of taxonomic and gene composition

Understanding the differences in species and taxonomic composition of different ecosystems is a multifaceted problem which requires a range of techniques. At the broadest level, differences in composition can be assessed by multivariate techniques which normalise the abundance of each taxon, and then compare the (dis-)similarity in abundance of these elements. Other important taxon-level questions (Figure 2) are how the dominant (most abundant) species change in relative abundance and which species had the largest relative changes in abundance. This latter question is more well suited to detecting species which are more rare, as large changes are less probable when a species has a high abundance (*i.e.,* a species with 50% abundance cannot display a positive fold change under a different experimental condition when using proportional abundance metrics), meaning that if a species-level picture of community composition changes is desired, neither approach should be used alone. Further to this, the question of whether the most abundant species remain the most abundant species should be addressed—to achieve this we examined rank abundance changes between arable and extensive fields.

Compositional changes in taxonomy and CAZy genes were assessed using permutational multivariate analysis of variance (PermMANOVA, adonis2 function) with land use type as a predictor. Permutations were kept within site to prevent pseudoreplication, and to account for site level effects. Environmental parameters which correlated with community composition (the initial model included pH, SOC, N:C, P:C, and N:P) were assessed using the same method, and the significance of predictors was checked using stepwise deletion of terms. Plots were produced

using nonmetric multidimensional scaling (NMDS) and canonical correspondence analysis (CCA), using the significant predictors from the PermMANOVA to constrain the ordination. Site was used as a as a conditioning factor.

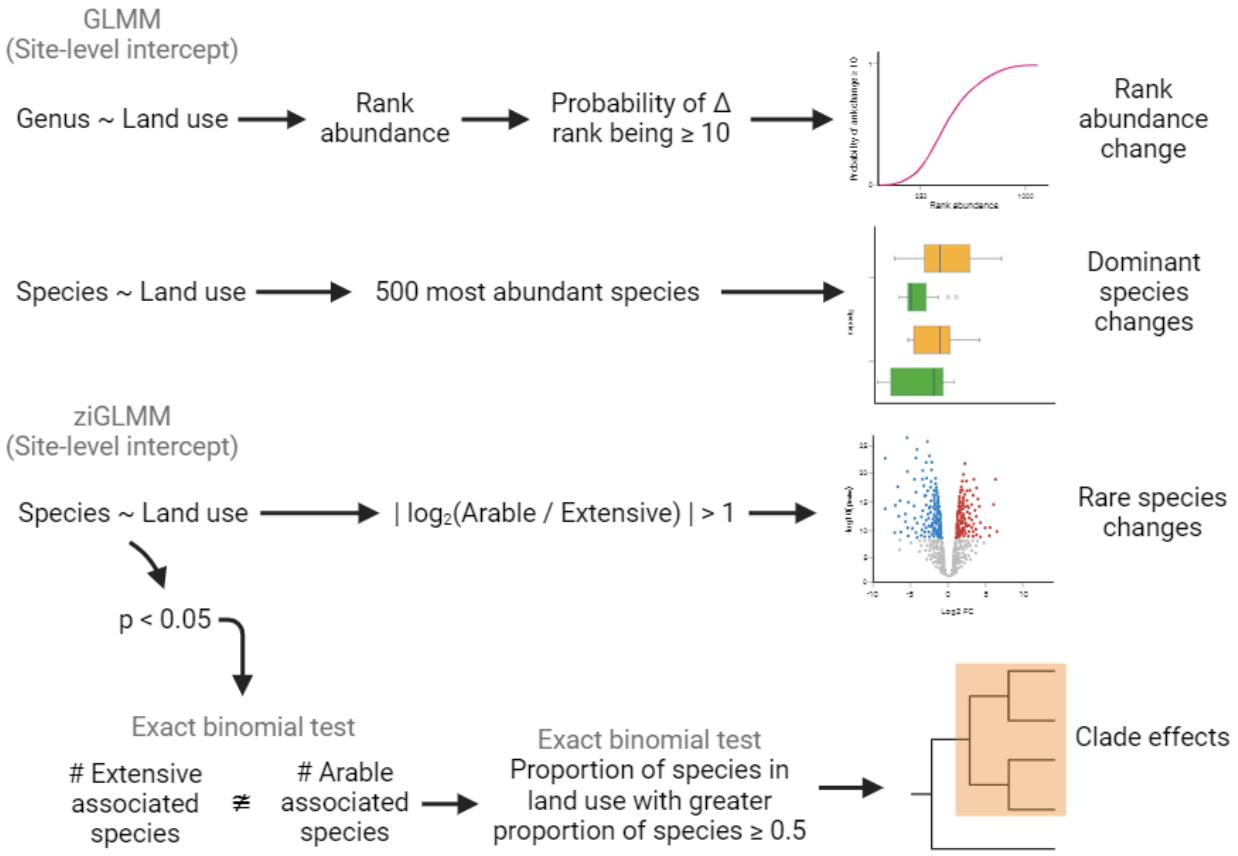

Figure 2: Overview of univariate methods used to assess changes to the taxonomic composition of the microbial community. Abbreviations: GLMM = Generalized linear mixed-effects model fitted with glmmTMB, ziGLMM = Generalized linear mixed-effects model with zero inflation parameter fitted with glmmTMB (ziformula = ~1).

To understand how individual species' rank abundances changed with land use, and the point at which rank changes become large, differences in the rank abundance of species between land use types in relation to their rank abundance in grassland were investigated using binomial generalized linear mixed-effects models (GLMMs; using glmmTMB 1.1.4) with logit link functions and the random intercept of site. The within-sample CPM values were converted to rank abundance (1 being the most abundant), and the site-level means for each land use type for each taxon was calculated. If the difference in mean rank abundance between land use types within a site was greater than or equal to 10, the taxon was deemed to have had a large rank change, and this categorical output was used as the response variable for a binomial model with a logit link

function. This was regressed with respect to the fixed effect of rank in grassland soils, with the random intercept of taxonomy and the intercept-correlated random slope of rank in grassland, allowing for similarly ranked species to respond differently according to the model. Model predictions were used to find the species rank-abundance in grassland where the mean rank abundance change was 10. The same procedure was repeated for a rank change of 5. To investigate how land use affected dominant species, associations between land-use and the 500 most abundant species in the dataset were tested using the same GLMMs (the logit link function accounts for the compositional nature of sequencing data); these models were used to estimate the abundance changes of these species. To understand which species were strongly affected by land-use change, differences in the relative abundance of each taxon (at species level, including unclassified sequences at each taxonomic rank) between land use types was checked using zero-inflated generalised linear mixed models (ziGLMM; glmmTMB ziformula = ~1) with beta error distributions (beta_family function) and logit link functions, and the random intercept of site. Species with strong changes in proportional abundance were identified by checking for species which had a significant $\log_2$ fold-change above 1 or below -1. Land use change effects on α-diversity was assessed using a model with a gaussian error distribution for the response variable of Shannon's diversity index of species and sequences with at least a domain level classification, the predictor of land use type, and the random intercept of site.

To understand the effect of land use on all clades at different taxonomic ranks, we used a series of exact binomial tests. First, the number of species which significantly increased in each land use type within each clade was compared. Second, if there were significantly more species which were associated with one land-use in a clade, two exact binomial tests were conducted to test whether the proportion of species in that clade which responded in the aforementioned direction was not significantly different from, or was or greater than, 0.5 (two-sided test and one-sided test respectively). Clades with roughly half or above half of species being significantly associated with one land use type were deemed to have been affected by land use type.

3.3.3.2 Do lignocellulase genes affect species responses to land use type?

We wanted to understand how land use change-induced differences in the relative abundance of soil microorganisms could alter the functional potential of the microbial community with regards to lignocellulose degradation. To achieve this, separate binomial GLMMs with the predictor of species association with land use (identified using the above ziGLMMs) with the random intercept

of site were used to quantify the proportion of the species with reads mapping to genes for cellulases, xylanases, and AAs. CPM values within each sample were converted to presence/absence for each lignocellulase gene type for a species. The analysis was performed for all species which were associated with a particular land use type (as well as those which had no association). Additionally, for species which had reads mapping to a lignocellulase gene in at least one sample, we fitted binomial GLMMs with logit link functions and the random intercept of site to test if the land-use association affected the average proportion of reads mapping to lignocellulase genes, within a species (CPM lignocellulase genes / CPM species). The purpose of this analysis was to test whether lignocellulase-rich species were significantly associated with a particular land-use type.

3.4 Results and discussion

The sequencing libraries had a mean number of quality-controlled sequences of 24.08 million paired-end reads (SD = 3.43 million paired-end reads) per sample. The assemblies had a total of 6928287 sequences, with an N50 of 1816 bp, and 1292 contigs longer than 50 kbp.

3.4.1 Effects of land use on microbial community composition

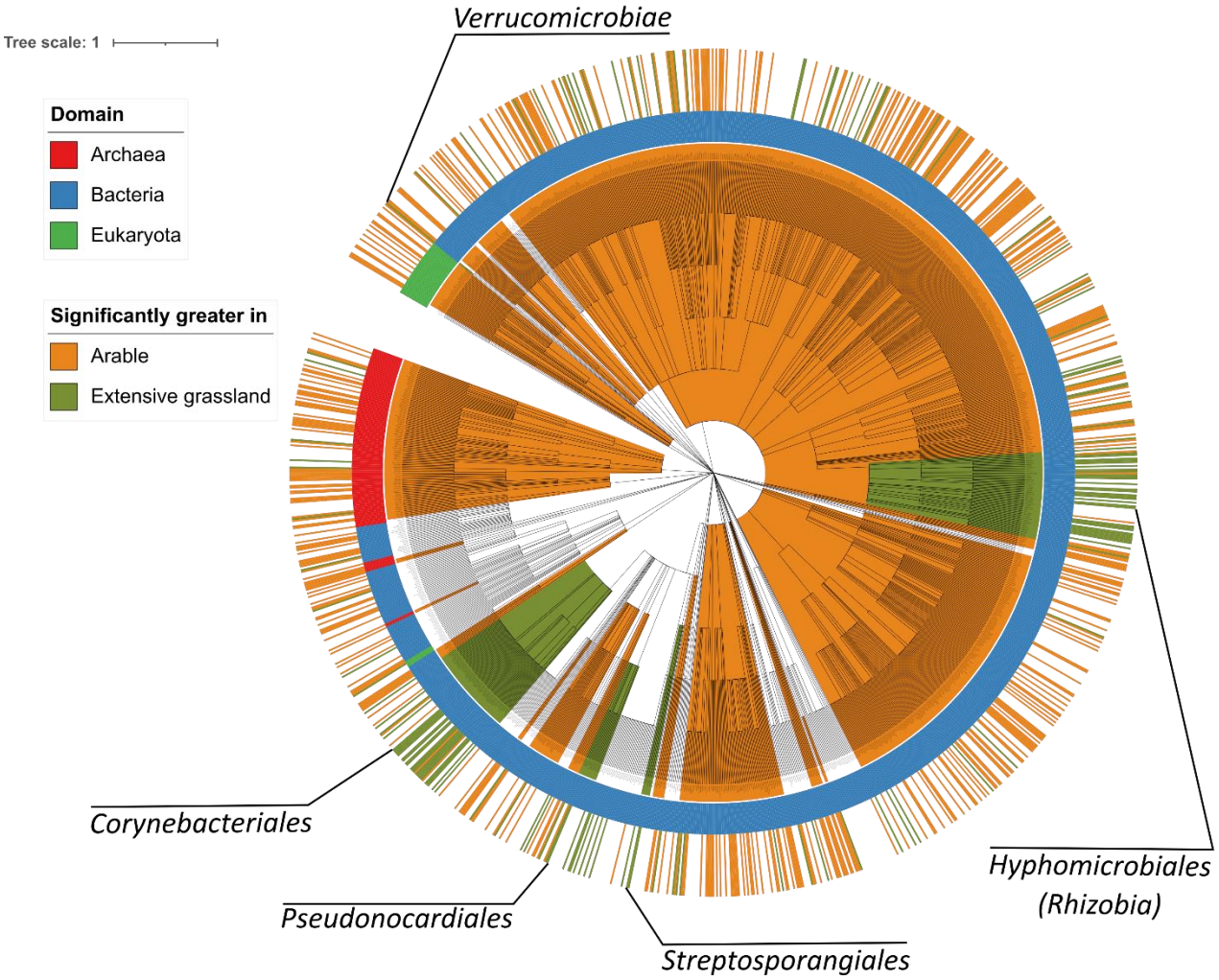

Figure 3: Taxonomic tree of species in the dataset. Broad phylogenetic groups which are highlighted represent clades which had roughly half, or over half of, the number of species in the clade being affected by land use in the same way, additionally there had to be significant bias towards an increase in one of the land use types in the highlighted clade. Taxa which were significantly associated with extensive grasslands are labelled. The outermost circle shows the phylogenetic distribution of species which were affected by land use change; species which were significantly more proportionally abundant (according to zero-inflated generalised linear mixed effects models) in extensive grasslands are highlighted in green, whereas those which were significantly more abundant in arable grasslands are highlighted in orange. The inner circle shows the domain level classification, and the middle circle shows phylum level classification. Only microbial sequences which could be identified at species level and which were present in more than 20% of samples are shown, although the groupings are based on the complete dataset of sequences which could be classified at a minimum of phylum level.

Land use type had strong effects on the relative abundance of broad phylogenetic groups across domains, with 22 clades from phylum to order level having consistent species-level increases in proportional abundance under arable management (Figure 3). The species level increases in arable soils amounted to significant increases in 1262 of the 3223 species or unassigned taxa (39%) in the study. By contrast, only 286 species in total (9%) were associated with extensive grasslands. The

majority of extensive grassland associated species were in one class (*Verrucomicrobiae*) and four orders (*Actinomycetia:Corynebacteriales*, *Actinomycetia:Pseudonocardiales*, *Actinomycetia:Streptosporangiales*, *Alphaproteobacteria:Hyphomicrobiales* (*Rhizobiales*)). These actinomycetal and proteobacterial taxa represent the most abundant classes of soil microorganism within *Proteobacteria* and *Actinobacteria* in British soils (Griffiths *et al.*, 2011). The proportional increase in the abundance of many widely taxonomically distributed species in arable soils in this study likely represent the decreased dominance of these grassland-associated taxa, as a result of reduced microbial biomass in arable soils due to reduced SOC content (the number of bacterial cells per gram of dry soil in this study decreased from 2.8 million to 1 million with agricultural intensification: GLMM: $\chi^2_1 = 52.74$, $p < 0.0001$), with increased detection of rare taxa and relic DNA (Carini *et al.*, 2016). Some species will likely have actually benefited from the decreased abundance of dominant taxa and changes to the environment, however, it is impossible using metagenomics to state which taxa are increased through increased cellular numbers or through increased detection because of negative effects on dominant community members—metatranscriptomics, metaproteomics or PLFA analyses are more suited to this application. These changes to the microbial community led to increased measured microbial diversity in arable ($H' = 5.37$, 95 % CI [5.27, 5.47]), relative to in extensive grasslands ($H' = 5.11$, 95% CI [5.01, 5.21]; $\chi^2_1 = 45.04$, $p < 0.001$), as seen in comparisons between other cereal and grassland systems (Tuck *et al.*, 2014; Delgado-Baquerizo, Maestre, *et al.*, 2016).

Species in the archaeal phyla *Euryarchaeota,* and *Thaumarchaeota* all consistently increased in relative abundance in response to arable management. Agricultural practices have been shown to increase archaeal cellular abundance (Gattinger *et al.*, 2002), gene copy number and N_2O emissions (Du *et al.*, 2019), and transcription of the *amoA* gene relative to bacterial transcription of this gene (Leininger *et al.*, 2006). Results from these previous studies provide evidence that the observed shift in *Archaea* may result at least partly from increased cellular abundance in arable soils.

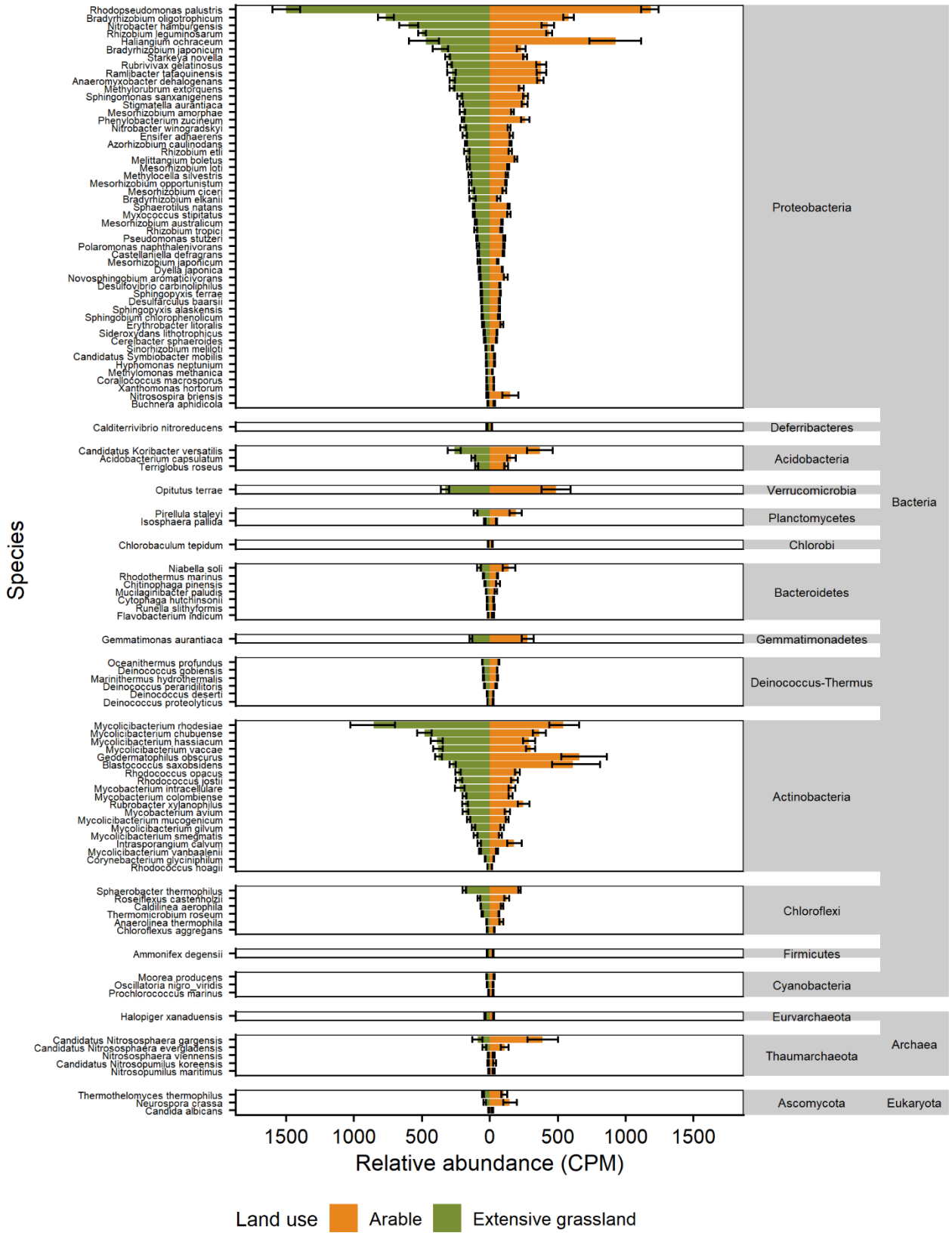

Figure 4: Dominant species (top 500 most abundant) which showed significant differences in relative abundance according to GLMMs (Bonferroni corrected p values). Species are ordered by abundance in extensive grassland, meaning rank abundance changes in arable soils are visible. Values are expressed in counts per million (CPM). Bars show mean relative abundance in each land use type across all samples, and error bars show the bootstrapped 95% confidence interval.

We wanted to investigate whether land use affected the relative abundance of dominant species, which might hopefully illuminate how more than 1000 broadly phylogenetically distributed species had increased relative abundance in this dataset. Totally unclassified sequences were the most abundant part of this dataset. On average, arable soil had roughly 15000 counts per million (CPM) more unclassified sequences than extensive grasslands (GLMM: χ^2_1 = 5.43, p = 0.020). Excluding unclassified sequences (although inclusion does not change the outcome of the test), extensive grasslands had significantly more of the proportional abundance data (63% 95% CI [60%, 65%]) occupied by the 500 most abundant taxa than arable (60% 95% CI [0.58%, 0.63%]), including taxa with assignment only to domain level (GLMM: χ^2_1 = 9.904, p = 0.002). Land use impacted the relative abundances of a high proportion of dominant soil taxa (Figure 4; Bonferroni-corrected p value < 0.05 for 111 of the 500 most abundant species (with species-level assignment), according to GLMMs). Dominant species of *Actinobacteria* (133 species) were highly negatively affected by land use intensification, with 15 of the 19 species which showed significant changes being negatively affected (Figure 4). Romdhane *et al*. (2022) also found strong effects of intensification on actinobacterial OTUs, which decreased from 7% in perennial grassland to 3% in continuously cropped soils, mirroring our results (Figures 3, 4). Roughly half of the dominant species of *Proteobacteria* (237 species total, 51 responsive) which responded to land use decreased in response to intensification. These extensive grassland associated *Proteobacteria* belonged exclusively, bar sequences assigned as *Methylomonas methanica*, to *Hyphomicrobiales* (*Rhizobia*; Figures 3, 4), which act as indicators of high SOM contents (Armbruster *et al.*, 2021)—this contrasts with findings from nutrient poor acidic soils in Brazil (Souza *et al.*, 2016). *Hyphomicrobiales* in soil comprise some of the most dominant species and genera, and exhibit both free-living, host-associated, and mixed life-histories. These include mutualistic associations with *Fabaceae* where they induce formation of root nodules and are major N fixers, and pathogenic associations such as those which cause root tumours (Wang *et al.*, 2020). As such they are a key component of terrestrial ecosystem functioning (Spehn *et al.*, 2002). *Hyphomicrobiales* account for 15% of the length scaled read abundance in extensive grasslands in this study, relative to only 11% in arable. The most stark species-level fold reduction in response to intensification comes from the dominant species *Bradyrhizobium elkanii,* with a mean abundance of 117 CPM in extensive grasslands, but of only 57 CPM in arable soils (Figures 4, 5). *Bradyrhizobium* has been shown to be one of the most abundant genera of soil microorganisms across continents and

biomes (Delgado-Baquerizo *et al.*, 2018), and so large abundance changes to species within this genus will likely have consequences to the broad functionality of the soil microbial community, however, we did not find any lignocellulolytic genes associated with this genus, suggesting it does not play a direct role in the decomposition of polymeric SOM.

All dominant species from all other phyla (except for sequences assigned to the extreme halophile species of *Euryarchaeota*, *Halopiger xanaduensis*) had increased proportional abundances in arable soils (Figure 4). The increased relative abundances in dominant species concur with those from globally distributed studies on the effects of agriculture on grasslands, with a study on the soil microbial community in Argentinean pampas showing that agriculture increases the relative abundance of *Gemmatimonadetes*, *Nitrospirae*, candidate division WS3, and potentially *Acidobacteria*, and decreases *Verrucomicrobia*, *Planctomycetes*, *Actinobacteria*, *Chloroflexota*, and *Bacteroidota* (Carbonetto *et al.*, 2014), and a French study showed increases to *Bacteroidota* with continuous cropping (Romdhane *et al.*, 2022). Interestingly, the study by Carbonetto *et al.* (2014) did not show any differences to the relative abundances of *Proteobacteria* at the class-level. Large granularity may have masked order-level or more fine resolution changes to the microbial community.

The major relative abundance changes to species of *Archaea* are evident; in *Thaumarchaeota* 16 out of the 20 detected species more than doubled in abundance, 18 of 150 species of *Euryarchaeota*, 9 of 48 species of *Crenarchaeota*, and one of ten Candidatus *Thermoplasmatota* showed the same response. The diversity and richness of *Thaumarchaeota* in agricultural soils has been related to pH and C:N ratio (Lu, Seuradge and Neufeld, 2016; Saghaï *et al.*, 2022) which may explain differences between the diversity of many archaea between land uses in this study.

Species of *Ascomycota* all also responded to land-use intensification by increasing in abundance in arable soils, concurring with results from George *et al.* (2019), who found increases in the ascomycotal classes *Sordariales*, *Eurotiomycetes*, *Dothideomycetes*, and *Pezizomycota* in cropland soils. Increased abundance of *Ascomycota* may be a result of their foraging capability and capabilities for the degradation of lignocellulose.

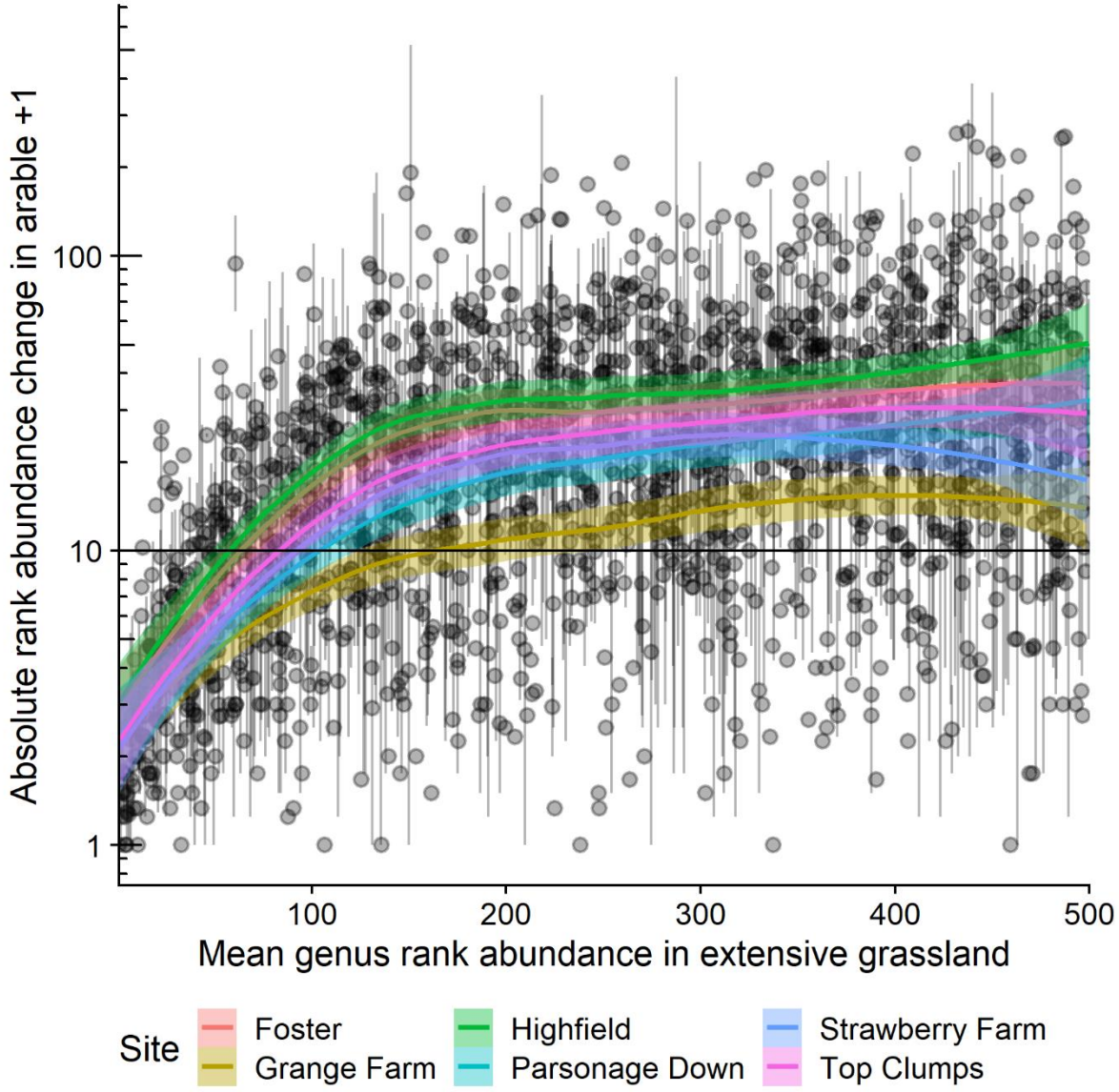

Figure 5: Decreasing abundance (x axis) increases the magnitude of rank-abundance change (irrespective of whether a genus increases or decreases in arable) due to agricultural intensification for the 100 most abundant species, at which point the size of the rank-abundance change asymptotes. Only the 500 most abundant genera in extensive grasslands are shown. Note the log scale on the y axis. Trendlines show LOESS curves and standard errors. Points show the mean rank abundance change due to arable, averaged over all sites. Error bars represent bootstrapped 95% confidence intervals. The horizontal line at 10 is included to delineate small changes in rank from larger changes.

While changes in relative abundance of different species and taxa, along with multivariate techniques, provide an insight as to how communities change, it is hard to interpret if changes in abundance cause large changes to the degree of dominance that species have in the community. Because of this, we examined how species rank abundance changed with agricultural intensification, as dependent upon rank abundance in extensive grasslands. The magnitude of

change in species rank-abundance from extensive grassland to arable increased as species became less abundant (Figure 5). Rank abundance changes were generally small until the 75th most abundant species in extensive grassland where the mean rank abundance change for a genus was 10; individual sites showed variability around this value (Figure 5). These findings suggests that the community composition of the most dominant species is relatively stable, while the community composition of less abundant taxa becomes highly variable.

Despite the stability of the most abundant taxa (Figure 5), the widespread species-level differences in abundance gave strong and consistent shifts in the microbial community composition at all sites in response to land use intensification ($p < 0.001$), even after modelling the influence of SOC, N:P, and N:C, according to *Perm*MANOVA (Figure 6a, b, Table 1). Other studies have shown similar effects on grassland microbial community composition (Manoharan *et al.*, 2017; Sünnemann *et al.*, 2021; Romdhane *et al.*, 2022).

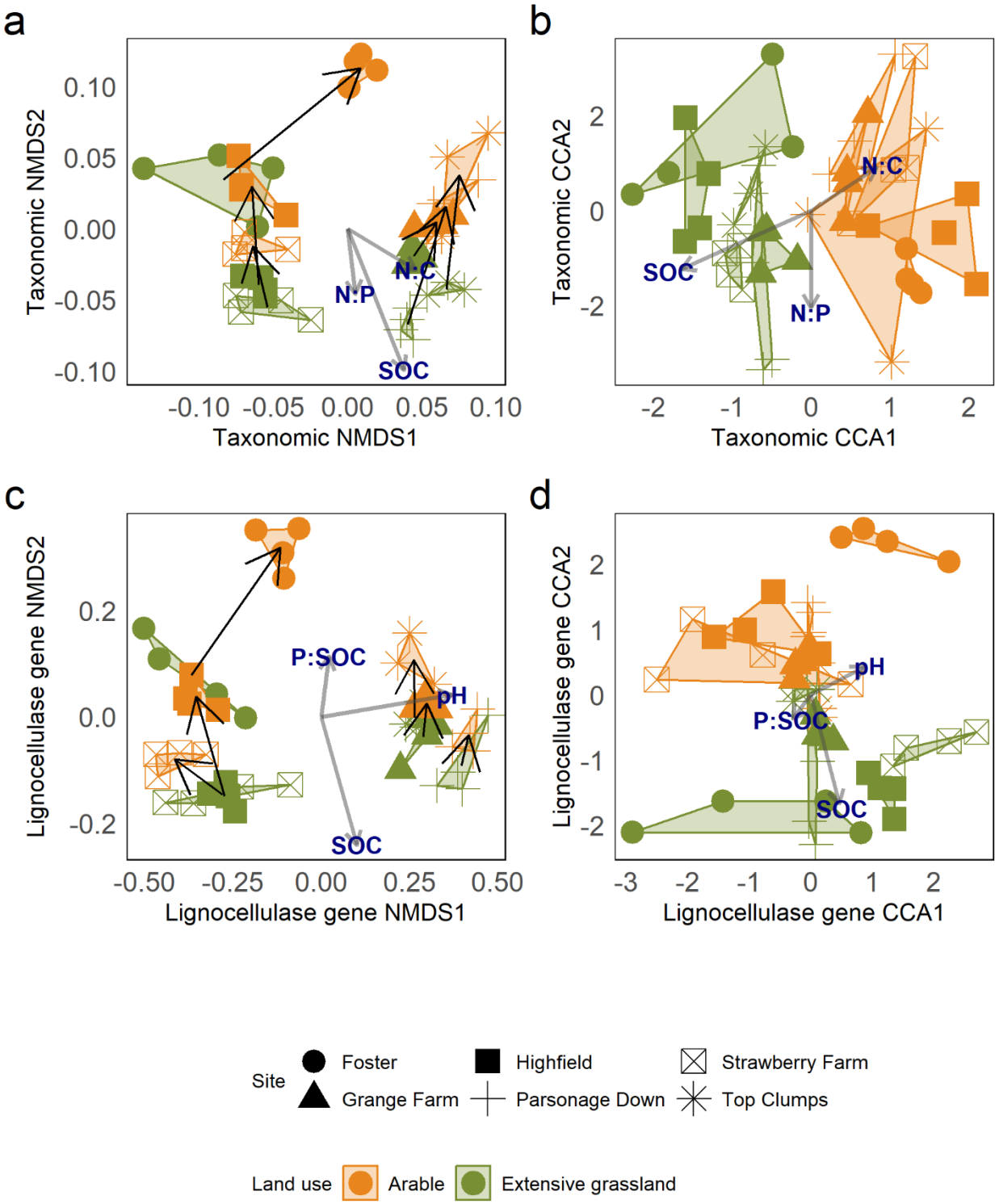

Figure 6: Effect of land-use change on microbial and lignocellulase gene community composition β-diversity. (a, b) NMDS and CCA of class-level microbial community composition. (c, d) NMDSA and CCA of lignocellulase gene composition respectively. Grey arrows show the directions which correlate with increasing values of different soil parameters, according to PermMANOVA, projected using envfit. Black arrows in the NMDS plots link centroids for the same site in different land use types for clarity. Site was a conditioning factor in the CCAs.

Understanding how changes in species composition relate to functional changes is a grand challenge for microbial ecology. Microbial biomass and composition are highly effective predictors of the broadly distributed function of soil carbon respiration (Graham *et al.*, 2016)—this likely depends on the abundance and composition of dominant taxa in the soil microbial community as well as depending on physical soil properties. More specialised functions (*e.g.*, denitrification) depend upon the abundance of specific community members with particular functional genes, with loss of particular species greatly impacting the rate of ecosystem functionality (Delgado-Baquerizo, Grinyer, *et al.*, 2016). Understanding where SOM degradation fits along the scale from generalist functionality to specialized functionality, and knowledge of the how species and the degree of interconnectedness between them alters the rate of this process, is a question which needs answering to improve SOM degradation estimates. Several studies have begun to effectively answer this question, using lignocellulosic substrates labelled with stable-isotopes to differentiate species which gain carbon from these polymers, *versus* those which don't. These studies identified *Proteobacteria*, *Actinobacteria*, *Bacteroidota*, and *Bacillota* as the dominant bacterial phyla in lignocellulolytic soil communities with significant lignocellulolytic potential from *Chloroflexota*, *Cyanobacteria*, *Verrucomicrobia*, and *Planctomycetes*, with *Ascomycota* and *Basidiomycota* dominating the lignocellulolytic fungal communities. Species-level genomic catalogues of the identified species from these studies would allow insight as to how and why the composition of functional genes differs with land use, aiding prediction of how soil microbial community functional potential responds to land use change globally (Wilhelm *et al.*, 2019; López-Mondéjar *et al.*, 2020; Weiss *et al.*, 2021). Testing the effects of land use on these dominant taxa should provide insights as to how and why the composition of functional genes (and thus predicted functionality) differs with land use.

3.4.1.1 Composition of lignocellulase genes

To understand how agricultural intensification might affect the composition of lignocellulolytic gene functions in grassland soils, we analysed changes in the relative abundances and composition of individual gene classes and gene classes which were aggregated by typical function. Lignocellulase and cellulase gene relative abundance were increased in arable soils relative to in extensive grasslands in this study (GLMMs, Lignocellulases: $\chi^2_1 = 5.886$, $p = 0.015$, 12% increase; Cellulases $\chi^2_1 = 9.894$, $p = 0.002$, 20% increase; Figure 7), suggesting greater community potential for cellulose degradation per unit biomass. Xylanase and auxiliary activity genes did not show the

same response (GLMM: $\chi^2_1 = 2.01$, $p = 0.157$, $\chi^2_1 = 0.052$, $p = 0.819$ respectively). This is contrary to the findings of Carbonetto *et al.* (2014), who found no differences in the carbon processing COG genes relating to carbohydrate transport and metabolism in cultivated and uncultivated Argentinean Pampas soils—although they did find differences in overall COG gene composition, which likely has implications for the ecology of the microorganisms in the community (Carbonetto *et al.*, 2014). Previous work has shown that arable farming decreases rates of β-glucosidase and dehydrogenase activity in sandy soils, following microbial biomass trends, making our findings of increased relative abundance of related genes surprising. These differences may reflect the number and type of genes analysed, with this study focussing on specific subsets of genes which are known to frequently encode proteins with particular lignocellulolytic capabilities, whereas the other studies focussed on different functions and a broader selection of genes.

Lignocellulase gene composition also changed consistently with land use intensification, but this was less strong than the taxonomic community shifts relative to the effect of site (Figure 7a, c) which is expected because lignocellulase genes are shared across species. The composition of these genes was associated with SOC content, pH, and the ratio of P to SOC (P:SOC) (Figure 7c, d, Table 1). We hypothesise that this increase in cellulase gene abundance and shift in composition of these genes may reflect increases to microorganisms which utilise the scarce native SOC and the dead root matter which remains in arable systems after harvest, as this will represent an abundant source of nutrients.

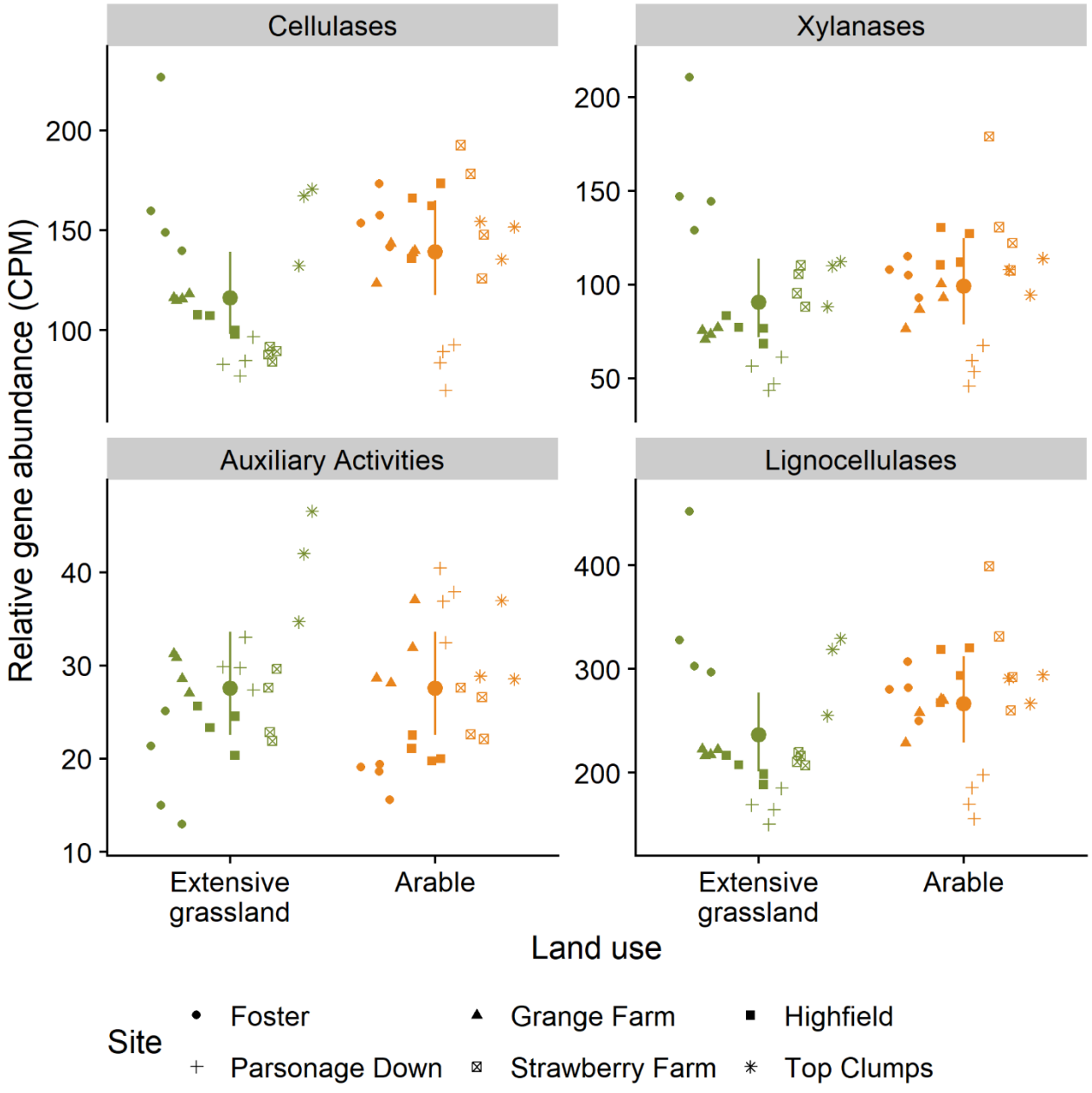

Figure 7: Relative abundance of lignocellulase gene types in each land use type. Values expressed in counts per million (CPM) of each sequencing library.

There were differences in the abundance of lignocellulase genes between land uses at the phylum level (Figure 8). The increased overall relative abundance of lignocellulase genes in arable soils is likely a result of increased lignocellulase gene abundance from *Proteobacteria* (ziGLMM: χ^2_1 = 6.47, p = 0.01), *Planctomycetes* (ziGLMM: χ^2_1 = 14.92, p < 0.001), and *Bacillota* (ziGLMM: χ^2_1 = 9.19, p = 0.002), mirroring the increased overall relative abundance of these phyla in arable soils (Figures 4, 8). Members of *Planctomycetes* and *Proteobacteria* have been shown to have

significant degradative ability, incorporating carbon from multiple lignocellulosic polymers in forest soils, and *Bacillota* have cellulolytic members (Wilhelm *et al.*, 2019).

Table 1: Permutational MANOVA results for microbial community composition and lignocellulase gene composition. P values were generated by permuting samples 60,000 times within site.

Response	**Predictor(s)**	**R^2**	**F**	**D.F.**	**p**
Taxonomic community (class level)	**Land use**	**0.17**	9.28	1, 44	**< 0.001**
Taxonomic community (class level)	**Land use**	**0.14**	11.34	1, 41	**< 0.001**
	N:P	**0.17**	13.56	1, 41	**< 0.001**
	Total carbon	**0.14**	11.68	1, 41	**0.002**
	N:C	0.05	3.68	1, 41	0.070
	pH	0.05	4.09	1, 40	0.547
	P:C	0.01	1.05	1, 39	0.564
Lignocellulases	**Land use**	**0.06**	2.69	1, 44	**<0.001**
Lignocellulases	**Land use**	**0.08**	7.78	1, 40	**<0.001**
	pH	**0.26**	24.58	1, 40	**<0.001**
	P:C	**0.02**	2.10	1, 40	**0.009**
	N:P	**0.03**	2.46	1, 40	**0.012**
	N:C	0.06	5.65	1, 40	0.077
	Total carbon	0.03	2.61	1, 39	0.209

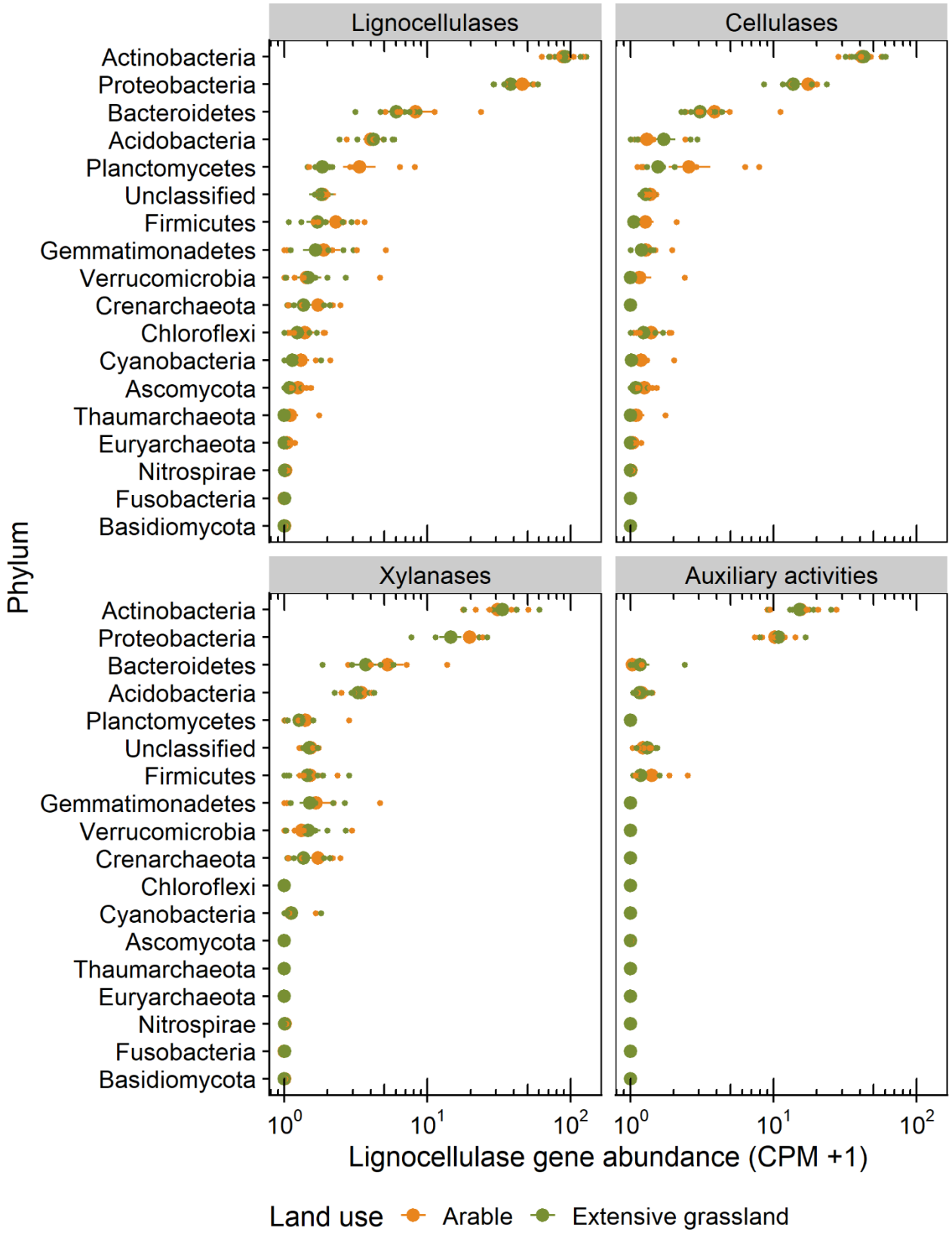

Figure 8: Taxonomic origins and relative abundance of lignocellulase genes. Values are expressed in counts per million (CPM). Error bars show bootstrapped 95% confidence intervals.

3.4.2 Lignocellulase genes do not the drive the microbial community composition response to land-use change

I wanted to understand whether the different species responses to land-use were associated with lignocellulase genetic traits, and so next compared the proportion of species with lignocellulase genes between land use indicator species types (extensive grassland or arable associated, and no association), as well as the relative within-species metagenomic lignocellulase content for each species associated with these groups. This comparison was made using the CPM mapping to lignocellulase genes for a species within a sample, divided by the CPM mapping to that species within a sample.

The two opposing outcomes I wished to explore were that (1) there would be increased lignocellulase gene content in extensive grassland-associated species because extensive grasslands generally possess higher organic matter stocks than arable soils, thus offering potential advantages in nutrient acquisition to competitive degraders, assuming a C-S-R microbial trait model (Grime, 1977; Wood, Tang and Franks, 2018; Malik *et al.*, 2020). (2) Alternatively, increased lignocellulase gene content in arable-associated species could result from the generally lower amounts of native organic matter present, with microorganisms being reliant on enzyme systems for scavenging low quantities of organic material, be it from native organic matter or dead plant material post-harvest. This corresponds with a microbial Y-A-S trait model over axes of resource abundance and stress, and suggests that extensive grasslands should favour species with high carbon use efficiencies rather than competitive degraders (Malik *et al.*, 2020). Given the observed increased total lignocellulase and cellulase abundance in arable fields, the second hypothesis seems more likely, contrary to our initial hypothesis. I therefore seek to specifically examine whether the taxa which change in relative abundance due to land use, and particularly those increased in arable soils, are more likely to possess cellulase degrading genes.

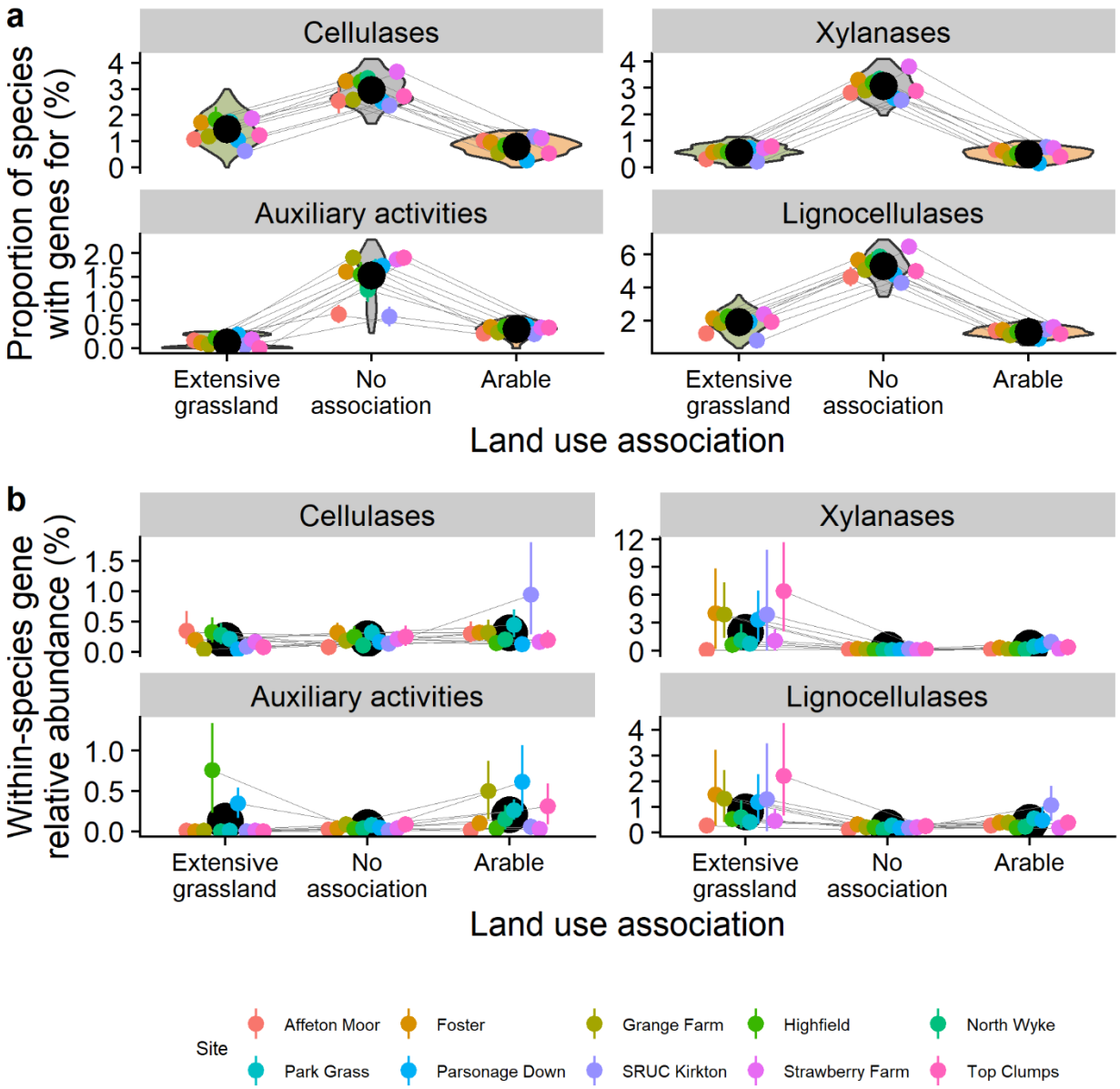

Figure 9: (a) Proportion of species associated with each land use type which had at least one lignocellulase gene. Points show site-level means for all species and error bars show bootstrapped 95% confidence intervals. (b) Percentage abundance of metagenomic lignocellulase genes per species, which implies higher copy number or higher coverage. Points show site-level means for all species with at least 1 lignocellulase gene, and error bars show bootstrapped 95% confidence intervals. In this figure "Lignocellulases" are the sum of the other three categories.

To test the above hypotheses, I firstly quantified the proportion of all species containing genes for each type of lignocellulase (*e.g.,* cellulases, xylanases, auxiliary activities), within each land use indicator classification (extensive grassland, arable, and no association). I secondly chose to explore another factor potentially relating to copy number, by also quantifying the relative abundance of detected genes for lignocellulases within each species (species-level metagenomic proportion). This analysis was performed to understand if changes in overall community-level

enzymic content may not only be due to change in overall numbers of species with enzymic traits, but could also relate to differences in within-species enzyme gene content (*i.e.,* copy number within strains).

I found no evidence that the difference in overall community-level lignocellulase abundances (Figure 7) was linked to changes in specific land use indicator taxa (Figure 9a). Most notably, the proportion of species with lignocellulase of all functional classes were significantly higher for species for which were not associated with a land use type than those associated with either land use (Figure 9a, Table 2), suggesting that the community of microorganisms which are involved in the decomposition of lignocellulosic polymers is relatively stable and does not respond drastically to agricultural intensification. Contrasting indicator species of each land-use type showed that indicator species of extensive grasslands were significantly more likely to have lignocellulase genes (2.16%) than indicator species of arable soils (1.31%, Figure 9a, Table 2), and extensive grassland-indicator species which possessed lignocellulase genes were more likely to have a higher mean within-species lignocellulase gene relative abundance, concurring with sequence capture results from Manoharan *et al*. (2017), and making interpretation of our community-level results (Figure 7) more challenging. This points to abundant lignocellulosic resources driving small changes to the microbial community which increase the abundance of competitive degraders, concurring with the C-S-R model as suggested by Wood *et al*. (2018). Cellulase and xylanase genes followed similar trends to lignocellulases, however the mean values for the proportion of indicator species with xylanase genes in arable and extensive grasslands was not significantly different, and the within-species proportional abundance was not increased for cellulases (Figure 9a, Table 2). Auxiliary activities followed the broader pattern of being more prevalent in species with no land-use association, however, there was a higher proportion of arable indicator species (0.45%) than extensive grassland indicator species (0.19%) with these genes, suggesting greater utilisation of lignin in arable soils. These slight differences in auxiliary activity gene abundance follow the Y-A-S model (Malik *et al.*, 2020) of microbial carbon utilisation more closely than the C-S-R model, suggesting that the frameworks for discussing microbial traits are not complete. There was no difference in within-species relative abundance of cellulase genes between indicator types (Figure 9b, Table 2) which could be used to explain the increase in lignocellulases and cellulases at the community level (Figure 7).

Table 2: Upper table: Contrasts between the proportion of species with genes for lignocellulase functional classes between species which were associated with extensive grassland and arable, or neither soils. Lower table: The proportion of length-scaled lignocellulase gene abundance within each species (CPM species lignocellulases / CPM species) between species which were associated with extensive grassland and arable soils. Letters after z and p denote contrasts being shown, with the first letter showing the reference category. (A = arable, E = extensive grassland, N = no association).

Percentage of species with genes for:	**z (E - A)**	**p (E - A)**	**z (N - E)**	**p (N - E)**	**z (N - A)**	**p (N - A)**	**Mean [95% CI] % for extensive grassland**	**Mean [95% CI] % for arable**	**Mean [95% CI] % for no association**
Lignocellulases	-6.35	**<0.001**	-15.58	**<0.001**	-26.86	**<0.001**	2.16 [1.84, 2.54]	1.31 [1.12, 1.52]	5.59 [5.00, 6.25]
Cellulases	-7.56	**<0.001**	-9.27	**<0.001**	-19.78	**<0.001**	1.62 [1.31, 2.01]	0.79 [0.64, 0.97]	3.16 [2.67, 3.72]
Xylanases	-1.77	0.077	-14.42	**<0.001**	-21.3	**<0.001**	0.64 [0.50, 0.83]	0.50 [0.41, 0.63]	3.20 [2.78, 3.68]
Auxiliary activities	3.853	**<0.001**	-10.62	**<0.001**	-13.94	**<0.001**	0.19 [0.13, 0.28]	0.45 [0.38, 0.54]	1.68 [1.57,1.79]
Within-species prevalence of genes for:	**z (E - A)**	**p (E - A)**	**z (N - E)**	**p (N - E)**	**z (N - A)**	**p (N - A)**	**Mean [95% CI] % for extensive grassland**	**Mean [95% CI] % for arable**	**Mean [95% CI] % for no association**
Lignocellulases	-1.54	0.123	2.87	**0.004**	0.51	0.607	0.87 [0.38, 1.97]	0.29 [0.10, 0.88]	0.21 [0.13, 0.34]
Cellulases	-0.05	0.957	-0.17	0.865	-0.29	0.775	0.19 [0.02, 1.50]	0.18 [0.03, 0.96]	0.23 [0.13, 0.42]
Xylanases	-2.01	**0.044**	4.59	**<0.001**	1.08	0.28	2.07 [0.02, 5.12]	0.29 [0.05, 1.55]	0.10 [0.04, 0.25]
Auxiliary activities	-0.24	0.807	0.95	0.341	0.86	0.389	0.33 [0.01, 15.68]	0.18 [0.01, 0.31]	0.04 [0.00, 0.31]

Together, these results suggest that the lignocellulose decomposition potential of temperate soil microbial communities is largely resilient to the effects of land use change, but that lignocellulose as a whole, cellulose, and xylan, decomposition may be marginally more associated with species in extensive grassland soils, however lignin decomposition may be marginally more associated with indicator species of arable soils. The observed increases in the community-level relative abundance of lignocellulolytic and cellulolytic genes in arable soils (Figure 7) are therefore unexplained. The data in this study points to neither hypothesis about the ways in which land use could affect lignocellulase gene-rich taxa as a succinct explanation of the observed different patterns. The analyses of the proportion of species containing genes for total lignocellulases and cellulases, and the within-species metagenomic lignocellulase and xylanase content, (Figure 9) provide weak evidence for the argument of Wood *et al*. (2018) who suggest that high availability of resources (high-native SOC contents in extensive grassland) which require enzymatic conversion should drive community selection for enzyme-rich competitive species. Conversely, our analyses of the proportion of species containing auxiliary activity genes (Figure 9) provide weak evidence for the argument of Malik *et al*. (2020), who suggest that scarcity of total resources (low native SOC contents in arable soils) benefits species with the best resources for acquisition, but not those which produce high numbers of enzymes (although we did not detect any differences to the mean gene abundance within-indicator species of different land uses). These hypotheses require further study to conclusively understand the life-history strategies of lignocellulolytic microorganisms at the community level.

Given the decreased proportion of species with cellulase genes in arable soils, it is possible that the increased relative abundance of these cellulase genes does not relate to any specific cellulase containing taxon being more abundant than in grassland, but reflects a more general characteristic of the molecular data—namely that due to the reduction in dominant species, arable soils comprise proportionally more taxa at rare abundances (Figure 4), and more of these are likely to contain lignocellulases (*e.g.* the non-land use associated taxa observed in Figure 9a). The activity status of these diverse and rare species is unknown and warrants further investigation. More generally these results demonstrate that genomic lignocellulase content is not a critical driver of most microbial species' responses to land use intensification in the case of transitions between extensive grasslands and arable fields, where reductions to plant community diversity, manure and/or fertilizer application, tillage, and loss of SOC stocks have the potential to alter the most beneficial nutrient acquisition or life-history strategies (Ho, Di Lonardo and Bodelier, 2017). Future work should aim to link metagenomic and metatranscriptomic data with the realised functional capacity of different land use classes, firstly to elucidate which members of the community are active, and to quantify how expression of genes in each land-use may translate into ecosystem processes.

3.5 Conclusions

Our analysis uses multiple methods to highlight key changes to the composition of microorganisms, at both broad and fine-resolutions, as well as the composition of lignocellulolytic genes in response to land use intensification in grasslands. We find that extensively managed grasslands have highly specialised microbial communities with strongly dominant species, and that intensive agriculture decreases overall bacterial numbers, the relative abundance of dominant species, and radically increases diversity, possibly as an artefact of these factors. We found a greater proportion of extensive grassland-associated species with lignocellulase and cellulase genes. Lignocellulolytic, and particularly xylanolytic species, had a higher relative abundance of genes for the decomposition of lignocellulose than arable-associated species. This was in line with our hypothesis about microbial utilisation of abundant SOC resources. A larger proportion of arable-associated species had auxiliary activity genes than grassland-associated species, suggesting that different life-history strategies may be employed by glycoside hydrolase and auxiliary activity producing microbial species. Despite these findings, our hypotheses overestimated the importance of lignocellulase genes in determining species responses to land

use, as lignocellulase genes of all functional classes were more rare in species which were responsive to land use type than in those which were not. We therefore find that lignocellulolytic capabilities are relatively buffered against the effects of agricultural intensification. There was a 20% increase in cellulase, and a 10% increase in lignocellulase, relative abundance at the community-level in intensive agricultural soils, relative to in extensive grasslands. This is likely due to the increased relative (but not actual) abundance of many diverse species. Future research should address the functional relevance of apparent increases in key carbon cycling genes, such as for cellulases, which may arise not through increases to cellulose degraders, but rather because of decreased biomass and proportional abundance of dominant taxa. Such changes may lead to increased detection of diverse rare species and DNA from senescing or dead cells which are likely to possess relatively common lignocellulase genes.

4

High-throughput *in situ* cultivation, genomics and phenotypic characterisation of lignocellulolytic soil microorganisms

4.1 Abstract

Our knowledge of lignocellulose decomposition in soil is largely limited to a constrained diversity of microorganisms phenotypically characterised *in vitro*. To increase the cultivated diversity of lignocellulosic microorganisms from soil, we combined enrichment of soil microbiota on lignocellulosic biomass buried *in situ* with high-throughput *in situ* cultivation of soil microorganisms with an iChip. This provided hundreds of viable lignocellulolytic isolates in pure-, or near-pure-culture. Based on the outputs of phenotypic screening on a range of lignocellulosic polymers, we genome sequenced 65 isolates, finding a high degree of taxonomic novelty (possibly 9 novel species), albeit within the well-studied and abundant genera *Pseudomonas*, *Pantoea, Ochrobactrum,* and *Agrobacterium*. Through *a priori* knowledge of the substrate specificities of gene products, and pangenome-wide association (pan-GWA), we identified genes in the isolates which are likely causative of the substrate utilisation phenotypes. Additionally, we identified genetic pathways and broad functional groups associated with growth on each of the lignocellulosic polymers studied. We identify pathways and genes in *Pseudomona*s which correlated with the substrate utilisation patterns. Synthesis of B-vitamins and iron chelation *via* siderophores and oxidative redox potential were commonly associated with lignocellulose polymer utilisation. Therefore, different metabolic strategies (production of proteins, or cell growth and reproduction) may determine growth of isolates on particular lignocellulosic polymers. Overall, this study highlights the complexities of relating phenotype to genotype, and shows that pan-GWA is an appropriate technique for understanding microbial life-history traits, but is relatively poor at finding causative genes for particular processes. Increasing the diversity of

cultivated microorganisms may benefit from combination of new cultivation media and high-throughput *in situ* cultivation.

4.2 Introduction

Soils are Earth's largest store of terrestrial carbon, containing 1500 gigatonnes of organic carbon, predominantly in the form of decaying lignocellulosic plant material (Lal, 2008). Consequently, the microbial-mediated decomposition of lignocellulose in soils is a key feature of the global carbon cycle. Understanding the genomics and physiology of diverse lignocellulolytic microbiota is a research priority, given the importance of soil microbiota for the hydrolysis and oxidation of lignocellulosic plant biomass, helping to regulate global carbon cycling. Microbiologists have cultivated several tens of thousands of microbial species, yet estimates of Earth's microbial species richness reach into the trillions (Locey and Lennon, 2016). These cultivated organisms are the basis of our knowledge of microbial physiology and function in ecosystems. Recent nucleotide sequencing advances such as single-gene community profiling, assembly of metagenome-assembled genomes (MAGs) and single-cell amplified genomes (SAGs), and environmental metatranscriptomics have quickly accelerated our understanding of the roles that microorganisms play in ecosystems through homology-based annotation of the species and genes (Parks *et al.*, 2021). However, at present the number of species with high-quality assembled genomes does not even reach 100,000 (Parks *et al.*, 2021). Despite the ever-expanding knowledge of uncultivated and undetected 'dark microbial diversity', relatively little work has focussed on expanding and confirming the known phenotypes of never-before cultivated microorganisms. The probable functions that this uncultivated microbial diversity can perform can be inferred from genes with known function, found by cultivating and manipulating microorganisms, or through experimentation with heterologously translated genes (Chistoserdova, 2009; Kobras, Fenton and Sheppard, 2021). The reliance on gene homology from a maximum of one percent of the total microbiota is a major limitation to our knowledge of functional enzymatic domains, and thus ecosystem functions. Indeed, because the majority of microbial life from all ecosystems has never been cultivated, we cannot expect to have a working understanding of the genetic basis for complex ecosystem functions. This means we have a very poor understanding of how ecosystems will respond to future climates and anthropogenic disturbance.

Overcoming this knowledge boundary will involve significant scientific effort, and investment in diverse cultivation techniques. Traditional cultivation methods on average produce isolates from

less than 1% of the microorganisms plated onto an agar Petri-dish across many environmental types (Lloyd *et al.*, 2018), a phenomenon which has become known as the 'great plate count anomaly' (Staley and Konopka, 1985). While it should be noted that there is significant variability in cultivability, with cultivability of microorganisms from soil generally being less than 5% (Sait, Hugenholtz and Janssen, 2002), the diversity of the uncultivated component of the microbiota deserves attention. The species which are successfully cultivated are typically highly competitive for resources, fast-growing, and tolerant of oxidative stress, meaning that less competitive species which can grow in these conditions have been overlooked until relatively recently.

Strategies for increasing the novelty of cultivated lineages include use of oligotrophic media (Watve *et al.*, 2000), careful selection of cultivation medium, long incubation times (Kato *et al.*, 2018), reducing the inoculum size to reduce the prevalence of fast-growing competitors, physical separation of individual cells (Zhang *et al.*, 2021), and physical but not chemical separation of individual cells from their native environment (Kaeberlein, Lewis and Epstein, 2002; Vartoukian, Palmer and Wade, 2010). The iChip is a tool which has been successfully used to cultivate up to 40% of cells from sea sediment (Kaeberlein, Lewis and Epstein, 2002), and has famously been used to discover a novel class of antibiotic (Ling *et al.*, 2015a). It achieves cultivation of novel microorganisms through a combination of high-throughput dilution to extinction and '*in situ*' incubation. That is, a single non-commercial iChip (Figure 2b) incubates up to 96 individual cells in agar plugs in the environment that the microorganisms originate from, allowing the transfer of small molecules into the cultivation medium, but preventing transfer of microbial cells. Cultivation of novel species with this method is possible partly because of the coevolution of species within syntrophic networks, with isolates growing in response to molecules from microorganisms with which they share a common resource (Nichols *et al.*, 2008).

Targeted isolation and cultivation of particular microbial trophic groups can be achieved through enrichment cultures or 'baiting'. This technique has been widely used and is an effective method for isolating rare species from communities that have particular functional traits. For instance, it has been used to enrich and isolate lignocellulose degrading microbiota from landfill sites (McDonald, Allison and McCarthy, 2010), mycophages from the rhizosphere (Ballhausen *et al.*, 2015), thermophiles from deep-sea hydrothermal vents (Harmsen, Prieur and Jeanthon, 1997), and biosurfactant producers from oil wells (Araújo *et al.*, 2020), among others.

To address the lack of cultivated lignocellulolytic microbial diversity, we combined enrichment of microorganisms on lignocellulosic biomass with high-throughput microbial cultivation in an iChip to isolate novel lignocellulolytic microorganisms. Additionally, we set out to perform *in vitro* lignocellulose substrate utilisation tests and genome sequencing of the isolate collection, to relate phenotype to genotype and gain an understanding of the genes involved in lignocellulose degradation by these isolates.

The aims of the study were to isolate novel lineages of microorganisms with lignocellulolytic capabilities. We expected to achieve this outcome because of the combination of lignocellulosic enrichment, the iChip, and high throughput screening techniques. Additionally, we expected to identify genes which were causative of the observed phenotypes through genomic techniques.

4.3 Methods

All microbiological and DNA work was conducted using aseptic technique in either a biological safety cabinet, a laminar flow hood, or a PCR cabinet. Surfaces were sterilised with bleach, methylated spirits, and where an ultraviolet light was installed, surfaces were irradiated with ultraviolet light-.

4.3.1 Isolation of microorganisms

Because lignocellulolytic microorganisms often bind strongly to the lignocellulolytic substrate (Neumann, McCormick and Suen, 2017), detachment of microorganisms from this substrate may need to utilise harsh sample processing methods. In this study we chose to use three different cell detachment methods to detach microorganisms from the matrix (shaking, blending, and bead-beating; Figure 1c). An initial purpose of this study was to compare the microorganisms obtained by each cell detachment and cultivation method (high-throughput *in situ* cultivation *versus* traditional cultivation and colony picking from plates), using the 16S rRNA gene. Unfortunately, a series of failed sequencing runs and time constraints meant that achieving this aim was not possible. Instead, the following methods section focusses on retrieval of microorganisms from the iChip devices, cultivation, phenotypic screening, and genomics of the cultivated isolates.

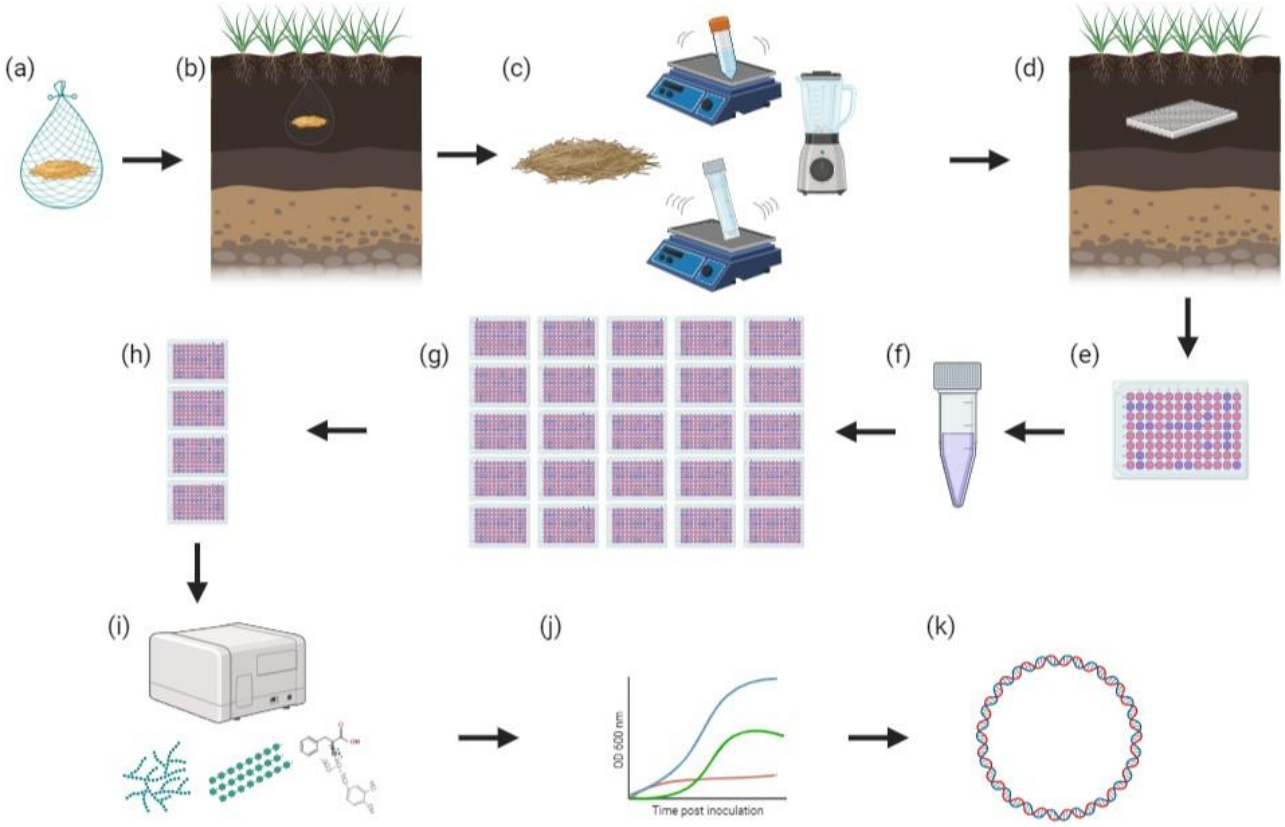

Figure 1: Schematic of the workflow used within this study. (a) Dried mature grass (hay) from Henfaes farm was bagged and (b) incubated in the soil for four months. (c) Partially degraded hay was processed by three methods to detach microorganisms with the cell suspension being (d) incubated in iChip devices in the soil for one month. I Cultures were screened for the ability to degrade amorphous cellulose, with positive hits being (f) transferred to a combined glycerol stock. (g) High-throughput isolation was used to increase the purity of the in-situ cultivated microorganisms, and (h) microorganisms from highly dilute wells were kept. (i) Phenotypic growth tests were performed for four lignocellulosic carbon sources, (j) and these were modelled, with the outputs being used to inform which isolates should be (k) genome sequenced.

Microbial isolation and *in situ* cultivation took place at Bangor University's Henfaes Research Centre, Abergwyngregyn, United Kingdom (53.24°N, 4.02°W; Elevation 12 m a.s.l.). Hay from the site, representing a lignocellulosic bait, was placed into four porous nylon bags (approximately 10 g per bag; Figures 1a, b, 2a, b) before being buried in a sandy clay loam textured Eutric Cambisol soil in a field which was previously grazed by sheep (*Ovis aries* L.) on 2018-06-25. The soil around the buried bags was kept moist by weekly watering with roughly 0.6 L of water per bag because the soil was so dry at the beginning of the experiment, with this amount reducing as rainfall increased. The nylon bags were retrieved on 2018-10-25 (four months), and hay samples were immediately processed by three methods (shaking, bead-beating, blending) to detach microorganisms and produce cell suspensions for inoculation into the iChip. Each hay sample and cell detachment method was inoculated into a separate iChip (12 iChips in total; Figure 2c) as follows: Hay was weighed into sterile weighing boats before being transferred to the appropriate container for each detachment method. For the shaking treatment, 1 g of each sample was shaken

at 300 rpm in a sterile 50 mL centrifuge tube on a rotating platform shaker in 25 mL sterile phosphate buffered saline, for 30 minutes. The blending treatment involved transferring 1 g of each sample in 200 mL PBS into a George GJB101B blender, sterilised with bleach and industrial methylated spirits. Samples were processed on setting "1" for two minutes. The bead-beating treatment used 0.1 g of each sample in an autoclaved Q-biogene purple-top multimix tube containing 0.5 g acid washed glass beads (425 – 600 µm), and 1.25 mL PBS. These were processed three times in a FastPrep FP120 bead-beater at 5.5 $m \cdot s^{-1}$ for 30 s).

The processed cell suspensions were underwent twelve 5-fold serial dilutons in PBS, with each resulting dilution being used to fill one row (8 wells) of an iChip device, so that each row was at a different dilution factor. These cell suspensions were mixed with double-strength Czapek-Dox agar + 0.1% (w/v) carboxymethyl cellulose (CMC) as a cellulosic carbon source to form the agar plugs of the iChip device. iChips were sealed with silicone glue and a poly carbonate track etched membrane with 0.03 µm diameter pores. iChip devices were buried at 10-15 cm depth 8 hours after the hay samples were retrieved (Figures 1a, 2c). iChip devices remained buried until 2018-11-05 (one month), and were watered weekly to keep the agar moist (Berdy *et al.*, 2017). Isolates were transferred into glycerol stocks using a pin replicator. As an initial screening for lignocellulolytic microorganisms, the isolates were grown from the glycerol stocks, using a pin replicator to transfer them on to on sterilised Q trays containing Czapek-Dox agar + 0.1% (w/v) CMC (Figures 1e, 2c). This allowed testing of the cultures for CMCase activity using congo red once colonies were visible. Colonies with zones of clearing were deemed to have lignocellulolytic potential and were marked for further investigation (Carder, 1986). Glycerol stocks (70% glycerol) of the CMCase activity positive microorganisms were taken and were kept at -80 for further work. The proportion of CMCase positive wells in each treatment was analysed using a binomial generalised linear model (GLMs) with a logit link function and cell detachment method as a predictor.

Because the glycerol stocks contained many non-pure cultures, we chose to re-isolate from the CMCase positive cultures. These glycerol stocks were mixed by adding 4.8 µL of each glycerol stock, to give a final volume of 2 mL (Figure 1f). This 2 mL stock was diluted in a liquid version of the Czapek-Dox + CMC medium on which the isolates were initially isolated and cultivated (pH 7.3) so to ensure we retained the full diversity present in the sample at concentrations of between 10^{-3} and 10^{-8} of the initial concentration. From each dilution, 200 µL was added to each well of five 96-

well microplates (Figure 1g). Plates in which below roughly 50% of wells showed microbial growth were deemed to be dilute enough for dilution to extinction to have worked effectively (Figure 1h). The purity of the microorganisms was checked by streaking out onto solid Czapek-Dox + CMC medium, and checking colony and cell morphology on microscope slides.

4.3.2 Quantification of lignocellulosic carbon source utilisation by soil microbial isolates

Purified microbial isolates were transferred from microplates containing liquid Czapek-Dox + CMC after reaching stationary phase to microplates containing liquid Czapek-Dox medium with no sucrose (LCD-media), but with the addition of either sucrose and CMC, CMC, Avicel microcrystalline cellulose, xylan, and kraft lignin as carbon sources, as well as a microplate with LCD-medium but with no carbon source as a control (no carbon control). There was no treatment of sucrose as a single carbon source. Wells with each substrate, but with no microorganisms were included in each treatment (no microbe control) to ensure that there was no growth in these wells due to contamination. These microplates were sealed with parafilm (Bemis Company, Inc.), and were incubated at 25°C with shaking at 150 rpm.

The optical density at 600 nm (OD) of each well from each microplate was measured using a Cerillo Stratus plate daily (three measurements per timepoint) for 6 days, to view trends in microbial growth on each substrate. To give an unbiased viewpoint about the microorganisms' abilities to utilise each substrate, a variety of population growth models (aomisc package for R, as well as custom functions) were fitted to these data (Figure 1j). For each well in each substrate treatment, OD was used as the response and time as the predictor, and the models fitted used were: linear regression, linear regression with a 2^{nd} order polynomial, linear regression with a 3^{rd} order polynomial (all fitted using the lm and poly functions), Bragg model, Michaelis-Menten model, logistic model, log-logistic model, Gompertz model, Holling IV model, Ricker model, negative exponential model, Shepherd model, Hassel model (all fitted using the nls function). The population growth model with the lowest AIC was chosen, although where the model predictions of population growth against time were clearly poor, the model with the next lowest AIC was chosen. Microorganisms which had predicted mean optical density 0.1 OD greater than the no-carbon control at any point on each substrate were deemed to be able to utilise the carbon source in the well.

4.3.3 Genome sequencing and annotation of lignocellulose degraders

To check the purity of the isolates which could utilise lignocellulosic carbon sources, microorganisms were plated on Czapek-Dox agar with the addition of 0.1% CMC (w/v), or the carbon substrate they grew best on if there was no growth on Czapek-Dox + CMC. Plates with a single colony morphology were deemed pure, and subcultures from single colonies of each type were picked and streaked onto Czapek-dox agar with the single lignocellulosic carbon source that the culture grew best on, to increase the chances of obtaining truly pure isolates. Single colonies were then picked and were grown in LCD with the same single lignocellulosic carbon source, in microcentrifuge tubes at 25°C in a shaking incubator (150 rpm) for two weeks. The contents of these were pelleted at 5000 *g* for 5 mins, and the pellet was used for DNA extraction following the MoBio PowerLyzer PowerSoil protocol. Glass beads were used for the bead beating step in a MoBio Power Lyzer which was set to 2000 rpm for 30 s. DNA extract purity was checked using a NanoDrop, and concentration was checked using a Qubit 3.0. DNA extracts underwent library preparation following the protocol for the NEBNext Ultra II FS DNA Library Prep with 10 ng of input DNA, with a fragmentation time of 15 minutes. The library indexes used were from the NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs) kit. Sequencing was performed using the Illumina NextSeq 1000 platform with a P1 reagent kit, after quantification and insert size checking using a TapeStation. Using a Pippin prep did not improve the distribution of sequence lengths and so the original library was used for sequencing.

The returned sequences were quality controlled using bbduk with options ref=adapters,artifacts,phix ktrim=r k=23 mink=11 hdist=1 ftl=3 maq=25 minlen=35 t=20. Human reads from the masked human genome available at https://drive.google.com/file/d/0B3llHR93L14wd0pSSnFULUlhcUk/edit?resourcekey=0-PsIKmg2q4EvTGWGOUjsKGQ, as well as PhiX, and reads found in the negative were removed from both paired-end and unpaired reads using BBMap (Bushnell, 2014) with options minid=0.95 maxindel=3 bwr=0.16 bw=12 quickmatch fast minhits=2 path=. Qtrim=rl trimq=10 untrim -Xmx23g -t=20. The bbTools (Bushnell, 2014) programs dedupe, reformat, and bbmerge were used to remove PCR duplicates, reformat the output, and merge overlapping reads. Remaining adapter sequences were removed using cutadapt (options -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT) before assembly. Genomes were assembled from each library using SPAdes 3.15.5 on merged, and

unmerged paired-end and unpaired reads, using 20 threads, 180 GB RAM, and t– --isolate flag (Figure 1k). CheckM v1.1.3 (lineage_wf -t 40) and CheckM2 0.13 (predict function with 40 threads and all models) were used to estimate genome completeness and contamination for all assemblies.

As second-generation sequencing technologies commonly return non-target DNA sequences (human or laboratory contaminants, sequences from the "kitome" and "splashome" (Olomu *et al.*, 2020)), assemblies with high CheckM contamination scores were investigated in more detail. Contaminated assemblies were further assessed using blobplots; coverage and GC content were calculated using InfoSeq from EMBOSS 6.6.0 and pileup.sh from bbTools. These contaminated assemblies were split using MetaBat2 (options: --minContig=2500, --minCV=0.1, --maxP=99, --minS=95, --maxEdges=2000), using sorted bam files produced by minimap2 with the short read pre-set as the input. Taxonomic annotation was performed using GTDB-Tk 2.1.1 using the classify_wf option, and functional annotation of lignocellulase genes was performed using a custom version of dbCAN2 which replaces hmmscan with hmmsearch for increased speed, with evaluation cut-off of $1e^{-15}$ (similar to the hmmsearch cut-off when accounting for format induced database size changes) with a coverage of 0.35.

The best 16S rRNA gene match for each isolate was found after assembling these gene sequences. Reads mapping to 16S rRNA gene sequences were obtained using phyloFlash, although these did not assemble properly. To overcome this we mapped the reads obtained from phyloFlash to the 16S rRNA gene of *Pseudomonas protegens* isolate CHA0 (Hida *et al.*, 2020, GenBank accession CP003190.1) using bbMap with settings minid=0.7, maxindel=10, bwr=0.16, bw=12, vslow, k=8, minhits=2, path=., -Xmx23g, and -t=40. Consensus sequences were obtained using angsd/0.935 with options -doFasta 2 and -doCounts 1. The resulting consensus sequences were concatenated, 'N' characters, whitespace, and newlines within the sequence were removed using sed. These sequences were searched against the NCBI rRNA/ITS 16S ribosomal RNA sequences (Bacteria and Archaea) using megablast, excluding sequences from Models and uncultured or environmental sample sequences. Because the sequences were not full length, but contained gaps and some had short overall lengths, we decided upon the best matching sequence by taking the sequence with the highest mean of the similarity score and the alignment length/10.

4.3.4 Finding links between genotype and phenotype

We conducted a pan-GWA analysis on the 68 isolates which were identified as *Pseudomonas_E* by GTDB-Tk. Whole genome trees were constructed from the pangenome using gubbins (generate_ska_alignment.py for alignment using 40 threads and using the longest assembly as a reference; gubbins options we– --threads 40, --tree-builder raxml, --first-tree-builder iqtree, --model GTRGAMMA, --recon-model GTRGAMMA, --seq-recon raxml). Pangenomes were created by first annotating assemblies using Prokka 1.14.5 (--cpus – --genus Pseudomonas --usegen– --mincontiglen 200) to create GFF files, and then passing these into Roary 3.12.0 (options -cd 95, -p 40 -e -n). The option for the percentage of organisms which contained a gene for it to be included as part of the core pangenome was lowered from 99% to 95% as the higher cut-off returned too few genes for reliable phylogenomics. Pan-GWA was performed with Scoary (40 threads) using the substrate utilisation phenotypes determined in the above section, with the gene presence/absence table produced by Roary, and the whole-genome phylogenetic trees produced by Gubbins. We used 100,000 label-switching permutations to calculate the p-values for the gene-phenotype links. The COG (clusters of orthologous groups) categories and pathways of all phenotype-correlated genes were assessed (Galperin *et al.*, 2020). Chi-squared goodness-of-fit tests were conducted to test whether the proportion of genes in each of the COG categories were statistically likely to have come from the same population between substrates. Additionally, we searched for relationships between *a priori* lignocellulolytic genes (found using dbCAN2, with known lignocellulolytic activities) and lignocellulolytic phenotypes using linear models, where maximum relative OD prediction (compared to the same isolate on the no-carbon control) was the response, and number of putatively lignocellulolytic genes was the predictor.

4.4 Results and discussion

4.4.1 High throughput *in situ* cultivation yielded a high proportion of unclassified lignocellulolytic isolates

An initial aim of the study was to compare the community of microorganisms retrieved through traditional and *in situ* cultivation methods using replicate iChips. Due to failed sequencing runs, and lack of time, this aim was not pursued, and therefore we present results only from the iChips. From the 12 iChip devices (Figure 2b), microorganisms and simple microbial communities from 416 out of the 1152 wells showed CMCase activity in an initial screening for lignocellulosic activity (Figure 2c), and these were transferred to glycerol stocks for further work. Cell detachment

method had no detectable effect on the proportion of CMCase positive wells, which was 36% across all samples (95% CI [15%, 65%]; binomial GLM: $\chi^2_2 = 3.30$, $p = 0.829$). Microorganisms from these glycerol stocks were combined and then reisolated using high throughput isolation techniques which yielded 304 isolates from plates with microbial growth in fewer than half of the wells. An example picture of one of the pure cultures is presented in Figure 2d.

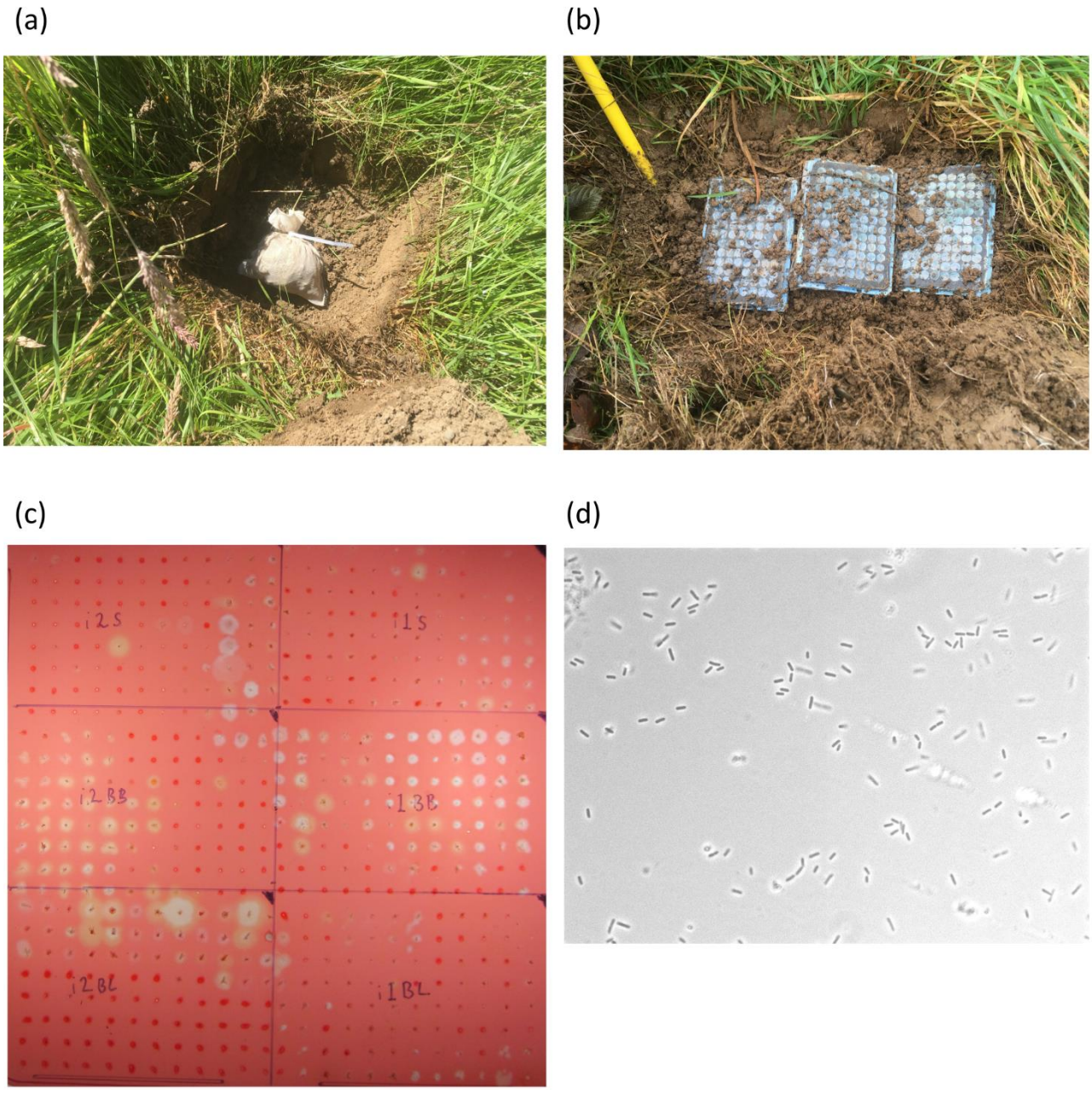

Figure 2: Photos from the project. (a) A porous nylon mesh bag containing hay, about to be buried and watered. (b) iChip devices inoculated with hay from a single nylon bag. The three iChips shown were inoculated with cell suspensions from the same degraded hay, which were obtained using different methods (shaking, blending, and bead beating in phosphate buffered saline). (c) Results of phenotypic tests for CMCase activity. Czapek-Dox cue trays with CMC as an additional carbon source were inoculated with cultures from the iChips, and a congo red stain was used to find isolates which had degraded the carboxymethyl cellulose. (d) Microscope image of a visually pure microbial culture following reisolation using dilution to extinction from one of the 10^{-7} dilution plates.

Phenotypic screens for utilisation of lignocellulosic polymers as sole carbon sources revealed growth of 173 isolates (56%) on one or more of the lignocellulosic substrates (Figure S1), showing the efficacy of the baiting and iChip technique for recovering lignocellulolytic microorganisms. Overall, 36% were found to grow on CMC, 22% on Avicel, 15% on xylan, and 25% on lignin. Predictions from the models used to determine if a microorganism could utilise a particular carbon source can be found in Figure S2.

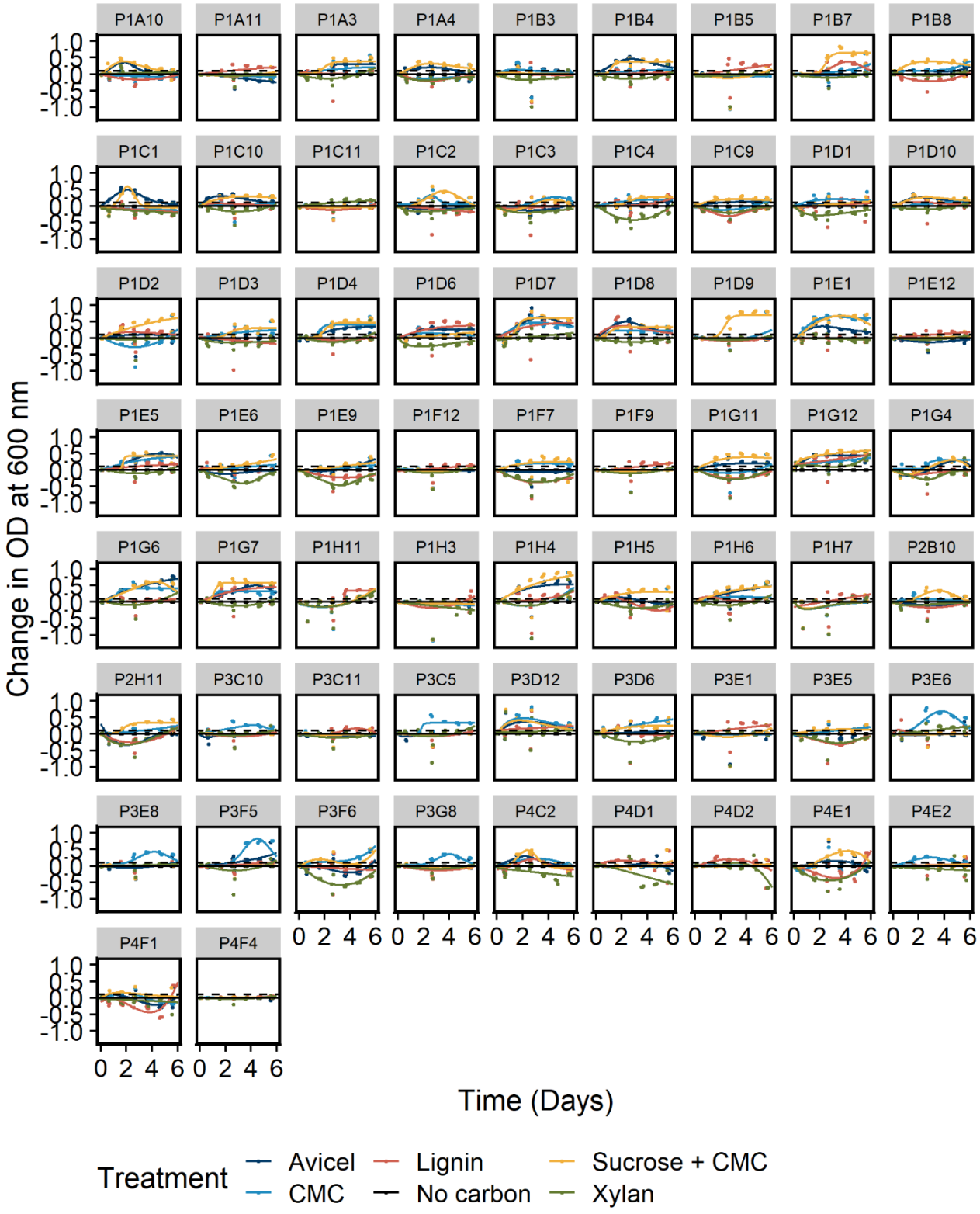

Figure 3: Change in optical density (OD) of the genome-sequenced isolates growing on different carbon sources, relative to a control without any carbon sources (after correcting for media optical density). Plotted lines show the mean predictions from the best population growth model for each curve. The dashed horizontal line shows the OD cut-off which was used to determine if a microorganism utilised a substrate or not (+ 0.1 OD). All growth models are shown in Figure S2.

Genome sequences were obtained for 83 of the isolates which exhibited different substrate utilisation patterns (Figure 3). Growth was observed for 78% of the sequenced isolates on CMC, 60% on Avicel, 30% on xylan, and 53% on lignin. We recovered many *Pseudomonas_E* isolates which will be discussed further below. The genomes had a mean size of 5.7 Mbp (SD = 2.03 Mbp), largest contigs of 0.23 Mbp (SD = 0.29 Mbp), a mean number of contigs longer than 1 kbp of 670 (SD = 722 kbp), and an N50 of 69.18 kbp (SD = 107 kbp). The total length of the GTDB representative genomes and NCBI type strains of *Pseudomonas_E fluorescens_A* (GCF_000802965.1), *Pseudomonas_E putida* (GCA_000412675.1), *Pseudomonas_E brassicae* (GCA_010671725.1), and *Pseudomonas_E protegens* (GCA_000397205.1) lie between 5.5 and 6.9 Mbp. This shows that our estimates of genome size are concordant with those observed for other *Pseudomonas* isolates. A single library, P1F4b had a very high contamination score according to CheckM, and so we plotted its per-contig % GC *versus* it's per-contig coverage as a means of identifying separate organisms (Laetsch and Blaxter, 2017)—this revealed two distinct clusters suggesting the presence of two organisms. Processing of the assembly with MetaBat2 effectively separated these groups of sequences, reducing the contamination score from 149% to 0.83% in the larger of the two bins (estimated completeness of 88%). After refinement of this genome, 38 out of 85 genomes passed the criteria for inclusion into GTDB, however, further work on this dataset could improve this number. None of the assemblies were single contig or ungapped. Strategies to improve assembly quality include the genome binning approach used on P1F4b, the one refined genome in this dataset, removal of sequences shorter than a defined cut-off (*e.g.,* 1 kbp), or re-sequencing of the isolates using long-read sequencing technology as most of the genomes which did not meet the GTDB criteria did not meet the criterion of < 1000 contigs, or N50 of the contigs being less than 5 kbp. Long-read sequencing and genome refinement helps to resolve repetitive regions such as rRNA operons, detect misassemblies (Utturkar *et al.*, 2017), fill scaffolding gaps (Boetzer and Pirovano, 2012), and give further insight into the physiology and life history of a microorganism based on its gene coordinates (Sonnenschein *et al.*, 2009; Sobetzko, Travers and Muskhelishvili, 2012; Romeo, Vakulskas and Babitzke, 2013; Gerganova *et al.*, 2015; Lato and Golding, 2020). Unfortunately, the work described above is beyond the scope of the current study.

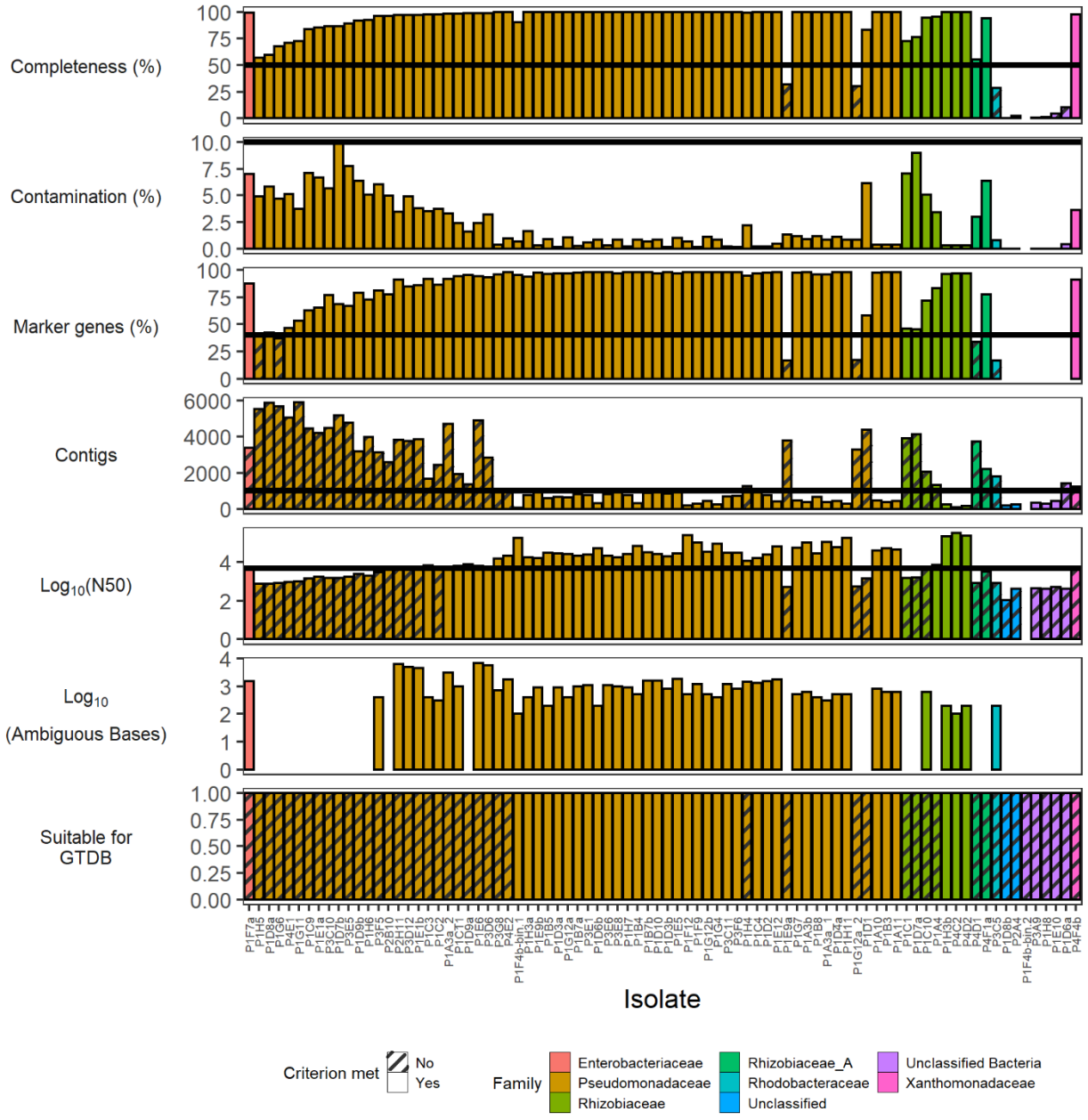

Figure 4: Genome Taxonomy Database criteria for genome inclusion for each of the sequenced genomes. Hatched bars show genomes which did not satisfy the inclusion criterion in question. Horizontal black lines show the cut-off value for each variable according to GTDB. The bottom panel shows is when all criteria were met, and therefore if the genome is suitable for inclusion as a novel isolate genome. Criteria were: CheckM completeness > 50%, CheckM contamination < 10%, Quality score > 50, > 40% of the bac12 marker genes, < 1000 contigs, N50 > 5 kbp, < 100,000 ambiguous bases.

Genomes were taxonomically classified mostly (65 isolates) into the GTDB genus *Pseudomonas_E* (Table 1, Figure 4), with isolates from the species *Pseudomonas_E neuropathica, Pseudomonas_E sp000633255*, *Pseudomonas_E sp002286815*, *Pseudomonas_E sp003097075*, and *Pseudomonas_E sp014851765*. Many of the *Pseudomonas_E* isolates remained unclassified, despite high genome completeness values (Figure 4, Table 1)—these isolates were placed as multiple sister clades to known species in whole genome trees (Figure 5), suggesting that the current isolate collection may

contain multiple novel species of *Pseudomonas_E*. In addition to isolates from the genus *Pseudomonas_E*, we cultivated seven isolates belonging to *Agrobacterium fabacearum*, *Ochrobactrum_A quorumnocens* and an isolate of *Ochrobactrum_A* (55% genome completeness) which was not classified at the species level, an isolate of *Paracoccus* which was not given a species designation (28% genome completeness), *Pantoea agglomerans*, and *Stenotrophomonas sp003484865*. Taxonomic classification of the isolates using recovered 16S rRNA sequences revealed that 31 of the 78 isolates had less than 99% sequence similarity (four *Agrobacterium*, one *Pantoea*, and 26 *Pseudomonas*; considering only comparisons with over 500 bp of alignment) to cultured bacteria. For 97% sequence similarity, this number halved to 15 isolates (three *Agrobacterium*, the *Pantoea*, and 11 *Pseudomonas*). The whole-genome taxonomic assignments and the dissimilarity of the recovered partial 16S rRNA genes (> 500 bp alignment coverage) from cultivated species' sequences gives reasonable evidence that many of these isolates belong to previously uncultivated species. Further work should be undertaken to validate this claim, and should include further phenotypic characterisation, sequencing of the 16S RNA gene directly and phylogeny building using this gene and those of cultivated representatives, lipid analysis, and proteomic analyses. This relatively small sampling effort using iChips, and taxonomically blind selection of isolates for genome sequencing based solely off phenotype has yielded isolates from six families, and has found potentially novel species from two genera outside of *Pseudomonas_E,* a genus in which we expect up to have found up to five novel species, based on the whole genome trees (Figure 5). The rate of discovery of potentially novel classified species of soil microorganism in this dataset is 9% (7 out of 78 isolates), which is between a three- and seven-fold increase relative to the rate given by traditional solid-medium cultivation and solid-medium cultivation focussing on isolation of slow-growing colonies, respectively, as reported in other studies (Kato *et al.*, 2018). Alterations to medium preparation methods can also substantially increase the rate of discovery of novel species, with autoclaving of phosphate other medium elements separately giving a 0.4-fold increase for fast-growing species, and an 8.4-fold increase for slow-growing species (Kato *et al.*, 2018). Future combination of alterations to media-preparation methods, high-throughput *in situ* cultivation, high-throughput isolation techniques, and targeting of slow-growing microorganisms could vastly increase the diversity of cultivated microbial life, increasing our understanding of ecosystem functioning as we study these organisms and their metabolic capabilities.

Of the 65 *Pseudomonas_E* isolates, our OD measurements suggest that 21 could utilise both one and two of the lignocellulosic polymers as a sole carbon source, 9 could utilise three of the polymers, and 14 could utilise all four polymers. Further quantification of substrate utilisation capabilities of the isolates will improve the capability for discovering and interpreting meaningful genetic associations with substrate utilization. Eight of the *Pseudomonas_E* isolates which could degrade all 4 polymers belonged to unidentified species which were identified as novel according to GTDB average nucleotide identity classification. One fully lignocellulosic isolate belonged to each of *Pseudomonas_E neuropathica* and *Pseudomonas_E sp002286815*, and four belonged to *Pseudomonas_E sp002018875*. Of the 20 non-*Pseudomonas* isolates, six could utilise one, two, and three of the polymers, one (P1D7a, *Agrobacterium fabacearum*) could utilise all of the polymers, and one (P4F4b, *Stenotrophomonas sp003484865*) was unable to utilise any of the polymers, even though this isolate has a GH8 gene. The P4F4b isolate was included for genome sequencing because there were spare sequencing library indexes.

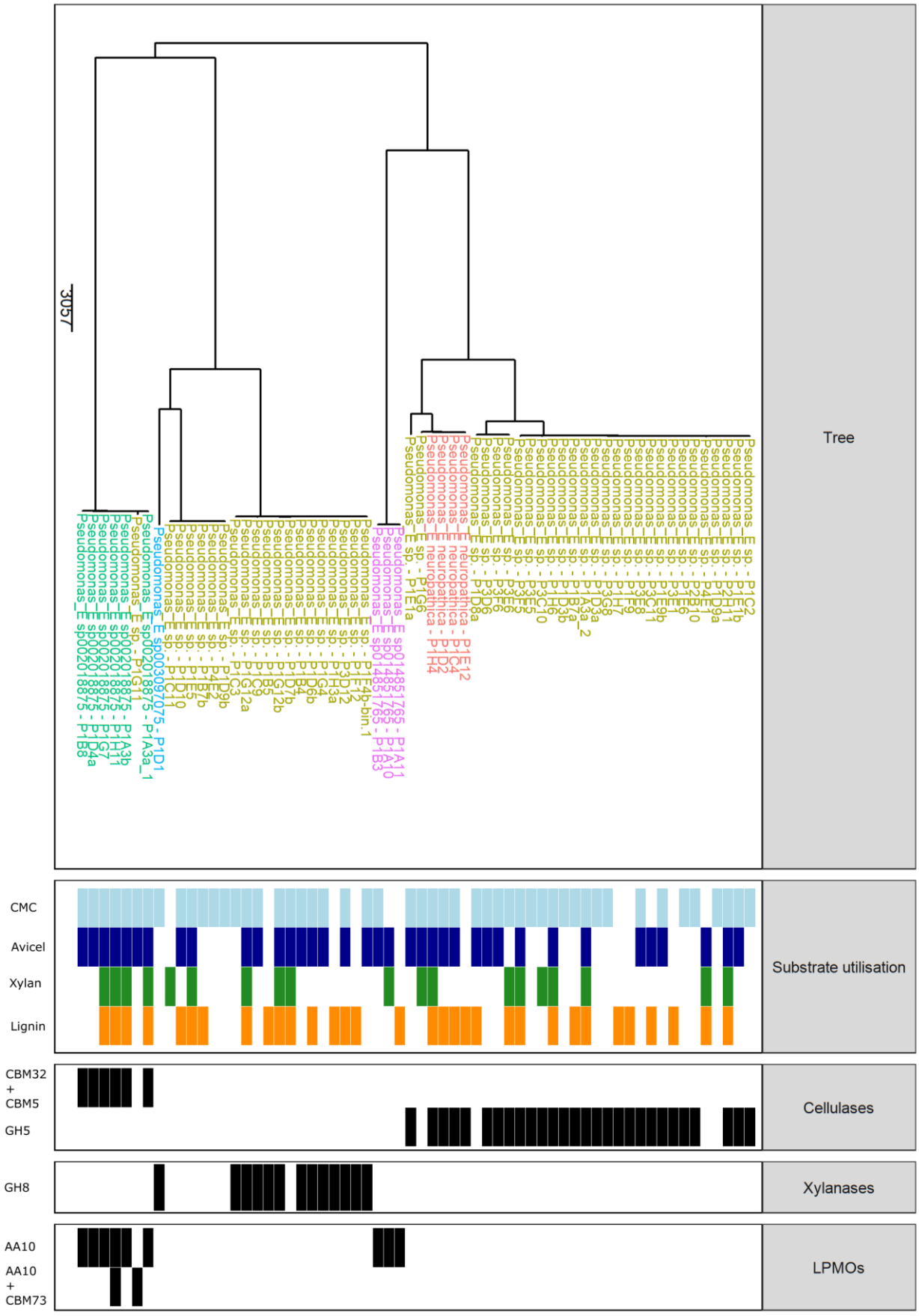
3057
Tree
CMC
Avicel
Xylan
Lignin
Substrate utilisation
CBM32
+
CBM5
GH5
Cellulases
GH8
Xylanases
AA10
AA10
+
CBM73
LPMOs

Figure 5: Phylogenomic relationships between Pseudomonas_E isolates, showing substrates on which the isolates had increased optical density, and the presence of CAZy genes which commonly have activities on different lignocellulosic polymers. Taxonomy was built based whole genomes aligned using SKA, and trees built using Gubbins. The scale bar represents the number of point mutations.

Table 1: Taxonomic classification of the sequenced isolates according to the Genome Taxonomy Data Base and the best BLAST match. Isolates which were unclassified at the family level had poor assembly statistics. The best BLAST match to the 16S rRNA gene (minimum 60% query coverage, "best" was quantified as the result with the largest mean of coverage and 0.1 times the number of aligned base pairs—this was done because full length 16S genes could not be recovered in most cases).

Isolate	Family	Genus	Species	Closest 16S rRNA species match	16S rRNA gene % identity	16S rRNA gene alignment length
P1F7a	*Enterobacteriaceae*	*Pantoea*	*Pantoea agglomerans*	*Pantoea agglomerans*	**96.871**	**703**
P1E12	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E neuropathica*	*Pseudomonas germanica*	98.754	**1525**
P1D2	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E neuropathica*	*Pseudomonas germanica*	99.521	**1461**
P1C4	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E neuropathica*	*Pseudomonas germanica*	**97.229**	**830**
P1H4	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E neuropathica*	*Pseudomonas soyae*	99.355	310
P1G4	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas allii*	98.333	**1020**
P3F6	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas atagonensis*	**96.129**	**620**
P3C10	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas bharatica*	100	305
P1B4	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas canadensis*	**95.749**	494
P1G11	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas cerasi*	**94.046**	**739**
P3F5	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas chlororaphis*	**97.358**	**530**
P3E6	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas chlororaphis*	**96.009**	426
P1C11	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas chlororaphis*	**91.188**	261
P1H6	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas chlororaphis*	100	246
P4E1	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas chlororaphis*	100	168
P1E5	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas extremaustralis 14-3*	**96.481**	**1421**
P1B7b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas extremaustralis 14-3*	**96.231**	398
P1C9	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas fildesensis*	99.644	281
P1D7b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas fildesensis*	98.024	253
P1F12	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas fluorescens*	99.606	**1521**
P1D6b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas fluorescens*	**96.94**	**719**
P1G12a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas fluorescens*	99.371	477

P1D8a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas germanica*	**92.917**	480
P3E8	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas germanica*	99.558	453
P1D9a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas germanica*	**95.946**	444
P2B10	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas germanica*	99.764	423
P1A3a_2	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas germanica*	99.443	359
P1E6	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas helmanticensis*	**97.52**	**1129**
P1D3b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas helmanticensis*	98.457	**1102**
P1B7a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas helmanticensis*	**97.108**	**830**
P3E1	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas helmanticensis*	97.917	**624**
P2H11	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas helmanticensis*	98.885	**538**
P3D12	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas hutmensis*	98.086	209
P1C3	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas hydrolytica*	100	82
P1D10	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas kairouanensis*	**96.97**	**825**
P1G12b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas kairouanensis*	98.872	**709**
P1H3a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas lurida*	**93.702**	**651**
P1C2	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas oleovorans*	**91.182**	**533**
P1D9b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas pisciculturae*	**90.16**	376
P1B5	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas pisciculturae*	99.666	299
P4E2	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas silesiensis*	99.605	**506**
P1F9	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	99.607	**1527**
P1H7	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	**96.139**	**1062**
P3E5	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	**93.774**	**1060**
P1D3a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	98.555	**969**
P3D6	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	98.846	**520**
P3G8	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	99.225	**516**
P1E1a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	**92.505**	467
P1E9b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	98.783	411
P3C11	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	99.625	267
P1H5	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	**89.583**	240
P1E1b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas soyae*	99.558	226

P1G6	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp. (unidentified)*	*Pseudomonas uvaldensis*	**96.078**	357
P1E9a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp000633255*	*Pseudomonas soyae*	100	96
P1A3b	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002018875*	*Pseudomonas cerasi*	**94.56**	**625**
P1B8	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002018875*	*Pseudomonas cerasi*	97.724	**615**
P1A3a_1	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002018875*	*Pseudomonas chlororaphis*	98.108	**740**
P1G7	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002018875*	*Pseudomonas coronafaciens*	99.671	304
P1D4a	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002018875*	*Pseudomonas fildesensis*	100	386
P1H11	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002018875*	*Pseudomonas petroselini*	**96.975**	**595**
P1G12a_2	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp002286815*	*Pseudomonas luteola*	100	128
P1D1	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp003097075*	*Pseudomonas pisciculturae*	99.375	160
P1A10	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp014851765*	*Pseudomonas kielensis*	98.21	**782**
P1A11	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp014851765*	*Pseudomonas migulae*	98.896	**634**
P1B3	*Pseudomonadaceae*	*Pseudomonas_E*	*Pseudomonas_E sp014851765*	*Pseudomonas pratensis*	100	355
P1H3b	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Agrobacterium arsenijevicii*	**93.625**	**549**
P1A4	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Agrobacterium arsenijevicii*	99.01	303
P1C1	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Agrobacterium arsenijevicii*	99.522	209
P4C2	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Agrobacterium fabacearum*	**96.309**	**1436**
P4D2	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Agrobacterium fabacearum*	**95.247**	**1031**
P1C10	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Agrobacterium fabacearum*	99.296	284
P1D7a	*Rhizobiaceae*	*Agrobacterium*	*Agrobacterium fabacearum*	*Peteryoungia desertarenae*	100	112
P4F1a	*Rhizobiaceae_A*	*Ochrobactrum_A*	*Ochrobactrum_A quorumnocens*	*Ochrobactrum quorumnocens*	94.07	371
P4D1	*Rhizobiaceae_A*	*Ochrobactrum_A*	*Ochrobactrum_A sp. (unidentified)*	*Pseudochrobactrum algeriensis*	100	118
P1D8b	*Unclassified*	*Unclassified*	*Unclassified*	*Lysobacter silvisoli*	100	39
P1H8	*Unclassified Bacteria*	*Unclassified Bacteria*	*Unclassified Bacteria*	*Pseudomonas jilinensis*	100	81
P1D6a	*Unclassified Bacteria*	*Unclassified Bacteria*	*Unclassified Bacteria*	*Pseudomonas toyotomiensis*	96.063	127
P4F4b	*Xanthomonadaceae*	*Stenotrophomonas*	*Stenotrophomonas sp003484865*	*Stenotrophomonas nematodicola*	99.748	397

The cut-offs for highlighting with bold text were: Pseudomonas 16S rRNA gene similarity < 97.69%, (https://journals.asm.org/doi/10.1128/mSystems.00704-21), *Pantoea and Agrobacterium* 16S rRNA gene similarity < 97%, 16S rRNA gene alignment length > 500 bp. Only gene similarities where the alignment criterion is satisfied are highlighted.

The isolates which passed all GTDB inclusion criteria were identified as *Agrobacterium fabacearum*, *Pseudomonas_E neuropathica*, *Pseudomonas_E sp002018875*, *Pseudomonas_E*

sp014851765, as well as 23 isolates from the GTDB genus *Pseudomonas_E* which were not assigned a species level taxonomy. In total, 65 of the 85 isolates were identified as belonging to the genus *Pseudomonas_E* which may be a result of the suitability of the sucrose-rich Czapek-dox medium used for cultivation of the isolates for species in this taxon. Alternatively, the prevalence of *Pseudomonas_E* may reflect their high cellular prevalence in the soil lignocellulolytic or plant material-associated community (de Lima Brossi *et al.*, 2016; Sah and Singh, 2016; Chiniquy *et al.*, 2021). We suggest that varying the medium used for *in situ* cultivation of microorganisms (*e.g.*, use of a minimal medium, or media with specific compositions used to target phenotypically constrained taxa) is an important method for increasing the broad taxonomic diversity recovered by isolation and cultivation techniques.

The *Pseudomonas* isolates had several genes which were classified into lignocellulolytic CAZy families (Figure 5). These included genes for cellulases (GH5), xylanases (GH8), LPMOs (AA10, AA10+CBM73), and carbohydrate binding modules (cellulose binding CBM32+CBM5, CBM13). All of the seven isolates from *Rhizobiaceae* were identified as the recently described *Agrobacterium fabacearum* which was isolated from the root nodules of plants in *Fabaceae* (Delamuta *et al.*, 2020). These isolates generally had increased OD on Avicel (5/7) and lignin (5/7). Additionally, all isolates had genes for GH8 (cellulase and xylanase activity), and all but two had genes for GH10 (xylanase or cellulase activity). *A. fabacearum* has been shown to utilise l-arabinose, d-xylose, l-xylose, methyl-β-d-xylopyranoside, d-galactose, d-mannose, l-rhamnose, and cellobiose, suggesting that the originally isolated strains may be able to grow on lignocellulose hydrosylates (Delamuta *et al.*, 2020). We did not detect any lignolytic CAZy genes in the *A. fabacearum* isolates, meaning that the mechanisms which gave them increase OD on lignin remain unidentified. Species of *Agrobacterium* with ligninolytic capabilities have previously been isolated with the genes implied in this process producing nonheme chloroperoxidase and benzaldehyde dehydrogenase. Aromatic ring-oxidizing genes such as 4-hydroxybenzoate polyprenyltransferase and 4-hydroxyphenylacetate 3-monooxygenase (Faisal *et al.*, 2021). The two isolates belonging to the GTDB genus *Rhizobiaceae_A*, one of which (P4D1) was an *Ochrobactrum_A* isolate which was unidentified at species level (genome completeness 55%), and the other (P4F1a) which was identified as *Ochrobactrum_A quorumnocens*, were both able to utilise lignin, but only the *Ochrobactrum_A quorumnocens* isolate could utilise CMC, and only P4D1 could utilise xylan (Figures 3, 5). No CAZy genes which implied lignin utilisation capability were found in either

genome. Xylan utilisation in the unidentified species was explained by presence of a GH8 (xylanase) gene in *Ochrobactrum* isolate P4D1, however, no lignocellulase genes were found in the *O. quorumnocens* isolate. A novel multi-copper polyphenol oxidoreductase enzyme which enables lignin degradation was recently found in *Ochrobactrum sp.* J10 (Chenxian Yang *et al.*, 2021), and so there is the possibility that the screening for lignocellulolytic genes missed less well classified lignocellulolytic systems. The only isolated member of *Enterobacteriaceae*, *Pantoea agglomerans* had increased growth only on CMC, and had one gene annotated as GH8. Isolate P3C5, an unidentified species of *Paracoccus* (*Rhodobacteraceae*) was able to utilise CMC, xylan, and lignin, but we detected no lignocellulase genes in this isolate, likely because of low genomic completeness (28% according to CheckM). Isolate P4F4b was identified as *Stenotrophomonas sp003484865* from the family *Xanthomonadaceae*. This isolate had two GH8 genes, although we did not detect increases to its growth on any of the lignocellulosic substrates relative to growth of the isolate on a medium containing no carbon source. In addition there were genomes with very low completeness values (between 0 and 10%; P1D6a, P1E10, P1H8, P3A9, P1D8b, P2A4) which were not given taxonomic assignments beyond "unclassified bacteria" we will not investigate these isolates further here.

4.4.2 Lignocellulase genes are poor predictors of growth on lignocellulolytic substrates

The large number of *Pseudomonas* isolates meant it was possible to identify patterns in phenotypic variation as a result of natural genotypic variation across the genus. We wanted to see how the number of genes which *a priori* likely have known relevant lignocellulolytic functions affected the growth of *Pseudomonadaceae* isolates on each substrate.

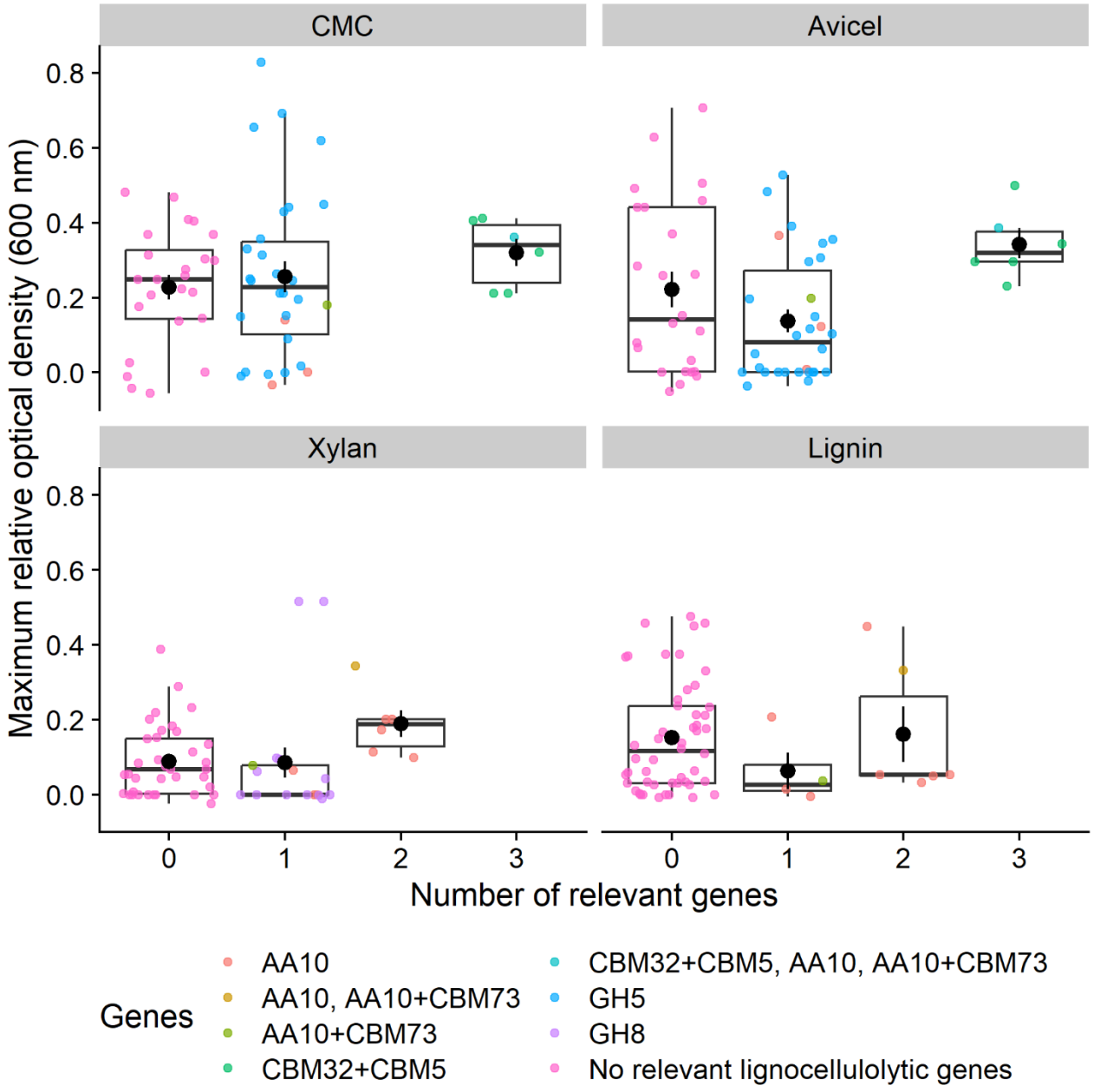

Figure 6: Maximum growth of isolates measured by optical density against number of lignocellulolytic genes which should be relevant to utilisation of the lignocellulosic polymer which acts as a sole carbon source. Except for with xylanases and AA10 genes, increasing numbers of lignocellulase genes does not increase growth on lignocellulosic substrates. Maximum predicted optical density according to population growth models relative to the same isolate grown in media with no carbon source, adjusting for medium optical density.

The number of extracellular xylanase genes (GH8, AA10, AA10+CBM73) influenced maximum OD on xylan (Kruskal-Wallis: χ^2_2 = 9.84, p = 0.007), with two extracellular xylanase genes giving increased maximum OD over zero or one xylanase gene. For individual gene annotations, this same result was observed for AA10 (Kruskal-Wallis: χ^2_2 = 5.23, p = 0.07; Dunn's test 2 genes *versus* 0 genes: z = -2.25, p = 0.02, Dunn's test 2 genes *versus* 1 gene: z = -1.73, p = 0.08; medians and interquartile ranges: 0 genes = 0.05 (0.12), 1 gene = 0.03 (0.13), 2 genes = 0.17 (0.09)), although no

other genes had a detectable effect. Increasing numbers of extracellular cellulolytic genes (GH5, CBM32+CBM5, AA10, AA10+CBM73) were also associated with higher maximum optical densities for Avicel (Kruskal-Wallis: χ^2_1 = 4.31, p = 0.050), with presence of two extracellular cellulolytic genes being associated with increased maximum OD, relative to with one (Dunn's test: z = -2.35, p = 0.019). This is mostly driven by the positive effect of CBM32+CBM5 genes on maximum OD (Kruskal-Wallis: χ^2_1 = 4.31, p = 0.038; medians and interquartile ranges: 0 genes = 0.11 (0.34), 1 gene = 0.32 (0.08)), as the only other significant association with maximum OD on Avicel was a negative relationship with extracellular GH5 presence (Kruskal-Wallis: χ^2_1 = 4.26, p = 0.039). Increasing numbers of relevant extracellular genes for the degradation of CMC or lignin (ligninases: AA10, AA10+CBM73) had no effect on the maximum optical density of the isolates on these substrates.

While we detected a significant effect on the median values, the relationship between number of relevant lignocellulolytic genes and maximum optical density on a substrate was highly variable across isolates, resulting from the fact that growth on a carbon source is a complex trait which is controlled by many interacting genes and transcriptional, translational, and post-translational factors. Because of this, the finding of high variation in maximum optical density is not surprising. Additionally, we tested only a single type of each polymer (xylan from beechwood, kraft lignin), whereas in soils there are many forms of these polymers which the genome-encoded lignocellulases may have increased rates of activity against. Further, the energetic cost of increased lignocellulase gene load, and increased transcription of lignocellulases could incur a reduction to microbial growth rate in they are highly expressed (Malik *et al.*, 2020), constituting a flaw in our initial hypothesis about increasing growth rate with more lignocellulase genes. Additionally, novel lignocellulase mechanisms, and more certainly, recently discovered ligninase mechanisms (Chenxian Yang *et al.*, 2021) in some of the isolates may have gone undetected due to the incompleteness of the databases used—transcriptomic analyses of these isolates could be used to check for any novel or recently discovered lignocellulolytic mechanisms. To gain insight into the genes which influence the observed growth traits, we employed pan-genome-wide association (pan-GWA) of genes to phenotypes.

4.4.3 Alternate cellular functions are associated with increased growth on different plant cell wall polymers in *Pseudomonas*

4.4.3.1 Broad metabolic trends for all substrates and analysis of GWA performance

In the absence of any strong relationships between the number of lignocellulase genes in a genome, and maximum relative growth on lignocellulosic polymers, we utilised the large number of genome sequenced *Pseudomonas_E* to conduct a pan-GWA with the aim of furthering our understanding of the ways in which the *Pseudomonadas_E* isolates utilise the lignocellulosic substrates. Pan-GWA analysis searches the pangenome of the isolates for correlations between presence/absence of accessory genes and presence/absence of the phenotype of interest. This identified variable numbers of genes which significantly correlated with utilisation of each of the lignocellulosic polymers in the 68 *Pseudomonadaceae* isolates. Utilisation of CMC was associated with 9 genes (7 unique annotations, 3 genes with no assigned function, 38 genes before phylogenetic and permutation tests), Avicel with 35 (15 unique annotations, 21 genes with no assigned function; 3523 genes before phylogenetic and permutation tests), xylan with 217 (59 unique annotations, 158 genes with no assigned function, 426 genes before phylogenetic and permutation tests), and lignin with 64 genes (42 unique annotations, 19 genes with no assigned function, 136 genes before phylogenetic and permutation tests). Despite the high number of significant trait-associated genes, there were no genes where the models of substrate utilisation and gene presence/absence aligned perfectly (Figure 7). More specifically, we observed a high proportion of genes where the presence of correlated genes was not accompanied by substrate utilisation (low positive predictive value), as well as absence of correlated genes where the substrate utilisation phenotype was observed (low negative predictive value). There were a high proportion of gene annotations of the correlated genes which were unlikely to be causative agents of the hydrolysis or oxidation of the substrates, or proteins which indirectly aid with this process.

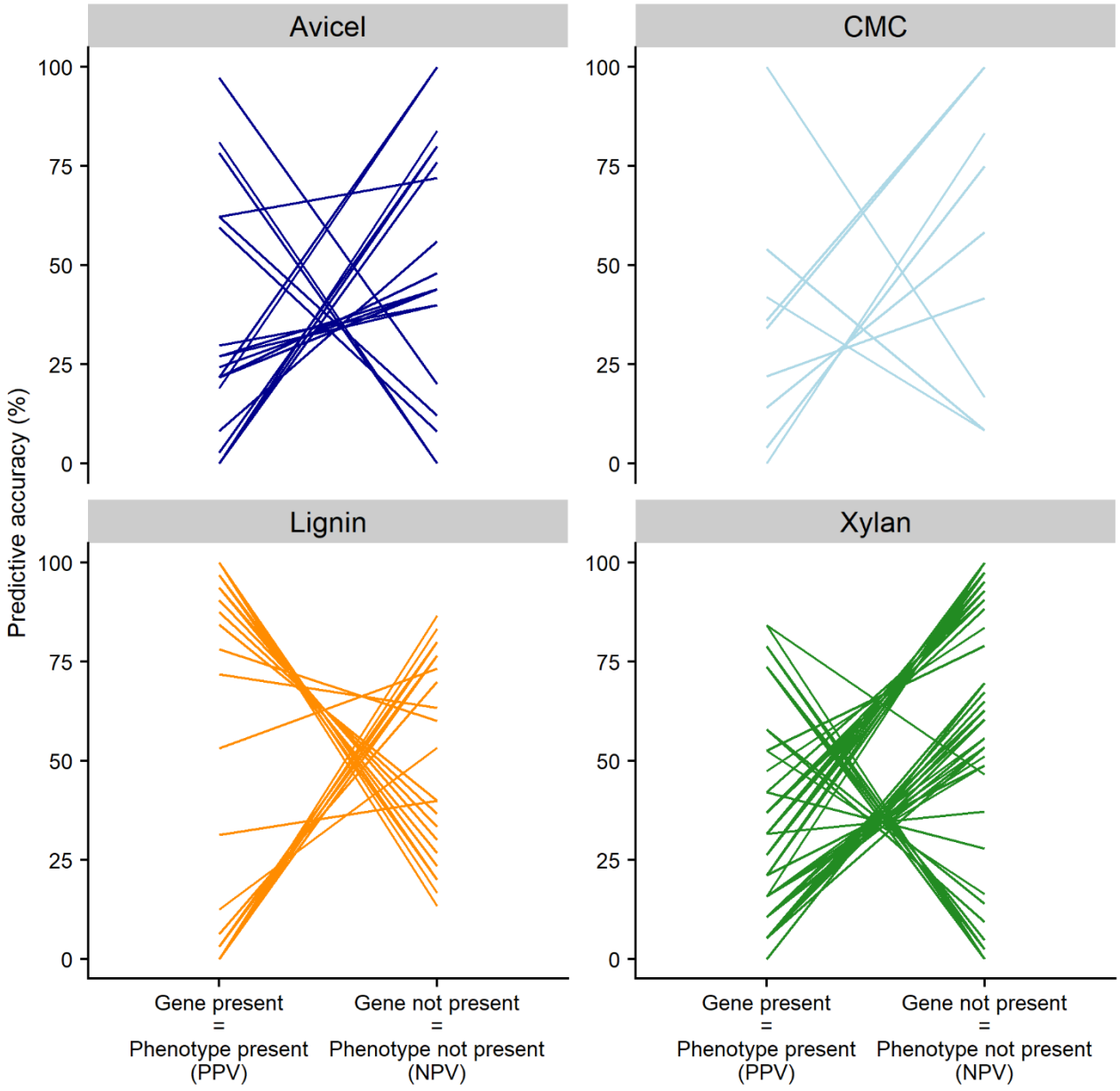

Figure 7: Evaluation of pan-genome-wide association in finding causative genes for substrate utilisation. Lines link the positive predictive value (PPV) and negative predictive value (NPV) for each gene which had a permutational p value less than 0.05, for each substrate.

There were no 'silver bullet' genes in the *Pseudomonadaceae* isolates, with no genes having both a positive predictive value and a negative predictive value above 80%, there were only three genes where both values were equal to or above 60%. These genes were the lignin-utilisation correlated *rne* (PPV = 78%, NPV = 60%) and *ribA* (PPV = 72%, NPV = 63%), which encode ribonuclease E, riboflavin biosynthesis protein, respectively, and the Avicel utilisation correlated *tldD* (PPV = 62%, NPV = 72%) which encodes Metalloprotease TldD. The products of these genes are essential for many cellular processes, but do not have obvious ties to lignocellulolytic activity. Riboflavin

(vitamin B2) is a precursor for coenzymes of many crucial cell processes including the metabolism of carbohydrates, lipids, ketone bodies, and proteins (Averianova *et al.*, 2020). Early work identified increased rates of cellulose degradation by rumen communities with the addition of B vitamins, including riboflavin (Hall, Cheng and Burroughs, 1955). The accessory *rne* gene for Ribonuclease E was correlated with lignin utilisation; it forms part of a membrane-bound degradosome, and plays the key role in the processing of stable-RNA and the decay of mRNA, exerting control over most cellular processes (Mackie, 2013). Finally, the *tldD* gene encodes proteins involved in the production of the antimicrobial compound microcin B17 (Ghilarov *et al.*, 2017).

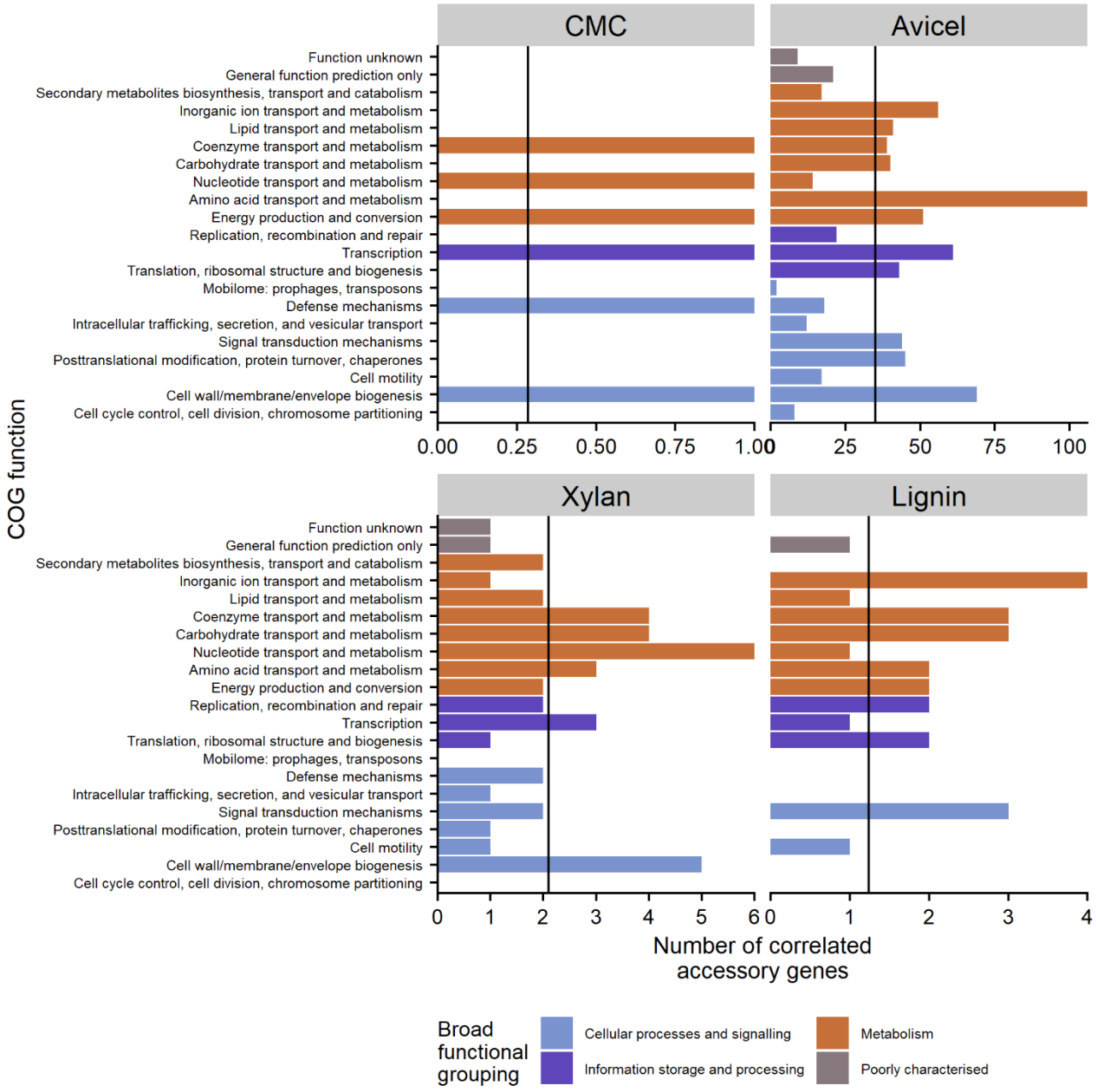

Figure 8: Importance of different COG categories for polymer utilisation. i.e., frequencies of the Cluster of Orthologous Groups (COG) proteins annotations from accessory genes which were correlated with increased optical density when isolates were grown on each carbon source shown above. Horizontal bars show the mean number of phenotype-correlated genes across COG categories for each observed phenotype.

Because of the poor-predictive power of individual genes in elucidating the mechanisms that the *Pseudomonas_E* isolates use for growth on particular substrates, and to understand how the different genomic strategies associated with utilisation of each lignocellulosic polymer, the COG functional classifications of the entire set of correlated genes (phylogenetically naïve $p < 0.05$) were obtained where annotations were available (Figure 8). Such an approach may produce many spurious associations between phenotype and genotype. High powered GWA studies on humans

commonly find correlations between thousands of genes and complex traits (as in this study). These associations often prove to be statistically robust, even when the means of causality is not obvious (Uffelmann *et al.*, 2021). Because the purpose of the present pan-GWA study is to generate testable hypotheses about how the accessory genes and pathways involved in how utilisation of polymers translates into growth in *Pseudomonas_E* isolates, and thus life-history strategy, spurious correlations are not a great concern. Metabolism was the COG supercategory with the largest number of phenotype-correlated genes for all single carbon sources in this experiment, followed by cellular processes, and signalling and information storage and processing, however, the relative gene frequencies in each COG category differed according to the single carbon source the microorganisms were grown on (xylan had a different gene distribution to Avicel: $\chi^2_{21} = 52.4$, $p < 0.001$). This approach of broad categorisation of many phenotype-correlated genes allows for understanding of the cellular strategies associated with the polygenic trait being studied. This is similar to the categorisation of genes in a genome into COG categories being used to characterise the functions of genomes (Galperin and Kolker, 2006; Galperin *et al.*, 2017). Amino acid transport and metabolism (E) and inorganic ion transport and metabolism (P) were the most abundant COG categories which correlated with increased growth on both Avicel and lignin, and both coenzyme transport and metabolism (H), and carbohydrate transport and metabolism (G) were in the topmost abundant COG categories for increased growth on xylan and lignin (Figure 8). These broad cellular processes likely represent the most important functions for substrate utilisation, cell growth, and reproduction for *Pseudomonas_E spp.* utilising lignocellulosic substrates as carbon sources. A wide range of phenotypes are controlled by these four broad COG groupings, including production and efficacy of lignocellulase enzymes among many others.

Increased growth on Avicel was additionally correlated with many genes involved in the regulation of transcription (K), while increased growth on xylan was mostly correlated with genes involved in nucleotide transport and metabolism (F), and increased growth on lignin was correlated with genes for signal transduction mechanisms (T). This suggests that for the isolates in the present collection, Avicel-utilising isolates relatively favour increased protein production, whereas the xylan-utilisers have accessory genomes which are relatively focussed on reproduction, and lignin-utilisers on signal transduction mechanisms and inorganic ion transport and metabolism (P; 75 genes, Figure 8) which includes the COG assignment from this dataset COG1496 for laccase domains—possibly corresponding to one of the AA10 genes found by dbCAN. Presence of this

domain, however, may in reality translate to weak, or strong, activity on aromatic compounds depending on the exact protein sequence of the active site (Beloqui *et al*., 2006).

4.4.3.2 Genes associated with growth on Avicel

We identified 196 genes from 54 known metabolic pathways which were correlated with increased growth on Avicel. The pathways with the most correlated genes were fatty acid biosynthesis (18 genes), tRNA modification (11 genes), isoleucine, leucine, valine biosynthesis (10 genes), cobalamin/B12 biosynthesis (8 genes), lysine biosynthesis (7 genes), phospholipid biosynthesis (7 genes), and RNA polymerase (7 genes). The gene products from the identified genes are known to be involved in a vast diversity of cellular processes. These include colonization of root tissues by *Pseudomonas* (Vílchez *et al.*, 2000), cell motility, biofilm formation, and pathogenicity against plants (Romeo, Vakulskas and Babitzke, 2013; Thakur *et al.*, 2013; Huertas-Rosales, Ramos-González and Espinosa-Urgel, 2016), production of cell membranes, cofactors secondary metabolites and siderophores, as well as quorum sensing (Yuan *et al.*, 2012), and electron transport in anaerobic respiration (Vo *et al.*, 2020), as well as glucose catabolism (Chavarría *et al.*, 2013). Some of these functions have clear general links to utilisation of lignocellulose (colonisation of root tissue, siderophore production) and cellular growth and reproduction, while for others the links are less clear and may represent spurious correlations.

4.4.3.3 Genes associated with growth on lignin

Lignin utilisation was most strongly correlated with genes for methionine import protein MetP and the periplasmic spermidine binding SpuE. The polyamine spermidine has been shown to increase growth of *Saccharomyces cerevisiae* in lignocellulosic hydrosylates, possibly by generally increasing the ability of cells to tolerate diverse environmental stresses (Kim *et al.*, 2015), we detected several polyamine binding genes correlated with lignin utilisation, all of which are known to increase tolerance to reactive oxygen species, such as those produced during lignin degradation through the Fenton reaction (Chou *et al.*, 2008). Other functions which correlated with lignin utilisation were production of type III secretion systems which are key virulence determinants for mammalian and plant hosts (Park *et al.*, 2010; Wu *et al.*, 2012). It seems reasonable that the presence of plant-based compounds as a sole carbon source would induce virulence factors so that the *Pseudomonas* isolate is primed to infect a plant host. Transport of siderophores was another related lignin-utilisation-correlated function. Siderophores act as mediators of oxidative reactions, and can enhancing depolymerisation of lignin through the Fenton reaction (Qin *et al.*,

2018). This mechanism seems likely because we did not detect any non-LPMO auxiliary activity genes in the present dataset using dbCAN2, meaning we likely did not detect any enzymatic mechanism for lignin decomposition; this contrasts with the Prokka annotation of a laccase domain (COG1496). Unfortunately, the Prokka output used to generate this assignment does not give information about the genomic location of this gene, and finding it to compare with the dbCAN2 output would require extensive further analyses. Additionally, we detected significant correlation with genes involved in pathways for the utilisation of lignin breakdown products (Pelmont *et al.*, 1989; Nonaka *et al.*, 2006; Y. Zheng *et al.*, 2019; Ling *et al.*, 2022). Finally, we detected correlations with genes involved in chemotaxis towards plants and the production of secondary metabolites (Grant, 2018; Hida *et al.*, 2020). Pan-GWA again appears to have found relevant genes for the utilisation of a specific lignocellulosic polymer, and provides interesting hypotheses about future genetic pathways to investigate.

4.4.3.4 Genes associated with growth on xylan

The most abundant COG categories associated with increased growth on xylan were nucleotide transport and metabolism (F) and cell wall/membrane/envelope biogenesis, suggesting that growth on xylan promotes cellular division, rather than production of many enzymes (Nuccio *et al.*, 2020). Included in this category is the *oprB_2* porin gene which is involved in transport of glucose and xylose in *Pseudomonas aeruginosa* (Trias, Rosenberg and Nikaido, 1988). Alink to xylan utilisation potential are correlations with in the xylose isomerase and *phnN* accessory genes. These are both involved in the metabolism of pentoses and form part of the pentose phosphate pathway. The pentose phosphate pathway is associated with efficient growth in *Pseudomonas aeruginosa* (Berger *et al.*, 2014), providing further evidence that the growth and reproduction life-history strategies discussed above are valid explanations for how the accessory genome influences the way *Pseudomonas_E* isolates utilise xylan. Another potentially important xylan-utilisation correlated gene was hemH. This gene encodes ferrochetalase, which regulates the production of siderophores and the active sites of heme-containing auxiliary activity enzymes (Baysse *et al.*, 2001). This is analogous to the function of the lignin-correlated *fhuA* and *yfhA* gene products which alter siderophore production, potentially implicating an oxidative mechanism of xylan utilisation for the present isolates. Another oxidative degradative enzyme for flavin-dependent trigonelline monooxygenase was correlated with xylan utilisation. This degrades the common plant root exudate trigonelline (Perchat *et al.*, 2018), although we did not find evidence that it has

more hemicellulose-specific degradative roles. Once again, accessory genes related to B vitamins (B6) were correlated with increased growth of the isolates on a lignocellulosic substrate, although this gene may have a generic, rather than a specific function *Pseudomonas aeruginosa* (Kim and Hong, 2016).

4.5 Conclusions and future directions

We utilised high-throughput *in situ* cultivation with the iChip, to isolate and cultivate lignocellulolytic soil microorganisms. We cultivated hundreds of viable pure lignocellulolytic cultures and screened them for the ability to utilise carboxymethyl cellulose in an agar screening assay. Purified isolates from the wells where a member of the isolated community could utilise CMC were subsequently tested in growth assays containing lignocellulosic compounds. Comparison of 65 whole-genome sequences, GTDB marker set gene sequences, and 16S rRNA sequences with closely related species suggests that we may have isolated 9 novel lignocellulolytic species from the genera *Pseudomonas*, *Pantoea*, *Ochrobactrum*, and *Agrobacterium*. Genes which encode lignocellulolytic enzymes were identified in the genomes of all but one of the 65 isolates, although there were some unexplained phenotypes which warrant further *in vitro* and genomic investigation. Pan-GWA identified genes, pathways, and the broad functions of genes which were associated with lignocellulolytic phenotypes, pointing to the different life-history strategies being favoured by microorganisms which utilise different lignocellulosic polymers. Interestingly, pan-GWA identified a single laccase gene as being correlated with growth on lignin, but identified no other relevant lignocellulase genes as being correlated with the growth traits. Further validation of these findings through transcriptional knockdown, genetic manipulation, and genome-scale metabolic modelling of these isolates' growth on lignocellulosic substrates would demonstrate how well or how poorly we understand the interplay of these genes, and further our knowledge on the genetic basis of species-species and species-resource interactions in ecosystems. Additionally, it seems that next generation physiology techniques hold significant promise for quantifying physiological traits of individual cells in their native environment before separating these for downstream manipulation and analysis (Hatzenpichler *et al.*, 2020). Combination of *in situ* and *in vitro* physiological testing of diverse isolates, single-cell genomics, and bulk metatranscriptomics could vastly improve our understanding of complex soil microbial community functioning.

4.6 Supplementary information

Figure S1:

https://github.com/fidlerdb/iChip_plots/blob/ce1c4ece0a49043acbb3e323763082cba3d6f4b0/All_isolates_substrate_utilisation.pdf

Figure S2:

https://github.com/fidlerdb/iChip_plots/blob/fbced52b9ca9436cbb0e8585507c2cad555a0d96/All_microbial_wells_growth_models_AIC_refined.pdf

5

Synthesis and future research

5.1 Introduction

Global challenges such as climate change, land use change, a growing population, and the mass extinction of species, all affect the way in which humans use, or interact with soils. Because of this, soils are linked with 7 of the United Nations' Sustainable Development Goals, including reducing inequality within and among countries, ending hunger, climate action, and sustaining biodiversity on land and at sea (Jayaraman *et al.*, 2021). Increasing food demand requires that high-yield farming practices are adopted globally. Provided these practices involve soil, they must aim to reduce the negative impact on soil organic carbon (SOC) content that is associated with agriculture, to ensure that productivity can continue into the future (Jayaraman *et al.*, 2021; Lowe, 2021). The benefits of soil organic carbon, particularly the lignocellulosic portion of this, are multi-fold, enabling food production, and buffering against biodiversity loss and rapid climatic change, through the provision of nutrients, habitats, and carbon stores which are relatively resistant to degradation. The degradation of lignocellulosic plant biomass is predominantly mediated by microorganisms, and the genes which afford them this capability. Our increasing understanding of microbially-mediated lignocellulose turnover is improving the predictive accuracy of global soil carbon flux models, which allow more effective management of the soil carbon resource (Wieder *et al.*, 2015; Kyker-Snowman *et al.*, 2020; Zhang *et al.*, 2020).

Throughout this thesis I have aimed to contribute to some of the remaining challenges associated with understanding the microbial utilisation of lignocellulose degradation in soils. These challenges were:

1) Knowledge about the relative contributions of species and broad groups to the degradative potential of lignocellulose in soils.
2) Knowledge about how the genetic potential of the community of soil microorganisms is affected by global changes.

3) Cultivation of diverse and uncultivated species for further diverse phenotypic characterisation so that we can begin to characterise and predict community dynamics with fewer unknowns.

This thesis has summarised and advanced our knowledge around these challenges throughout its chapters.

Chapter 1 gave a broad overview of the global importance of lignocellulose in soils and the organisms and mechanisms involved in lignocellulose degradation, using current literature to explore this. Chapter 2 utilised a decade-long plant-exclusion experiment to understand the effect of plant inputs on microbial community composition and their lignocellulolytic genes. We hypothesised that microbial community function in bare plots would transition towards microbiota with many genes for plant biomass degradation, when compared to grassland plots which had high labile carbon inputs. We explore this using a combination of metagenomics, fibre analysis and metabolomics. We show plant exclusion consistently favours genera in *Bacillales*, *Thermoproteota*, and diverse lineages of *Proteobacteria*, alters the repertoire of lignocellulolytic genes present, increases auxiliary activity relative abundance, and that taxonomic and genetic changes are not clearly linked. In chapter 3 we used the existing national scale UGRASS dataset and metagenomic sequences to investigate the effect of agricultural intensification on grassland soil microbial communities and their lignocellulolytic genes. Land use change was associated with drastic changes to the composition of microorganisms and lignocellulolytic genes, as well as large increases to the total relative abundance of lignocellulase genes. Despite this we found that agricultural intensification did not specifically benefit lignocellulolytic species, with most lignocellulolytic species showing no differences between land uses. We are sceptical that the measured increased proportional abundances of many diverse species and genes represent increased cellular/enzymatic abundance in arable soils, due to the large reductions to bacterial biomass. In chapter 4 we used high-throughput *in situ* cultivation techniques to cultivate thousands of microbial isolates, screened hundreds for lignocellulolytic capabilities, and genome sequenced 83. While these genome-sequenced isolates were not broadly diverse and are from commonly isolated genera (*Pseudomonas_E*, *Agrobacterium*, *Ochrobactrum_A, Paracoccus*, *Pantoea*, and *Stenotrophomonas*), we estimate a high rate of discovery of novel species, and have cultivated isolates potentially belonging to 7 new species. The many isolates from the GTDB genus *Pseudomonas_E* afforded us the ability to conduct a pangenome-wide association (pan-GWA)

study to associate genome-encoded accessory genes with *in vitro* phenotypic tests to confirm the utilisation of several lignocellulosic polymers. This provided valuable insights into the mechanisms of degradation of particular substrates by this genus. For instance, it seems likely that that the *Pseudomonas_E* isolates degraded lignin using weakly oxidative lytic polysaccharide monooxygenases which propagated lignin decomposition through the Fenton reaction using iron in the cultivation medium.

5.2 Synthesis of findings

5.2.1 Knowledge of the relative contributions of species and broad groups to the degradative potential and actual turnover of lignocellulose in soils.

As species- and gene-level characterisation of soils becomes cheaper, spatially explicit functional prediction of ecosystem processes based on the composition of genes and species will become a powerful technique for prioritising conservation efforts, habitats, and ecosystem service evaluation (Griffiths *et al.*, 2016). Firstly, however, scientists must agree upon the relative importance of different taxa and genes—and the link between the two—for ecosystem processes. This is particularly pertinent for lignocellulose degradation, as genes for its decomposition are widely taxonomically distributed, yet broad groups of microorganisms have differing lignocellulolytic potentials, and different strains within species show large variation within this also, as we show in chapter 3.

In chapter 2 the combination of metagenomics, metabolomics, and fibre analysis provides a multi-level view of how plant communities and carbon inputs affect the composition of lignocellulolytic genes in grassland soils, and point to the relative degradative capabilities of the microbial community. Chapter 3 examined the effect of agricultural intensification on the taxonomic origins of lignocellulase genes. In both chapters 2 and 3, the dominant taxa *Actinobacteria*, *Proteobacteria*, *Bacteroidota* and *Planctomycetes* contributed the majority of lignocellulase genes, with *Bacillota* and *Acidobacteria* being the next most important contributors in chapters 2 and 3 respectively. Chapter 2 showed that there were more cellulase, xylanase, and carbohydrate binding module genes from *Acidobacteria* than there were from phyla with similar abundances as detected by metagenomics, likely because *Acidobacteria* have a high genomic content of lignocellulase genes (Kalam *et al.*, 2020). *Acidobacteria* may therefore act as keystone species in lignocellulosic soil microbial communities. Agricultural intensification affected the relative abundance of lignocellulase genes from differing taxonomic origins. There were notably more

lignocellulase genes from *Planctomycetes, Proteobacteria,* and *Bacteroidota* in arable soil than in extensive grasslands. The increases to relative abundance however must be balanced with the knowledge that the number of bacterial cells per gram of arable soil was almost reduced to one third of the value in extensive grasslands, meaning that the importance of compositional changes for functioning may be reduced by the reduction to the number of cells—the realised functional implications of this require further testing.

In chapter 4, I isolated many lignocellulolytic strains from the genus *Pseudomonas_E*, as well as several other taxa with lignocellulolytic capabilities. Because of the lignocellulosic enrichment, dilution to extinction, and phenotypic screening techniques used, it is likely that members of *Pseudomonas_E* are some of the major degraders of plant cell wall polymers in grassland soils—although the cultivation medium used will provide some bias. Data from chapters 2 and 3 show *Pseudomonas* in the top 5 and 15 genera in terms of lignocellulase gene relative abundance, respectively. Minimal effective lignocellulolytic communities with plastic degrading capabilities have been isolated from forest soils which include members of *Pseudomonas_E* and *Ochrobactrum* which we also isolated in chapter 4 (Díaz Rodríguez *et al.*, 2022), adding some support to this claim.

In all study systems across the chapters, the importance of dominant taxa for lignocellulose decomposition potential became clear, as pointed to by independent DNA and PLFA stable-isotope probing studies (Wilhelm *et al.*, 2019). This may however be a result of the trends that bulk metagenomics and cultivation studies highlight (enrichment in this study using buried hay before inoculation of the colonising cells into iChip devices should have biased the community towards degradative taxa), although the principle of microbial dominance controlling processes which can be mediated by broadly distributed species is worthy of further investigation. Indeed, quantifying the functional importance for the degradation of lignocellulose in soils of dominant groups, and the biomass of these taxa, remains a research priority for deepening our understanding and predictive abilities regarding soil management and carbon storage. For this, we suggest that metatranscriptomic analyses focusing on native SOC degradation potential, coupled with data about the biomass of particular groups, will help to advance knowledge in this area of soil microbial ecology. Combination of '-omics' and cultivation approaches will likely make significant contributions to our functional understanding of ecosystems in the coming years.

5.2.2 Knowledge about how the genetic potential for, and realised rate of, degradation of lignocellulose by microorganisms is affected by global changes.

Soils globally have the potential to recapture and sequester anthropogenic-derived carbon dioxide (CO_2) which is altering the global climate and accelerating biodiversity loss (Soto-Navarro *et al.*, 2020). Increased atmospheric CO_2 concentrations can reduce soil carbon storage potential by altering plant-microorganism relationships (Carney *et al.*, 2007). An understanding of whether agricultural practices have a similar effect may lead to utilisation of more sustainable farming practices. In chapters 2 and 3 I utilised contrasting soil management regimes from the same localities to observe changes to the community of soil microorganisms and genes in response to different management practices. An interesting observation is that both plant-exclusion (chapter 2) and agricultural intensification (chapter 3) increased soil microbial diversity and decreased soil microbial abundance. The seemingly most important common factor between these management strategies (undisturbed soil with no nutrient inputs *versus* tilled soil with nutrient inputs) is the reduction to soil organic carbon (SOC), which is a major driver of microbial biomass and diversity (Bastida *et al.*, 2021). Perhaps then the commonly cited intermediate disturbance hypothesis (Grime, 1973; Connell, 1978) is not the major cause of increased microbial diversity in agricultural grassland soils, but it is rather a product of reduced dominance of species which thrive in the niches provided by high SOC contents, and the widespread ability of other less competitive species to utilise the carbon that is remaining leading to increases. Another related hypothesis is that reduced dominance and biomass, coupled with similar sequencing depths of the DNA extracts increases the proportion of reads from rare, senescing, or dead cells as an artefact of using compositional data (Carini *et al.*, 2016). The true reason for increased diversity in these low SOC, low biomass, systems is likely a combination of these two processes—measurement of soil microbial community activity *via* temporally replicated metatranscriptomics in combination with biomass markers such as phospholipid fatty-acids could help to address this issue, allowing quantification of microbial populations as in macro- and meso-ecological studies. The proposed effect of SOC and microbial biomass on diversity is further evidenced by the short-term (1 year) plant exclusion treatment in chapter 2 having equal microbial diversity and similar biomass to the grassland treatments (George *et al.*, 2021), showing that plant exclusion *per se* does not quickly increase the measured diversity of soil microorganisms. Additionally, both systems saw large increases to the relative abundance of *Archaea*, particularly the ammonia oxidizing order

Nitrososphaerales, for which other studies provide evidence of increased cellular abundance and activity (Gattinger *et al.*, 2002; Leininger *et al.*, 2006; Du *et al.*, 2019). While we detected a gene for chitin degradation from this order, alterations to the nitrogen cycle (resulting from plant removal or fertilizer application) seem a more likely cause for their increase in both systems (Zhalnina *et al.*, 2013; Sheridan *et al.*, 2020).

Decade-long plant exclusion was associated with increased auxiliary activity gene relative abundance, and reduced cellulase diversity and richness relative to after one year of plant-exclusion—biasing the composition towards GH5 cellulolytic domains. A single year of plant-exclusion increased the diversity of xylanase genes, altered the composition of lignocellulase genes, and increased hemicellulose breakdown product abundances. By contrast, we detected increased overall lignocellulase and cellulase gene relative abundance in agricultural soils at the national scale, with the massively increased diversity of *Proteobacteria* and other groups likely being the cause. It is possible that while the two study systems have decreased SOC contents, the arable grasslands are more similar in physicochemical composition to the short-term plant exclusion treatment, having greater aeration (increased pore size and tillage), carbon inputs/standing stock, and nutrient availabilities. Such physicochemical similarities between short-term plant exclusion and arable, and therefore dissimilarities between arable systems and the long-term plant-exclusion treatment, may partly explain the difference in directionality of cellulase abundance in systems with similar SOC contents within this thesis. The well-known reduction to microbial biomass in arable systems (de Vries *et al.*, 2013; Sun *et al.*, 2016; Malik *et al.*, 2018) likely also has a role in determining the direction of these relationships. Further work to understand how microbial (ligno)cellulase gene abundance (and the relationship between compositional changes, changes to cellular numbers, and functionality) varies with differing soil physicochemical properties will be key for improving the accuracy of metagenome-informed soil carbon flux models.

The decreased total biomass and relative abundance of dominant taxa observed in chapter 3 in response to land use change represent only an extremely limited single facet of the complex changes being experienced by soil communities worldwide. Climatic changes, land-use changes, and biodiversity loss may all drive different additive or interacting changes to the microbial communities which govern the soil carbon resource. In chapter 3 we identify that the proportion of species with lignocellulase genes, and the metagenomic density of genes for lignocellulose

processing in grasslands is relatively resilient to agricultural intensification, but may be reduced slightly in arable ecosystems—concurring with enzymatic data (Frąc *et al.*, 2020). Discovering the ways in which microorganisms, lignocellulase genes and functions interact with land-use types, and the outcomes for ecosystems and global services will require concerted effort from governments and the research community.

5.2.3 Increasing the diversity of cultivated microbial species to improve characterisation and prediction of microbial community dynamics.

Functional classification of species and communities within ecosystems relies upon our knowledge of the function of proteins produced by microorganisms which are cultivated in laboratories. The outcomes of often complex phenotypic tests with knockout mutants (Uzman, 2003) or isolates with altered expression levels due to manipulation with altered CRISPR/Cas9 systems (Banta *et al.*, 2020) are used to build this knowledge base. The impressive catalogue of gene functions built by geneticists and microbiologists unfortunately does not scratch the surface of the diversity of microorganisms and genes present in ecosystems. One of the largest commercial strain collections in the world, the DMSZ German collection of Microorganisms (https://www.dsmz.de/), has bacterial strains from 20,324 species, from 3,698 genera belonging to 42 phyla. Metagenome-assembled genomes and single-cell amplified genomes have expanded the number of known species with high quality bacterial genome sequences to 62,291 with these belonging to 15,342 genera and 148 phyla (Parks *et al.*, 2021), however, global estimates of microbial diversity predict up to one trillion (10^{12}) species (Locey and Lennon, 2016). If we are to have a hope of understanding ecosystem functioning at the genetic level, the proportion of cultivated and phenotypically characterised species clearly needs to increase.

In chapter 4 of this thesis I utilised high-throughput *in situ* cultivation on lignocellulosic enrichments with multiple cell detachment methods. After isolate purification by dilution-to-extinction yielded 304 isolates, I identified 173 isolates which could utilise at least one of the lignocellulosic polymers as a sole carbon source. Of these lignocellulolytic isolates, 83 were taxonomically and genomically characterised. We found 7 genome-resolved evolutionary groups (containing multiple isolates) which had no species-level assignment in the GTDB. While these 7 potentially novel species constitute minor contributions to the global paucity of knowledge about microorganisms, the high proportion of novel species cultivated in this small study clearly shows the potential for high-throughput *in situ* cultivation to increase the number of cultivated species.

Achieving the cultivation of diverse and novel microbial species will bring a range of biotechnological and biomedical advances (antibiotics, enzymes for biotechnology, increased understanding of pathogenicity), as well as increasing our understanding of ecosystems. Chapter 4 is limited by the single sucrose rich medium type used and the relatively small sampling effort. These factors may have biased the likelihood of cultivating particular groups, however, further research using different medium types could increase the broad diversity of recovered lignocellulolytic microorganisms. Initial screens of the 16S rRNA gene sequence of cultivated microorganisms would be a cost-effective method for first selecting diverse community members before genome sequences are obtained, streamlining this process.

The analyses of genomic content highlighted the complexity of the relative growth phenotype I measured. The individual lignocellulase genes were poor predictors of microbial growth, which depends upon many interrelated pathways. Whilst this and other sources of variation such as initial inoculum cell number may have affected our findings somewhat, the combination of phenotypic and genotypic measurements has proven a useful initial screening step to identify microorganisms which can grow in media with lignocellulosic polymers as a sole carbon source. This was confirmed by the presence of lignocellulase genes in the sequenced genomes, giving a theoretical basis for our findings. The pan-GWA study identified a broad range of gene annotations associated with utilisation of different lignocellulosic elements. Genes for protein production, cellular division, expression, and other cellular processes all correlated with different substrate utilisation statuses, with the diversity in annotation of the correlated genes mimicking results from high-powered GWA studies which measure complex phenotypes in humans (Uffelmann *et al.*, 2021). The complexity of microbial growth (measured *via* optical density in liquid culture) as a phenotype, and the differences in the broad functions of the correlated genes highlights how utilisation of different elements of lignocellulose may be more profitable strategies for species with different life-history strategies. This parallels our interpretation of how differences in the proportion of indicator species for a land-use type with different functional classes of lignocellulase gene differ with native SOC content according to life-history strategy. More specific tests, such as tests for production of breakdown products may be more suitable methods for screening for lignocellulase activity, and the identification of causative genes and resultant metabolic pathways, however, there is still merit in understanding the total pathway between utilisation of a compound and population growth for microbial ecology. Careful thought and

further research must be given to the way in which we translate functional annotation of genotypes into ecological interpretations. Additionally, it shows the value that genome-scale metabolic models will have in predicting these in the long term, but highlights how far away we are from understanding the microbial community dynamics of complex ecosystems.

5.3 Concluding remarks

This thesis provides novel insights into the interactions of soil microorganisms with their environment. It answers fundamental questions about how the presence of plants, plant-derived carbon inputs, and agricultural intensification shape microbial communities and the composition of genes at local and national scales. It demonstrates how closely related species and strains of *Pseudomonas* utilise non-shared genes to effectively utilise different long-chain lignocellulosic polymers for growth and reproduction, and how high-throughput *in situ* cultivation can increase the proportion of novel species discovered. This knowledge contributes to the scientific understanding of the microorganisms which help to regulate our climate, and whose lignocellulolytic enzymes uphold multiple global multi-billion dollar industries.

To further our knowledge of this critical system, we must address the major knowledge and resource gaps about lignocellulose degradation in soil which have become apparent during the writing of this thesis. Firstly, there needs to be a concerted effort to cultivate and phenotypically categorise soil microorganisms to increase the cultivated diversity. Combination of the iChip with other successful cultivation advances should allow for rapid progress in this area. Secondly, there is a need to build genome-scale metabolic models of these microorganisms and validate the results using mock communities, to build a foundation of understanding about the genetic functioning and dynamics of complex microbial communities. Advances in machine-learning aided high-throughput cultivation and characterisation (Huang *et al.*, 2023), single-cell techniques (Hatzenpichler *et al.*, 2020), and *in silico* prediction of function from primary protein sequences using deep learning (Jumper *et al.*, 2021) should allow for rapid progress in this area. Thirdly, global soil datasets with appropriate metadata should be used to understand how soil microorganisms and the relative abundance of their lignocellulolytic genes relates to soil properties, climates, and land use categories. Fourthly, these relationships should be tested with a dataset which quantifies actual rates of carbon utilisation through respiration, microbial carbon use efficiency, stable isotope probing, and enzymatic assays to begin parametrising metagenome-informed soil carbon flux models with known error rates on samples which were not used to build

the models. Fifthly, to predict soil carbon fluxes globally, the information gained from the previous steps should be combined to predict soil SOC contents, carbon turnover and incorporation rates, and associations with different soil physicochemical properties in different land-use categories.

Societally, the discovery and cultivation of novel microorganisms will lead to greater understanding of soil ecosystems, giving biotechnological, biomedical, and geopolitical advances. Soil carbon management strategies which consider how microbial communities in different land uses interact with the carbon stocks will provide greater clarity of direction for governments working towards sustainable food production potential as the human population grows. Effective soil carbon management strategies will hopefully act as a major contributing factor in halting the mass extinction of species, and maintenance of ecosystems which support Earth's biodiversity, as well as the ecosystem services associated with soil and its biodiversity. Global sustainable soil management requires scientists of many disciplines, policy makers, marketing experts, large corporations, farmers, and politicians to work closely and cooperatively.

References

Abdallah, R.Z., Wegner, C.E. and Liesack, W. (2019) 'Community transcriptomics reveals drainage effects on paddy soil microbiome across all three domains of life', *Soil Biology and Biochemistry*, 133. Available at: https://doi.org/10.1016/j.soilbio.2019.01.023.

A'Bear, A.D. *et al.* (2014) 'Interactive effects of temperature and soil moisture on fungal-mediated wood decomposition and extracellular enzyme activity', *Soil Biology and Biochemistry*, 70, pp. 151–158. Available at: https://doi.org/10.1016/j.soilbio.2013.12.017.

Adsul, M. *et al.* (2020) 'Designing a cellulolytic enzyme cocktail for the efficient and economical conversion of lignocellulosic biomass to biofuels', *Enzyme and Microbial Technology*, 133, p. 109442. Available at: https://doi.org/https://doi.org/10.1016/j.enzmictec.2019.109442.

Agger, J.W. *et al.* (2014) 'Discovery of LPMO activity on hemicelluloses shows the importance of oxidative processes in plant cell wall degradation', *Proceedings of the National Academy of Sciences*, 111, pp. 6287–6292. Available at: https://doi.org/10.1073/pnas.1323629111.

Ahmad, M. *et al.* (2010) 'Development of novel assays for lignin degradation: comparative analysis of bacterial and fungal lignin degraders', *Molecular BioSystems*, 6, pp. 815–821. Available at: https://doi.org/10.1039/B908966G.

Akram, F. *et al.* (2016) 'Cloning with kinetic and thermodynamic insight of a novel hyperthermostable β-glucosidase from *Thermotoga naphthophila* RKU-10Twith excellent glucose tolerance', *Journal of Molecular Catalysis B: Enzymatic*, 124, pp. 92–104. Available at: https://doi.org/10.1016/j.molcatb.2015.12.005.

Akram, F. *et al*. (2022) 'Genus Thermotoga: A valuable home of multifunctional glycoside hydrolases (GHs) for industrial sustainability', *Bioorganic Chemistry*, 127, 105942. Available at: https://doi.org/10.1016/j.bioorg.2022.105942.

Álvarez, C., Reyes-Sosa, F.M. and Díez, B. (2016) 'Enzymatic hydrolysis of biomass from wood', *Microbial Biotechnology*, pp. 149–156. Available at: https://doi.org/10.1111/1751-7915.12346.

Amann, R.I., Ludwig, W. and Schleifer, K.H. (1995) 'Phylogenetic identification and in situ detection of individual microbial cells without cultivation.', *Microbiological Reviews*, 59, pp. 143–69. Available at: https://doi.org/10.1016/j.jip.2007.09.009.

Amore, A. *et al.* (2013) 'Industrial waste based compost as a source of novel cellulolytic strains and enzymes', *FEMS Microbiology Letters*, 339, pp. 93–101. Available at: https://doi.org/10.1111/1574-6968.12057.

Anderson, I. *et al.* (2012) 'Genomics of aerobic cellulose utilisation systems in *Actinobacteria*', *PLoS ONE*, 7, 39331. Available at: https://doi.org/10.1371/journal.pone.0039331.

Anderson, K.L. and Salyers, A.A. (1989) 'Genetic evidence that outer membrane binding of starch is required for starch utilisation by *Bacteroides thetaiotaomicron'*, *Journal of Bacteriology*, 171, pp. 3199–3204. Available at: https://doi.org/10.1128/jb.171.6.3199-3204.1989.

Arantes, V. *et al.* (2011) 'Lignocellulosic polysaccharides and lignin degradation by wood decay fungi: The relevance of nonenzymatic Fenton-based reactions', *Journal of Industrial Microbiology and Biotechnology*, 38, pp. 541–555. Available at: https://doi.org/10.1007/s10295-010-0798-2.

Arantes, V. and Goodell, B. (2014) 'Current understanding of brown-rot fungal biodegradation mechanisms: A review', *ACS Symposium Series*, 1158, pp. 3–21. Available at: https://doi.org/10.1021/bk-2014-1158.ch001.

Araújo, W.J. *et al.* (2020) 'Microbial Culture in Minimal Medium With Oil Favors Enrichment of Biosurfactant Producing Genes.', *Frontiers in Bioengineering and Biotechnology*, 8, 962. Available at: https://doi.org/10.3389/fbioe.2020.00962.

Armbruster, M. *et al.* (2021) 'Bacterial and archaeal taxa are reliable indicators of soil restoration across distributed calcareous grasslands', *European Journal of Soil Science*, 72, pp. 2430–2444. Available at: https://doi.org/https://doi.org/10.1111/ejss.12977.

Arntzen, M.Ø. *et al.* (2017) 'Outer membrane vesicles from *Fibrobacter succinogenes* S85 contain an array of carbohydrate-active enzymes with versatile polysaccharide-degrading capacity', *Environmental Microbiology*, 19, pp. 2701–2714. Available at: https://doi.org/10.1111/1462-2920.13770.

Aronesty, E. (2011) 'ea-utils: Command-line tools for processing biological sequencing data'. Available at: https://expressionanalysis.github.io/ea-utils/ .

Ausec, L. *et al.* (2011) 'Bioinformatic analysis reveals high diversity of bacterial genes for laccase-like enzymes', *PLoS ONE*, 6, 2572. Available at: https://doi.org/10.1371/journal.pone.0025724.

Averianova, L.A. *et al.* (2020) 'Production of Vitamin B2 (Riboflavin) by Microorganisms: An Overview', *Frontiers in Bioengineering and Biotechnology*, 8, 570828. Available at: https://doi.org/10.3389/fbioe.2020.570828.

Aspeborg, H. *et al*. (2012) 'Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5)'. *BMC Evol Biol* 12, 186. Available at: https://doi.org/10.1186/1471-2148-12-186

Baker, B.J. *et al.* (2020) 'Diversity, ecology and evolution of Archaea', *Nature Microbiology*, 5, pp. 887–900. Available at: https://doi.org/10.1038/s41564-020-0715-z.

Baker, P.W., Charlton, A. and Hale, M.D.C. (2019) 'Fibre degradation of wheat straw by *Pleurotus erygnii* under low moisture conditions during solid-state fermentation', *Letters in Applied Microbiology*, 68, pp. 182–187.

Ballhausen, M.-B. *et al.* (2015) 'Methods for Baiting and Enriching Fungus-Feeding (Mycophagous) Rhizosphere Bacteria', *Frontiers in Microbiology*, 6, 1416. Available at: https://doi.org/10.3389/fmicb.2015.01416.

Banta, A.B. *et al.* (2020) 'A High-Efficacy CRISPR Interference System for Gene Function Discovery in *Zymomonas mobilis*', *Applied and Environmental Microbiology*, 86, pp. e01621-20. Available at: https://doi.org/10.1128/AEM.01621-20.

Bastida, F. *et al.* (2021) 'Soil microbial diversity–biomass relationships are driven by soil carbon content across global biomes', *The ISME Journal*, 15, pp. 2081–2091. Available at: https://doi.org/10.1038/s41396-021-00906-0.

Baysse, C. *et al.* (2001) 'Impact of mutations in hemA and hemH genes on pyoverdine production by Pseudomonas fluorescens ATCC17400.', *FEMS Microbiology Letters*, 205, pp. 57–63. Available at: https://doi.org/10.1111/j.1574-6968.2001.tb10925.x.

Beloqui, A. *et al*. (2006) 'Novel polyphenol oxidase mined from a metagenome expression library of bovine rumen: biochemical properties, structural analysis, and phylogenetic relationships', *Journal of Biological Chemistry*, 281, pp. 22933–22942. Available at: https://doi.org/10.1074/jbc.M600577200.

Berdy, B. *et al.* (2017) 'In situ cultivation of previously uncultivable microorganisms using the ichip', *Nature Protocols*, 12, pp. 2232–2242. Available at: https://doi.org/10.1038/nprot.2017.074.

Berger, A. *et al.* (2014) 'Robustness and plasticity of metabolic pathway flux among uropathogenic isolates of *Pseudomonas aeruginosa*.', *PLoS one*, 9, 88368. Available at: https://doi.org/10.1371/journal.pone.0088368.

Berger, E. *et al.* (2007) 'Two noncellulosomal cellulases of *Clostridium thermocellum*, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically', *FEMS Microbiology Letters*, 268, pp. 194–201. Available at: https://doi.org/10.1111/j.1574-6968.2006.00583.x.

Berlemont, R. *et al.* (2014) 'Cellulolytic potential under environmental changes in microbial communities from grassland litter', *Frontiers in Microbiology*, 5, 639 . Available at: https://doi.org/10.3389/fmicb.2014.00639.

Berlemont, R. and Martiny, A.C. (2013) 'Phylogenetic Distribution of Potential Cellulases in Bacteria', *Applied and Environmental Microbiology*, 79, pp. 1545–1554. Available at: https://doi.org/10.1128/AEM.03305-12.

Bertini, L. *et al.* (2018) 'Catalytic Mechanism of Fungal Lytic Polysaccharide Monooxygenases Investigated by First-Principles Calculations', *Inorganic Chemistry*, 57, pp. 86–97. Available at: https://doi.org/10.1021/acs.inorgchem.7b02005.

Beylot, M.H. *et al.* (2001) 'The *Pseudomonas cellulosa* glycoside hydrolase family 51 arabinofuranosidase exhibits wide substrate specificity.', *The Biochemical Journal*, 358, pp. 607–614. Available at: https://doi.org/10.1042/0264-6021:3580607.

Bielińska, E. and Mocek-Płóciniak, A. (2012) 'Impact of the tillage system on the soil enzymatic activity', *Archives of Environmental Protection*, 38, pp. 75–82. Available at https://doi.org/10.2478/v10265-012-0006-8.

Bischof, R. *et al.* (2013) 'Comparative analysis of the *Trichoderma reesei* transcriptome during growth on the cellulase inducing substrates wheat straw and lactose', *Biotechnology for Biofuels*, 6. Available at: https://doi.org/10.1186/1754-6834-6-127.

Bissaro, B. *et al.* (2017) 'Oxidative cleavage of polysaccharides by monocopper enzymes depends on H2O2', *Nature Chemical Biology*, 13, pp. 1123–1128. Available at: https://doi.org/10.1038/nchembio.2470.

Boetzer, M. and Pirovano, W. (2012) 'Toward almost closed genomes with GapFiller', *Genome Biology*, 13, p. R56. Available at: https://doi.org/10.1186/gb-2012-13-6-r56.

Bollmann, A., Lewis, K. and Epstein, S.S. (2007) 'Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates', *Applied and Environmental Microbiology*, 73, pp. 6386–6390. Available at: https://doi.org/10.1128/AEM.01309-07.

Boraston, A.B. *et al.* (2004) 'Carbohydrate-binding modules: fine-tuning polysaccharide recognition', *Biochemical Journal*, 382, pp. 769–781. Available at: https://doi.org/10.1042/BJ20040892.

Bornscheuer, U., Buchholz, K. and Seibel, J. (2014) 'Enzymatic degradation of (ligno)cellulose', *Angewandte Chemie - International Edition*, 53, pp. 10876–10893. Available at: https://doi.org/10.1002/anie.201309953.

Bourbonnais, R. and Paice, M.G. (1990) 'Oxidation of non-phenolic substrates. An expanded role for laccase in lignin biodegradation', *FEBS Letters*, 267, pp. 99–102. Available at: https://doi.org/10.1016/0014-5793(90)80298-W.

Brandt, A. *et al.* (2013) 'Deconstruction of lignocellulosic biomass with ionic liquids', *Green Chemistry*, 15, p. 550. Available at: https://doi.org/10.1039/c2gc36364j.

Brandt, B.W. *et al.* (2004) 'Modelling microbial adaptation to changing availability of substrates', *Water Research*, 38, pp. 1003 - 1013. Available at: https://doi.org/10.1016/j.watres.2003.09.037.

Van Den Brink, J. and De Vries, R.P. (2011) 'Fungal enzyme sets for plant polysaccharide degradation', *Applied Microbiology and Biotechnology*, 91, pp. 1477–1492. Available at: https://doi.org/10.1007/s00253-011-3473-2.

Brooks, M.E. *et al.* (2017) '{glmmTMB} Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling', *The R Journal*, 9, pp. 378–400.

Brulc, J.M. *et al.* (2009) 'Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases', *Proceedings of the National Academy of Sciences*, 106, pp. 1948–1953. Available at: https://doi.org/10.1073/pnas.0806191105.

Brunecky, R. *et al.* (2013) 'Revealing nature's cellulase diversity: The digestion mechanism of *Caldicellulosiruptor bescii* CelA', *Science*, 342, pp. 1513–1516. Available at: https://doi.org/10.1126/science.1244273.

Buerger, S. *et al.* (2012) 'Microbial scout hypothesis and microbial discovery.', *Applied and Environmental Microbiology*, 78, pp. 3229–3233. Available at: https://doi.org/10.1128/AEM.07308-11.

Bugg, T.D.H. *et al.* (2011) 'Pathways for degradation of lignin in bacteria and fungi', *Natural Product Reports*, 12, pp. 1871–1960. Available at: https://doi.org/10.1039/c1np00042j.

Bull, I.D. *et al.* (2000) 'Detection and classification of atmospheric methane oxidizing bacteria in soil', *Nature*, 405, pp. 175–178. Available at: https://doi.org/10.1038/35012061.

Burnet, M.C. *et al.* (2015) 'Evaluating models of cellulose degradation by Fibrobacter succinogenes S85', *PLoS ONE*, 10, 143809. Available at: https://doi.org/10.1371/journal.pone.0143809.

Burns, R.G. and Dick, R.P. (2002) *Enzymes in the environment: activity, ecology, and applications*. Marcel Dekker.

Bushnell, B. (2014) 'BBMap A Fast, Accurate, Splice-Aware Aligner'.

Carbonetto, B. *et al.* (2014) 'Structure, Composition and Metagenomic Profile of Soil Microbiomes Associated to Agricultural Land Use and Tillage Systems in Argentine Pampas', *PLoS ONE*, 9, pp. 1–11. Available at: https://doi.org/10.1371/journal.pone.0099949.

Carder, J.H. (1986) 'Detection and quantitation of cellulase by Congo red staining of substrates in a cup-plate diffusion assay.', *Analytical Biochemistry*, 153, pp. 75–79. Available at: https://doi.org/10.1016/0003-2697(86)90063-1.

Carini, P. *et al.* (2016) 'Relic DNA is abundant in soil and obscures estimates of soil microbial diversity.', *Nature Microbiology*, 2, 16242. Available at: https://doi.org/10.1038/nmicrobiol.2016.242.

Carney, K.M. *et al.* (2007) 'Altered soil microbial community at elevated CO_2 leads to loss of soil carbon', *Proceedings of the National Academy of Sciences*, 104, pp. 4990–4995. Available at: https://doi.org/10.1073/pnas.0610045104.

Carriquiry, M.A., Du, X. and Timilsina, G.R. (2011) 'Second generation biofuels: Economics and policies', *Energy Policy*, 39, pp. 4222–4234. Available at: https://doi.org/10.1016/j.enpol.2011.04.036.

Carter, M.R. (2002) 'Soil Quality for Sustainable Land Management', *Agronomy Journal*, 94, pp. 38–47. Available at: https://doi.org/10.2134/AGRONJ2002.3800.

Castro, L. *et al.* (2016) 'Insights into structure and redox potential of lignin peroxidase from QM/MM calculations', *Organic and Biomolecular Chemistry*, 14, pp. 2385–2389. Available at: https://doi.org/10.1039/C6OB00037A.

Catucci, G. *et al.* (2020) 'Biochemical features of dye-decolorizing peroxidases: Current impact on lignin degradation.', *Biotechnology and Applied Biochemistry*, 67, pp. 751–759. Available at: https://doi.org/10.1002/bab.2015.

CAZypedia Consortium (2017) 'Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes', *Glycobiology*, 28, pp. 3–8. Available at: https://doi.org/10.1093/glycob/cwx089.

Chamberlain, S.A. and Szöcs, E. (2013) 'taxize: taxonomic search and retrieval in R', *F1000Research*, 2, 191. Available at: https://doi.org/10.12688/f1000research.2-191.v2.

Chavarría, M. *et al.* (2013) 'The Entner–Doudoroff pathway empowers *Pseudomonas putida* KT2440 with a high tolerance to oxidative stress', *Environmental Microbiology*, 15, pp. 1772–1785. Available at: https://doi.org/https://doi.org/10.1111/1462-2920.12069.

Chen, C.C. *et al.* (1997) 'Release of lignin from kraft pulp by a hyperthermophilic xylanase from *Thermatoga maritima*', *Enzyme and Microbial Technology*, 20, pp. 39–45. Available at: https://doi.org/10.1016/S0141-0229(97)82192-8.

Chen, C.Y. *et al.* (2013) 'Properties of the newly isolated extracellular thermo-alkali-stable laccase from thermophilic actinomycetes, *Thermobifida fusca* and its application in dye intermediates oxidation', *AMB Express*, 3, pp. 1–9. Available at: https://doi.org/10.1186/2191-0855-3-49.

Chen, J. *et al.* (2020) 'Soil carbon loss with warming: New evidence from carbon-degrading enzymes', *Global Change Biology*, 26, pp. 1944–1952. Available at: https://doi.org/10.1111/gcb.14986.

Cheng, L. *et al.* (2007) 'Dynamics of labile and recalcitrant soil carbon pools in a sorghum free-air CO2 enrichment (FACE) agroecosystem', *Soil Biology and Biochemistry*, 39, pp. 2250–2263. Available at: https://doi.org/10.1016/j.soilbio.2007.03.031.

Chiniquy, D. *et al.* (2021) 'Microbial Community Field Surveys Reveal Abundant Pseudomonas Population in Sorghum Rhizosphere Composed of Many Closely Related Phylotypes', *Frontiers in Microbiology*, 12, 598180. Available at: https://doi.org/10.3389/fmicb.2021.598180.

Chistoserdova, L. (2009) 'Functional Metagenomics: Recent Advances and Future Challenges', *Biotechnology and Genetic Engineering Reviews*, 26(1), pp. 335–352. Available at: https://doi.org/10.5661/bger-26-335.

Cholodny, N.G. (1934) 'A soil chamber as a method for the microscopic study of the soil microflora', *Archiv für Mikrobiologie*, 5, pp. 148–156. Available at: https://doi.org/10.1007/BF00409166.

Chou, H.T. *et al.* (2008) 'Transcriptome analysis of agmatine and putrescine catabolism in *Pseudomonas aeruginosa* PAO1', *Journal of Bacteriology*, 190, p. 1966—1975. Available at: https://doi.org/10.1128/jb.01804-07.

Christiansen, L., *et al*. (2020) 'A Multifunctional Polysaccharide Utilization Gene Cluster in Colwellia echini Encodes Enzymes for the Complete Degradation of κ-Carrageenan, ι-Carrageenan, and Hybrid β/κ-Carrageenan', *mSphere*, 5, 00792-19. Available at: https://doi.org/10.1128/msphere.00792-19.

Christopherson, M.R. *et al.* (2013) 'The Genome Sequences of *Cellulomonas fimi* and "*Cellvibrio gilvus*" Reveal the Cellulolytic Strategies of Two Facultative Anaerobes, Transfer of "*Cellvibrio gilvus*" to the Genus *Cellulomonas*, and Proposal of *Cellulomonas gilvus sp. nov*', *PLoS ONE*, 8, 53954. Available at: https://doi.org/10.1371/journal.pone.0053954.

Coleman, K. and Jenkinson, D.S. (1996) 'RothC-26.3 - A Model for the turnover of carbon in soil', in D.S. Powlson, P. Smith, and J.U. Smith (eds) *Evaluation of Soil Organic Matter Models*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 237–246.

Colpa, D.I., Fraaije, M.W. and Van Bloois, E. (2014) 'DyP-type peroxidases: A promising and versatile class of enzymes', *Journal of Industrial Microbiology and Biotechnology*, pp. 1–7. Available at: https://doi.org/10.1007/s10295-013-1371-6.

Conn, H.J. (1960) 'Staining procedures used by the Biological Stain Commission'. Maryland: Biological Stain Commission, University of Rochester Medical Center, Rochester, N.Y., by Williams & Wilkins.

Connell, J.H. (1978) 'Diversity in Tropical Rain Forests and Coral Reefs', *Science*, 199, pp. 1302–1310. Available at: https://doi.org/10.1126/science.199.4335.1302.

Costanza, R. *et al.* (2014) 'Changes in the global value of ecosystem services', *Global Environmental Change*, 26, pp. 152–158. Available at: https://doi.org/https://doi.org/10.1016/j.gloenvcha.2014.04.002.

Couradeau, E., Sasse, J., Goudeau, D. *et al*. (2019) 'Probing the active fraction of soil microbiomes using BONCAT-FACS'. Nat Commun 10, 2770. Available at: https://doi.org/10.1038/s41467-019-10542-0.

Cragg, S.M. *et al.* (2015) 'Lignocellulose degradation mechanisms across the Tree of Life', *Current Opinion in Chemical Biology*, 29, pp. 108–119. Available at: https://doi.org/10.1016/j.cbpa.2015.10.018.

Crowther, T.W. *et al.* (2016) 'Quantifying global soil carbon losses in response to warming', *Nature*, 540, pp. 104–108. Available at: https://doi.org/10.1038/nature20150.

Crowther, T.W. *et al.* (2018) 'Crowther *et al*. reply', *Nature*, 554, pp. E7–E8. Available at: https://doi.org/10.1038/nature25746.

Curtin, D., Peterson, M.E. and Anderson, C.R. (2016) 'pH-dependence of organic matter solubility: Base type effects on dissolved organic C, N, P, and S in soils with contrasting mineralogy', *Geoderma*, 271, pp. 161–172. Available at: https://doi.org/https://doi.org/10.1016/j.geoderma.2016.02.009.

Damon, C. *et al.* (2012) 'Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils', *PLoS ONE*, 7, 28967. Available at: https://doi.org/10.1371/journal.pone.0028967.

Das, N. *et al.* (2015) 'Progress in the development of gelling agents for improved culturability of microorganisms', *Frontiers in Microbiology*, 6, 698. Available at: https://doi.org/10.3389/fmicb.2015.00698.

Dashtban, M. *et al.* (2010) 'Fungal biodegradation and enzymatic modification of lignin', *International Journal of Biochemistry and Molecular Biology*, pp. 36–50.

Delamuta, J.R.M. *et al.* (2020) 'Genetic diversity of *Agrobacterium* species isolated from nodules of common bean and soybean in Brazil, Mexico, Ecuador and Mozambique, and description of the new species *Agrobacterium fabacearum sp. nov*.', *International Journal of Systematic and Evolutionary Microbiology*, 70, pp. 4233–4244. Available at: https://doi.org/https://doi.org/10.1099/ijsem.0.004278.

Delgado-Baquerizo, M., Maestre, F.T., *et al.* (2016) 'Microbial diversity drives multifunctionality in terrestrial ecosystems', *Nature Communications*, 7, p. 10541. Available at: https://doi.org/10.1038/ncomms10541.

Delgado-Baquerizo, M., Grinyer, J., *et al.* (2016) 'Relative importance of soil properties and microbial community for soil functionality: insights from a microbial swap experiment', *Functional Ecology*, 30, pp. 1862–1873. Available at: https://doi.org/https://doi.org/10.1111/1365-2435.12674.

Delgado-Baquerizo, M. *et al.* (2018) 'A global atlas of the dominant bacteria found in soil', *Science*, 359, pp. 320–325. Available at: https://doi.org/10.1126/science.aap9516.

Diamond, S. *et al.* (2019) 'Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms', *Nature Microbiology*, 4, pp. 1356–1367. Available at: https://doi.org/10.1038/s41564-019-0449-y.

Díaz Rodríguez, C.A. *et al.* (2022) 'Novel bacterial taxa in a minimal lignocellulolytic consortium and their potential for lignin and plastics transformation', *ISME Communications*, 2, p. 89. Available at: https://doi.org/10.1038/s43705-022-00176-7.

Din, N. *et al.* (1991) 'Non–Hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulase', *Nature Biotechnology*, 9, pp. 1096–1099. Available at: https://doi.org/10.1038/nbt1191-1096.

Divne, C. *et al.* (1994) 'The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*', *Science*, 265, pp. 524–528. Available at: https://doi.org/10.1126/science.8036495.

Du, Y. *et al.* (2019) 'Moderate Grazing Promotes Grassland Nitrous Oxide Emission by Increasing Ammonia-Oxidizing Archaea Abundance on the Tibetan Plateau', *Current Microbiology*, 76, pp. 620–625. Available at: https://doi.org/10.1007/s00284-019-01668-x.

Dudley, N. and Alexander, S. (2017) 'Agriculture and biodiversity: a review', *Biodiversity*, 18, pp. 45–49. Available at: https://doi.org/10.1080/14888386.2017.1351892.

Dunn, O.J. (1964) 'Multiple Comparisons Using Rank Sums', *Technometrics*, 6, pp. 241–252. Available at: https://doi.org/10.1080/00401706.1964.10490181.

Dworkin, M. and Gutnick, D. (2012), 'Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist', *FEMS Microbiology Reviews*, 36, pp. 364–379, Available at: https://doi.org/10.1111/j.1574-6976.2011.00299.x.

Earle, S.G. *et al.* (2016) 'Identifying lineage effects when controlling for population structure improves power in bacterial association studies.', *Nature Microbiology*, 1, 16041. Available at: https://doi.org/10.1038/nmicrobiol.2016.41.

Edwards, J.L. *et al.* (2010) 'Identification of carbohydrate metabolism genes in the metagenome of a marine biofilm community shown to be dominated by Gammaproteobacteria and Bacteroidetes', *Genes*, 1, pp. 371–384. Available at: https://doi.org/10.3390/genes1030371.

Eibinger, M. *et al.* (2014) 'Cellulose surface degradation by a lytic polysaccharide monooxygenase and its effect on cellulase hydrolytic efficiency', *Journal of Biological Chemistry*, 289, pp. 35929–35938. Available at: https://doi.org/10.1074/jbc.M114.602227.

Eichorst, S.A. and Kuske, C.R. (2012) 'Identification of cellulose-responsive bacterial and fungal communities in geographically and edaphically different soils by using stable isotope probing', *Applied and Environmental Microbiology*, 78, pp. 2316–2327. Available at: https://doi.org/10.1128/AEM.07313-11.

Eme, L. *et al.* (2017) 'Archaea and the origin of eukaryotes.', *Nature Reviews Microbiology*, 15, pp. 711–723. Available at: https://doi.org/10.1038/nrmicro.2017.133.

Espina, L. (2021) 'An approach to increase the success rate of cultivation of soil bacteria based on fluorescence-activated cell sorting', *PLoS ONE*, 15, e0237748. Available at: https://doi.org/10.1371/journal.pone.0237748.

Evans, S.G., Kelley, L.C. and Potts, M.D. (2015) 'The potential impact of second-generation biofuel landscapes on at-risk species in the US', *Global Change Biology Bioenergy*, 7, pp. 337–348. Available at: https://doi.org/10.1111/gcbb.12131.

Faisal, U.H. *et al.* (2021) 'Draft Genome Sequence of Lignin-Degrading *Agrobacterium* sp. Strain S2, Isolated from a Decaying Oil Palm Empty Fruit Bunch', *Microbiology Resource Announcements*, 10, pp. 00259-21. Available at: https://doi.org/10.1128/MRA.00259-21.

Falush, D. (2016) 'Bacterial genomics: Microbial GWAS coming of age', *Nature Microbiology*, 1, 16059. Available at: https://doi.org/10.1038/nmicrobiol.2016.59.

Fan, X. *et al.* (2021) 'Improved model simulation of soil carbon cycling by representing the microbially derived organic carbon pool', *The ISME Journal*, 15, pp. 2248–2263. Available at: https://doi.org/10.1038/s41396-021-00914-0.

FAO (2015) *Global soil status, processes and trends*, *Status of the World's Soil Resources*.

FAO (2020) *State of knowledge of soil biodiversity - Status, challenges and potentialities, Report*. Rome. Available at: https://doi.org/https://doi.org/10.4060/cb1928en.

FAO and ITPS (2021) *Recarbonizing global soils – A technical manual of recommended management practices. Volume 1 – Introduction and methodology*. Rome. Available at: https://doi.org/https://doi.org/10.4060/cb6386en.

Finzi, A.C. *et al.* (2011) 'Responses and feedbacks of coupled biogeochemical cycles to climate change: Examples from terrestrial ecosystems', in *Frontiers in Ecology and the Environment,* 9, 61–67. Available at: https://doi.org/10.1890/100001.

Floudas, D. *et al.* (2012) 'The Paleozoic Origin of Enzymatic Lignin Decomposition Reconstructed from 31 Fungal Genomes', *Science*, 336, pp. 1715–1719. Available at: https://doi.org/10.1126/science.1221748.

Frąc, M. *et al.* (2020) 'Structural and functional microbial diversity of sandy soil under cropland and grassland', *PeerJ*, 8, p. e9501. Available at: https://doi.org/10.7717/peerj.9501.

de França Passos, D., Pereira, N. and de Castro, A.M. (2018) 'A comparative review of recent advances in cellulases production by *Aspergillus*, *Penicillium* and *Trichoderma* strains and their use for lignocellulose deconstruction', *Current Opinion in Green and Sustainable Chemistry*, 14, pp. 60–66. Available at: https://doi.org/https://doi.org/10.1016/j.cogsc.2018.06.003.

Furukawa, T., Bello, F.O. and Horsfall, L. (2014) 'Microbial enzyme systems for lignin degradation and their transcriptional regulation', *Frontiers in Biology*, 9, pp. 448–471. Available at: https://doi.org/10.1007/s11515-014-1336-9.

Galperin, M.Y. *et al.* (2017) 'Microbial genome analysis: the COG approach', *Briefings in Bioinformatics*, 20, pp. 1063–1070. Available at: https://doi.org/10.1093/bib/bbx117.

Galperin, M.Y. *et al.* (2020) 'COG database update: focus on microbial diversity, model organisms, and widespread pathogens', *Nucleic Acids Research*, 49, pp. D274–D281. Available at: https://doi.org/10.1093/nar/gkaa1018.

Galperin, M.Y. and Kolker, E. (2006) 'New metrics for comparative genomics', *Current Opinion in Biotechnology*, 17, pp. 440–447. Available at: https://doi.org/https://doi.org/10.1016/j.copbio.2006.08.007.

Gänzle, M.G. and Follador, R. (2012) 'Metabolism of oligosaccharides and starch in lactobacilli: A review', *Frontiers in Microbiology*, 3, 340. Available at: https://doi.org/10.3389/fmicb.2012.00340.

Gattinger, A. *et al.* (2002) 'Microbial community structure varies in different soil zones of a potato field', *Journal of Plant Nutrition and Soil Science*, 165, pp. 421–428.

Gavrilov, S.N. *et al.* (2016) 'Isolation and characterization of the first xylanolytic hyperthermophilic euryarchaeon *Thermococcus sp*. strain 2319x1 and its unusual multidomain glycosidase', *Frontiers in Microbiology*, 7, 552. Available at: https://doi.org/10.3389/fmicb.2016.00552.

George, P.B.L. *et al.* (2019) 'Divergent national-scale trends of microbial and animal biodiversity revealed across diverse temperate soil ecosystems', *Nature Communications*, 10, p. 1107. Available at: https://doi.org/10.1038/s41467-019-09031-1.

George, P.B.L. *et al.* (2021) 'Shifts in Soil Structure, Biological, and Functional Diversity Under Long-Term Carbon Deprivation', *Frontiers in Microbiology*, 12, 2509. Available at: https://doi.org/10.3389/fmicb.2021.735022.

Gerganova, V. *et al.* (2015) 'Chromosomal position shift of a regulatory gene alters the bacterial phenotype.', *Nucleic Acids Research*, 43, pp. 8215–8226. Available at: https://doi.org/10.1093/nar/gkv709.

Gerlt, J.A. (2017) 'Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence–Function Space and Genome Context to Discover Novel Functions', *Biochemistry*, 56, pp. 4293–4308. Available at: https://doi.org/10.1021/acs.biochem.7b00614.

van Gestel, N. *et al.* (2018) 'Predicting soil carbon loss with warming', *Nature*, 554, pp. 4–5. Available at: https://doi.org/10.1038/nature25745.

Ghilarov, D. *et al.* (2017) 'The Origins of Specificity in the Microcin-Processing Protease TldD/E.', *Structure* , 25, pp. 1549-1561. Available at: https://doi.org/10.1016/j.str.2017.08.006.

de Gonzalo, G. *et al.* (2016) 'Bacterial enzymes involved in lignin degradation', *Journal of Biotechnology*, pp. 110–119. Available at: https://doi.org/10.1016/j.jbiotec.2016.08.011.

Gourlay, K. *et al.* (2015) 'The use of carbohydrate binding modules (CBMs) to monitor changes in fragmentation and cellulose fiber surface morphology during cellulase- And swollenin-induced

deconstruction of lignocellulosic substrates', *Journal of Biological Chemistry*, 290, pp. 2938–2945. Available at: https://doi.org/10.1074/jbc.M114.627604.

de Graaff, M.-A. *et al.* (2019) 'Chapter One - Effects of agricultural intensification on soil biodiversity and implications for ecosystem functioning: A meta-analysis', in D.L. Sparks (ed.). Academic Press (Advances in Agronomy), pp. 1–44. Available at: https://doi.org/https://doi.org/10.1016/bs.agron.2019.01.001.

Graham, E.B. *et al.* (2016) 'Microbes as Engines of Ecosystem Function: When Does Community Structure Enhance Predictions of Ecosystem Processes?', *Frontiers in Microbiology*, 7, 214. Available at: https://doi.org/10.3389/fmicb.2016.00214.

Graham, J.E. *et al.* (2011) 'Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment', *Nature Communications*, 2, p. 375. Available at: https://doi.org/10.1038/ncomms1373.

Grant, G.A. (2018) 'D-3-Phosphoglycerate Dehydrogenase', *Frontiers in Molecular Biosciences*, 5, 110. Available at: https://doi.org/10.3389/fmolb.2018.00110.

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

Griffiths, R.I. *et al.* (2016) 'Mapping and validating predictions of soil bacterial biodiversity using European and national scale datasets', *Applied Soil Ecology*, 97, pp. 61–68. Available at: https://doi.org/https://doi.org/10.1016/j.apsoil.2015.06.018.

Grime, J.P. (1973) 'Competitive Exclusion in Herbaceous Vegetation', *Nature*, 242, pp. 344–347. Available at: https://doi.org/10.1038/242344a0.

Grime, J.P. (1977) 'Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory', *The American Naturalist*, 111, pp. 1169–1194. Available at: https://doi.org/10.1086/283244.

Grondin, J.M. *et al.* (2017) 'Polysaccharide Utilization Loci: Fuelling microbial communities', *Journal of Bacteriology*. Available at: https://doi.org/10.1128/JB.00860-16.

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

Guo, T. *et al.* (2020) 'Analysis of microbial utilization of rice straw in paddy soil using a DNA-SIP approach', *Soil Science Society of America Journal*, 84, pp. 99–114. Available at: https://doi.org/10.1002/saj2.20019.

Haichar, F.E.Z. *et al.* (2007) 'Identification of cellulolytic bacteria in soil by stable isotope probing', *Environmental Microbiology*, 9, pp. 625–634. Available at: https://doi.org/10.1111/j.1462-2920.2006.01182.x.

Hall, G., Cheng, E.W. and Burroughs, W. (1955) 'B-Vitamins Stimulatory to Cellulose Digestion by Washed Suspensions of Rumen Microorganisms', *Proceedings of the Iowa Academy of Science*, 62, pp. 273–278.

Harmsen, H., Prieur, D. and Jeanthon, C. (1997) 'Distribution of microorganisms in deep-sea hydrothermal vent chimneys investigated by whole-cell hybridization and enrichment culture of thermophilic subpopulations', *Applied and Environmental Microbiology*, 63, pp. 2876–2883. Available at: https://doi.org/10.1128/aem.63.7.2876-2883.1997.

Hatzenpichler, R. *et al.* (2020) 'Next-generation physiology approaches to study microbiome function at single cell level', *Nature Reviews Microbiology*, 18, pp. 241–256. Available at: https://doi.org/10.1038/s41579-020-0323-1.

Hedlund, B.P. *et al.* (2015) 'High-Quality Draft Genome Sequence of *Kallotenue papyrolyticum* JKG1 T Reveals Broad Heterotrophic Capacity Focused on Carbohydrate and Amino Acid Metabolism', *Genome Announcements*, 3, pp. 14–15. Available at: https://doi.org/10.1128/genomeA.01410-15.Copyright.

Hemingway, J.D. *et al.* (2019) 'Mineral protection regulates long-term global preservation of natural organic carbon', *Nature*, 570, pp. 228–231. Available at: https://doi.org/10.1038/s41586-019-1280-6.

Hemsworth, G.R. *et al.* (2015) 'Lytic Polysaccharide Monooxygenases in Biomass Conversion', *Trends in Biotechnology*, pp. 747–761. Available at: https://doi.org/10.1016/j.tibtech.2015.09.006.

Henrissat, B. *et al.* (1985) 'Synergism of cellulases from *Trichoderma reesei* in the degradation of cellulose', *Nature Biotechnology*, 3, pp. 722–726. Available at: https://doi.org/10.1038/nbt0885-722.

Hida, A. *et al.* (2020) 'Characterization of methyl-accepting chemotaxis proteins (MCPs) for amino acids in plant-growth-promoting rhizobacterium Pseudomonas protegens CHA0 and enhancement of amino acid chemotaxis by MCP genes overexpression', *Bioscience, Biotechnology, and Biochemistry*, 84, pp. 1948–1957. Available at: https://doi.org/10.1080/09168451.2020.1780112.

Himmel, M.E. *et al.* (2007) 'Biomass recalcitrance: Engineering plants and enzymes for biofuels production', *Science*, 315, pp. 804–807. Available at: https://doi.org/10.1126/science.1137016.

Hirano, K. *et al.* (2016) 'Enzymatic diversity of the Clostridium thermocellum cellulosome is crucial for the degradation of crystalline cellulose and plant biomass', *Scientific Reports*, 6. Available at: https://doi.org/10.1038/srep35709.

Ho, A., Di Lonardo, D.P. and Bodelier, P.L.E. (2017) 'Revisiting life strategy concepts in environmental microbial ecology', *FEMS Microbiology Ecology*, 93. Available at: https://doi.org/10.1093/femsec/fix006.

Hofrichter, M. (2002) 'Review: Lignin conversion by manganese peroxidase (MnP)', *Enzyme and Microbial Technology*, 30, pp. 454–466. Available at: https://doi.org/10.1016/S0141-0229(01)00528-2.

Hori, C. *et al.* (2013) 'Genomewide analysis of polysaccharides degrading enzymes in 11 white- and brown-rot Polyporales provides insight into mechanisms of wood decay', *Mycologia*, 105, pp. 1412–1427. Available at: https://doi.org/10.3852/13-072.

Hottes, A.K. *et al.* (2004) 'Transcriptional Profiling of *Caulobacter crescentus* during Growth on Complex and Minimal Media', *Journal of Bacteriology*, 186, pp. 1448–1461. Available at: https://doi.org/10.1128/JB.186.5.1448-1461.2004.

Hou, P. Bin *et al.* (2006) 'Cellulolytic complex exists in cellulolytic myxobacterium *Sorangium*', *Enzyme and Microbial Technology*, 38, pp. 273–278. Available at: https://doi.org/10.1016/j.enzmictec.2004.08.044.

Huang, K. *et al.* (2018) 'Improving economics of lignocellulosic biofuels: An integrated strategy for coproducing 1,5-pentanediol and ethanol', *Applied Energy*, 213, pp. 585–594. Available at: https://doi.org/10.1016/j.apenergy.2017.11.002.

Huang, Y. *et al.* (2021) 'Global Simulation and Evaluation of Soil Organic Matter and Microbial Carbon and Nitrogen Stocks Using the Microbial Decomposition Model ORCHIMIC v2.0', *Global Biogeochemical Cycles*, 35, e2020GB006836. Available at: https://doi.org/https://doi.org/10.1029/2020GB006836.

Huang, Y. *et al.* (2023) 'High-throughput microbial culturomics using automation and machine learning', *Nature Biotechnology*. Available at: https://doi.org/10.1038/s41587-023-01674-2.

Huertas-Rosales, Ó., Ramos-González, M.I. and Espinosa-Urgel, M. (2016) 'Self-Regulation and Interplay of Rsm Family Proteins Modulate the Lifestyle of *Pseudomonas putida*', *Applied and Environmental Microbiology*, 82, pp. 5673–5686. Available at: https://doi.org/10.1128/AEM.01724-16.

Hugenholtz, P. and Tyson, G.W. (2008) 'Microbiology: Metagenomics', *Nature*, pp. 481–483. Available at: https://doi.org/10.1038/455481a.

Hyatt, D. *et al.* (2010) 'Prodigal: Prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11, 9 . Available at: https://doi.org/10.1186/1471-2105-11-119.

IPCC, I.P. on C.C. (2007) *Climate Change 2007 - The Physical Science Basis*, *Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Available at: https://doi.org/10.1260/095830507781076194.

Ivanova, A.A., Naumoff, D.G., *et al.* (2017) 'Comparative genomics of four Isosphaeraceae planctomycetes: A common pool of plasmids and glycoside hydrolase genes shared by *Paludisphaera borealis* PX4T, *Isosphaera pallida* IS1BT, *Singulisphaera acidiphila* DSM 18658T, and strain SH-PL62', *Frontiers in Microbiology*, 8, 412. Available at: https://doi.org/10.3389/fmicb.2017.00412.

Ivanova, A.A., Wegner, C.E., *et al.* (2017) 'Metatranscriptomics reveals the hydrolytic potential of peat-inhabiting *Planctomycetes*', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, pp. 1–9. Available at: https://doi.org/10.1007/s10482-017-0973-9.

Jäger, G. *et al.* (2011) 'How recombinant swollenin from *Kluyveromyces lactis* affects cellulosic substrates and accelerates their hydrolysis', *Biotechnology for Biofuels*, 4. Available at: https://doi.org/10.1186/1754-6834-4-33.

Jannasch, H.W. and Jones, G.E. (1959) 'Bacterial Populations in Sea Water as Determined by Different Methods of Enumeration', *Limnology and Oceanography*, 4, pp. 128–139. Available at: https://doi.org/10.4319/lo.1959.4.2.0128.

Janssen, P.H. *et al.* (2002) 'Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions *Acidobacteria*, *Actinobacteria*, *Proteobacteria*, and *Verrucomicrobia*.', *Applied and Environmental Microbiology*, 68, pp. 2391–6. Available at: https://doi.org/10.1128/AEM.68.5.2391-2396.2002.

Jayaraman, S. *et al.* (2021) 'Editorial: Sustaining Soil Carbon to Enhance Soil Health, Food, Nutritional Security, and Ecosystem Services', *Frontiers in Sustainable Food Systems*, 5, 777495. Available at: https://doi.org/10.3389/fsufs.2021.777495.

Jehmlich, N. *et al.* (2010) 'Protein-based stable isotope probing', *Nature Protocols*, 5, pp. 1957–1966. Available at: https://doi.org/10.1038/nprot.2010.166.

Jiménez, D.J., Chaves-Moreno, D. and Van Elsas, J.D. (2015) 'Unveiling the metabolic potential of two soil-derived microbial consortia selected on wheat straw', *Scientific Reports*, 5. Available at: https://doi.org/10.1038/srep13845.

Jin, Z. *et al.* (2006) 'Covalent linkages between cellulose and lignin in cell walls of coniferous and nonconiferous woods', *Biopolymers*, 83, pp. 103–110. Available at: https://doi.org/10.1002/bip.20533.

Jobbágy, E.G. and Jackson, R.B. (2000) 'The vertical distribution of soil organic carbon and its relation to climate and vegetation', *Ecological Applications*, 10, pp. 423–436. Available at: https://doi.org/10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2.

Johansen, K.S. (2016) 'Discovery and industrial applications of lytic polysaccharide mono-oxygenases', *Biochemical Society Transactions*, 44, pp. 143–149. Available at: https://doi.org/10.1042/BST20150204.

Jones, S.M., van Dyk, J.S. and Pletschke, B.I. (2012) '*Bacillus Subtilis* SJ01 produces hemicellulose degrading multi-enzyme complexes', *BioResources*, 7, pp. 1294–1309. Available at: https://doi.org/10.15376/biores.7.1.1294-1309.

Jónsson, J.Ö.G. and Davíðsdóttir, B. (2016) 'Classification and valuation of soil ecosystem services', *Agricultural Systems*, 145, pp. 24–38. Available at: https://doi.org/https://doi.org/10.1016/j.agsy.2016.02.010.

Joshi, N. and Fass, J. (2011) 'sickle – A windowed adaptive trimming tool for FASTQ files using quality', *(Version 1.33)* [Software]. Available at https://github.com/najoshi/sickle.

Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596, pp. 583–589. Available at: https://doi.org/10.1038/s41586-021-03819-2.

Jung, D., Aoi, Y. and Epstein, S.S. (2016) 'In Situ Cultivation Allows for Recovery of Bacterial Types Competitive in Their Natural Environment', *Microbes and Environments*, 31, pp. 456–459. Available at: https://doi.org/10.1264/jsme2.ME16079.

Juturu, V. and Wu, J.C. (2014) 'Microbial cellulases: Engineering, production and applications', *Renewable and Sustainable Energy Reviews*, 33, pp. 188–203. Available at: https://doi.org/10.1016/j.rser.2014.01.077.

Kaeberlein, T., Lewis, K. and Epstein, S.S. (2002) 'Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment.', *Science*, 296, pp. 1127–9. Available at: https://doi.org/10.1126/science.1070633.

Kalam, S. *et al.* (2020) 'Recent Understanding of Soil *Acidobacteria* and Their Ecological Significance: A Critical Review', *Frontiers in Microbiology*, 11, 580024. Available at: https://doi.org/10.3389/fmicb.2020.580024.

Kanokratana, P. *et al.* (2011) 'Insights into the Phylogeny and Metabolic Potential of a Primary Tropical Peat Swamp Forest Microbial Community by Metagenomic Analysis', *Microbial Ecology*, 61, pp. 518–528. Available at: https://doi.org/10.1007/s00248-010-9766-7.

Kapich, A.N. *et al.* (2005) 'Involvement of lipid peroxidation in the degradation of a non-phenolic lignin model compound by manganese peroxidase of the litter-decomposing fungus *Stropharia coronilla*', *Biochemical and Biophysical Research Communications*, 330, pp. 371–377. Available at: https://doi.org/10.1016/j.bbrc.2005.02.167.

Karimi, B. *et al.* (2018) 'Biogeography of soil bacteria and archaea across France', *Science Advances*, 4, eaat1808. Available at: https://doi.org/10.1126/sciadv.aat1808.

Kato, S. *et al.* (2018) 'Isolation of previously uncultured slowgrowing bacteria by using a simple modification in the preparation of agar media', *Applied and Environmental Microbiology*, 84. Available at: https://doi.org/10.1128/AEM.00807-18.

Kelly, R.H. *et al.* (1997) 'Simulating trends in soil organic carbon in long-term experiments using the century model', *Geoderma*, 81, pp. 75–90. Available at: https://doi.org/https://doi.org/10.1016/S0016-7061(97)00082-7.

Kim, M. Il and Hong, M. (2016) 'Crystal structure and catalytic mechanism of pyridoxal kinase from *Pseudomonas aeruginosa*', *Biochemical and Biophysical Research Communications*, 478, pp. 300–306. Available at: http://doi.org/https://doi.org/10.1016/j.bbrc.2016.07.007.

Kim, S. *et al.* (2020) 'High-throughput cultivation based on dilution-to-extinction with catalase supplementation and a case study of cultivating acI bacteria from Lake Soyang', *Journal of Microbiology*, 58, pp. 893–905. Available at: https://doi.org/10.1007/s12275-020-0452-2.

Kim, S.-K. *et al.* (2015) 'Enhanced tolerance of *Saccharomyces cerevisiae* to multiple lignocellulose-derived inhibitors through modulation of spermidine contents', *Metabolic Engineering*, 29, pp. 46–55. Available at: https://doi.org/https://doi.org/10.1016/j.ymben.2015.02.004.

Kim, S.K. *et al.* (2018) 'Expression of a cellobiose phosphorylase from *Thermotoga maritima* in *Caldicellulosiruptor bescii* improves the phosphorolytic pathway and results in a dramatic increase in cellulolytic activity', *Applied and Environmental Microbiology*, 84. Available at: https://doi.org/10.1128/AEM.02348-17.

Ko, K.C. *et al.* (2013) 'A novel multifunctional cellulolytic enzyme screened from metagenomic resources representing ruminal bacteria', *Biochemical and Biophysical Research Communications*, 441, pp. 567–572. Available at: https://doi.org/10.1016/j.bbrc.2013.10.120.

Kobras, C.M., Fenton, A.K. and Sheppard, S.K. (2021) 'Next-generation microbiology: from comparative genomics to gene function', *Genome Biology*, 22, p. 123. Available at: https://doi.org/10.1186/s13059-021-02344-9.

Koeck, D.E. *et al.* (2014) 'Genomics of cellulolytic bacteria', *Current Opinion in Biotechnology*, pp. 171–183. Available at: https://doi.org/10.1016/j.copbio.2014.07.002.

Kolby Smith, W. *et al.* (2015) 'Large divergence of satellite and Earth system model estimates of global terrestrial CO2 fertilization', *Nature Climate Change*, 6, pp. 306–310. Available at: https://doi.org/10.1038/nclimate2879.

Könneke, M. *et al.* (2014) 'Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation', *Proceedings of the National Academy of Sciences of the United States of America*, 111, pp. 8239–8234. Available at: https://doi.org/10.1073/pnas.1402028111.

Kopittke, P.M. *et al.* (2017) 'Global changes in soil stocks of carbon, nitrogen, phosphorus, and sulphur as influenced by long-term agricultural production', *Global Change Biology*, 23, pp. 2509–2519. Available at: https://doi.org/https://doi.org/10.1111/gcb.13513.

Korzhenkov, A.A. *et al.* (2019) 'Archaea dominate the microbial community in an ecosystem with low-to-moderate temperature and extreme acidity', *Microbiome*, 7, p. 11. Available at: https://doi.org/10.1186/s40168-019-0623-8.

Kracher, D. *et al.* (2016) 'Extracellular electron transfer systems fuel cellulose oxidative degradation.', *Science*, 3165, p. aaf3165. Available at: https://doi.org/10.1126/science.aaf3165.

Kuhad, R.C., Gupta, R. and Singh, A. (2011) 'Microbial Cellulases and Their Industrial Applications', *Enzyme Research*, 2011, pp. 1–10. Available at: https://doi.org/10.4061/2011/280696.

Kulichevskaya, I.S. *et al.* (2012) '*Telmatocola sphagniphila gen. nov*., sp. nov., a novel dendriform planctomycete from northern wetlands', *Frontiers in Microbiology*, 3, 146. Available at: https://doi.org/10.3389/fmicb.2012.00146.

Kurm V., van der Putten W.H., Hol W.H.G. (2019) 'Cultivation-success of rare soil bacteria is not influenced by incubation time and growth medium', *PLoS ONE* 14, e0210073. Available at: https://doi.org/10.1371/journal.pone.0210073Kyker-Snowman, E. *et al.* (2020) 'Stoichiometrically coupled carbon and nitrogen cycling in the MIcrobial-MIneral Carbon Stabilization model version 1.0 (MIMICS-CN v1.0)', *Geoscientific Model Development*, 13, pp. 4413–4434. Available at: https://doi.org/10.5194/gmd-13-4413-2020.

Laetsch, D.R. and Blaxter, M.L. (2017) 'BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]', *F1000Research*, 6. Available at: https://doi.org/10.12688/f1000research.12232.1.

Lal, R. (2008) 'Carbon sequestration.', *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 363, pp. 815–30. Available at: https://doi.org/10.1098/rstb.2007.2185.

Lal, R. (2009) 'Soil degradation as a reason for inadequate human nutrition', *Food Security*, 1, pp. 45–57. Available at: https://doi.org/10.1007/s12571-009-0009-z.

Lambertz, C. *et al.* (2016) 'Progress and obstacles in the production and application of recombinant lignin-degrading peroxidases', *Bioengineered*, 7, pp. 145–154. Available at: https://doi.org/10.1080/21655979.2016.1191705.

Langmead, B. and Salzberg, S. (2013) 'Bowtie2', *Nature methods* 9, pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Larsbrink, J. *et al.* (2016) 'A polysaccharide utilization locus from *Flavobacterium johnsoniae* enables conversion of recalcitrant chitin', *Biotechnology for Biofuels*, 9. Available at: https://doi.org/10.1186/s13068-016-0674-z.

Lato, D.F. and Golding, G.B. (2020) 'Spatial Patterns of Gene Expression in Bacterial Genomes', *Journal of Molecular Evolution*, 88, pp. 510–520. Available at: https://doi.org/10.1007/s00239-020-09951-3.

Lee, C.G. *et al.* (2011) 'Bacterial populations assimilating carbon from 13C-labeled plant residue in soil: Analysis by a DNA-SIP approach', *Soil Biology and Biochemistry*, 43, pp. 814–822. Available at: https://doi.org/10.1016/j.soilbio.2010.12.016.

Lo Leggio, L. *et al.* (2015) 'Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase', *Nature Communications*, 6. Available at: https://doi.org/10.1038/ncomms6961.

Lehmann, J. *et al.* (2020) 'Persistence of soil organic carbon caused by functional complexity', *Nature Geoscience*, 13, pp. 529–534. Available at: https://doi.org/10.1038/s41561-020-0612-3.

Leininger, S. *et al.* (2006) 'Archaea predominate among ammonia-oxidizing prokaryotes in soils', *Nature*, 442, pp. 806–809. Available at: https://doi.org/10.1038/nature04983.

Leis, B. *et al.* (2015) 'Functional Screening of Hydrolytic Activities Reveals an Extremely Thermostable Cellulase from a Deep-Sea Archaeon', *Front Bioeng Biotechnol*, 3, 00095. Available at: https://doi.org/10.3389/fbioe.2015.00095.

Leschine, S.B. (1995) 'Cellulose degradation in anaerobic environments', *Annual Review of Microbiology*, 49, pp. 399–426.

Leung, H.T. *et al.* (2016) 'Long-term effects of timber harvesting on hemicellulolytic microbial populations in coniferous forest soils', *The ISME Journal*, 10118, pp. 363–375. Available at: https://doi.org/10.1038/ismej.2015.118.

Lewin, A. *et al.* (2017) 'Novel archaeal thermostable cellulases from an oil reservoir metagenome', *AMB Express*, 7. Available at: https://doi.org/10.1186/s13568-017-0485-z.

Li, D. *et al.* (2016) 'MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices', *Methods*, 102, pp. 3–11. Available at: https://doi.org/10.1016/j.ymeth.2016.02.020.

Li, F., Zhao, H., *et al.* (2021) 'Enhanced Fenton Reaction for Xenobiotic Compounds and Lignin Degradation Fueled by Quinone Redox Cycling by Lytic Polysaccharide Monooxygenases', *Journal of Agricultural and Food Chemistry*, 69, pp. 7104–7114. Available at: https://doi.org/10.1021/acs.jafc.1c01684.

Li, F., Zhang, J., *et al.* (2021) 'Lytic polysaccharide monooxygenases promote oxidative cleavage of lignin and lignin–carbohydrate complexes during fungal degradation of lignocellulose', *Environmental Microbiology*, 23, pp. 4547–4560. Available at: https://doi.org/https://doi.org/10.1111/1462-2920.15648.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25, pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Li, S. *et al.* (2012) 'Technology Prospecting on Enzymes: Application, Marketing and Engineering', *Computational and Structural Biotechnology Journal*, 2, e201209017. Available at: https://doi.org/10.5936/csbj.201209017.

Liao, Y., Smyth, G.K. and Shi, W. (2014) 'FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30, pp. 923–930. Available at: https://doi.org/10.1093/bioinformatics/btt656.

Liebl, W. (2001) 'Cellulolytic enzymes from *Thermotoga* species', *Methods in Enzymology*, 330, pp. 290–300. Available at: https://doi.org/10.1016/S0076-6879(01)30383-X.

Lienhard, P. *et al.* (2013) 'Soil microbial diversity and C turnover modified by tillage and cropping in Laos tropical grassland', *Environmental Chemistry Letters*, 11, pp. 391–398. Available at: https://doi.org/10.1007/s10311-013-0420-8.

de Lima Brossi, M.J. *et al.* (2016) 'Soil-Derived Microbial Consortia Enriched with Different Plant Biomass Reveal Distinct Players Acting in Lignocellulose Degradation', *Microbial Ecology*, 71, pp. 616–627. Available at: https://doi.org/10.1007/s00248-015-0683-7.

Ling, C. *et al.* (2022) 'Muconic acid production from glucose and xylose in *Pseudomonas putida* via evolution and metabolic engineering', *Nature Communications*, 13, p. 4925. Available at: https://doi.org/10.1038/s41467-022-32296-y.

Ling, L.L. *et al.* (2015a) 'A new antibiotic kills pathogens without detectable resistance', *Nature*, 517, pp. 455–459. Available at: https://doi.org/10.1038/nature14098.

Ling, L.L. *et al.* (2015b) 'Erratum: A new antibiotic kills pathogens without detectable resistance', *Nature*, 520, pp. 388–388. Available at: https://doi.org/10.1038/nature14303.

Lloyd, K.G. *et al.* (2018) 'Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes', *mSystems*, 3, e00055-18. Available at: https://doi.org/10.1128/mSystems.00055-18.

Locey, K.J. and Lennon, J.T. (2016) 'Scaling laws predict global microbial diversity', *Proceedings of the National Academy of Sciences*, 113, pp. 5970–5975. Available at: https://doi.org/10.1073/pnas.1521291113.

Lombard, V. *et al.* (2014) 'The carbohydrate-active enzymes database (CAZy) in 2013', *Nucleic Acids Research*, 42, pp. 490–495. Available at: https://doi.org/10.1093/nar/gkt1178.

Lopes, L.D. *et al.* (2015) 'Exploring the sheep rumen microbiome for carbohydrate-active enzymes', *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 108, pp. 15–30. Available at: https://doi.org/10.1007/s10482-015-0459-6.

López-Mondéjar, R. *et al.* (2022) 'Global Distribution of Carbohydrate Utilization Potential in the Prokaryotic Tree of Life', *mSystems*, 7, pp. e00829-22. Available at: https://doi.org/10.1128/msystems.00829-22

López-Mondéjar, R. *et al.* (2020) 'Metagenomics and stable isotope probing reveal the complementary contribution of fungal and bacterial communities in the recycling of dead biomass in forest soil', *Soil Biology and Biochemistry*, 148, 107875. Available at: https://doi.org/10.1016/j.soilbio.2020.107875.

López-Mondéjar, R., Algora, C. and Baldrian, P. (2019) 'Lignocellulolytic systems of soil bacteria: A vast and diverse toolbox for biotechnological conversion processes', *Biotechnology Advances*, 37, 107374. Available at: https://doi.org/10.1016/j.biotechadv.2019.03.013.

Lowe, N.M. (2021) 'The global challenge of hidden hunger: perspectives from the field.', *The Proceedings of the Nutrition Society*, 80, pp. 283–289. Available at: https://doi.org/10.1017/S0029665121000902.

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. Available at: https://doi.org/10.1093/femsec/fiw246.

Ludwig, R. *et al.* (2010) 'Cellobiose dehydrogenase: A versatile catalyst for electrochemical applications', *ChemPhysChem*, pp. 2674–2697. Available at: https://doi.org/10.1002/cphc.201000216.

Lynd, L.R. *et al.* (2002) 'Microbial Cellulose Utilisation: Fundamentals and Biotechnology', *Microbiology and Molecular Biology Reviews*, 66, pp. 506–577. Available at: https://doi.org/10.1128/MMBR.66.3.506.

Lynd, L.R. *et al.* (2017a) 'Cellulosic ethanol: status and innovation', *Current Opinion in Biotechnology*, pp. 202–211. Available at: https://doi.org/10.1016/j.copbio.2017.03.008.

Lynd, L.R. *et al.* (2017b) 'Cellulosic ethanol: status and innovation', *Current Opinion in Biotechnology*, pp. 202–211. Available at: https://doi.org/10.1016/j.copbio.2017.03.008.

Mackie, G.A. (2013) 'RNase E: at the interface of bacterial RNA processing and decay', *Nature Reviews Microbiology*, 11, pp. 45–57. Available at: https://doi.org/10.1038/nrmicro2930.

Malgas, S. and Pletschke, B.I. (2019) 'The effect of an oligosaccharide reducing-end xylanase, BhRex8A, on the synergistic degradation of xylan backbones by an optimised xylanolytic enzyme cocktail.', *Enzyme and microbial technology*, 122, pp. 74–81. Available at: https://doi.org/10.1016/j.enzmictec.2018.12.010.

Malik, A.A. *et al.* (2018) 'Land use driven change in soil pH affects microbial carbon cycling processes', *Nature Communications*, 9, pp. 1–10. Available at: https://doi.org/10.1038/s41467-018-05980-1.

Malik, A.A. *et al.* (2020) 'Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change', *The ISME Journal*, 14, pp. 1–9. Available at: https://doi.org/10.1038/s41396-019-0510-0.

Manoharan, L. *et al.* (2017) 'Agricultural land use determines functional genetic diversity of soil microbial communities', *Soil Biology and Biochemistry*, 115, pp. 423–432. Available at: https://doi.org/https://doi.org/10.1016/j.soilbio.2017.09.011.

Mardanov, A. V. *et al.* (2012) 'Complete genome sequence of the hyperthermophilic cellulolytic crenarchaeon "*Thermogladius cellulolyticus*" 1633', *Journal of Bacteriology*, pp. 4446–4447. Available at: https://doi.org/10.1128/JB.00894-12.

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

Martînez, A.T. (2002) 'Molecular biology and structure-function of lignin-degrading heme peroxidases', *Enzyme and Microbial Technology*, 30, pp. 425–444. Available at: https://doi.org/10.1016/S0141-0229(01)00521-X.

Martínez, Á.T. *et al.* (2005) 'Biodegradation of lignocellulosics: Microbial, chemical, and enzymatic aspects of the fungal attack of lignin', in *International Microbiology*, pp. 195–204. Available at: https://doi.org/im2305029.

Martinez, C. *et al.* (2016) 'Belowground carbon allocation patterns as determined by the in-growth soil core ^{13}C technique across different ecosystem types', *Geoderma*, 263, pp. 140–150. Available at: https://doi.org/https://doi.org/10.1016/j.geoderma.2015.08.043.

Martinez, D. *et al.* (2008) 'Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*)', *Nature Biotechnology*, 26, pp. 553–560. Available at: https://doi.org/10.1038/nbt1403.

Martinez, N.D. (2023) 'Predicting ecosystem metaphenome from community metagenome: A grand challenge for environmental biology', *Ecology and Evolution*, 13, p. e9872. Available at: https://doi.org/https://doi.org/10.1002/ece3.9872.

Martins, L.F. *et al.* (2013) 'Metagenomic Analysis of a Tropical Composting Operation at the São Paulo Zoo Park Reveals Diversity of Biomass Degradation Functions and Organisms', *PLoS ONE*, 8, 61928. Available at: https://doi.org/10.1371/journal.pone.0061928.

McDonald, J.E., Allison, H.E. and McCarthy, A.J. (2010) 'Composition of the Landfill Microbial Community as Determined by Application of Domain- and Group-Specific 16S and 18S rRNA-Targeted Oligonucleotide Probes', *Applied and Environmental Microbiology*, 76, pp. 1301–1306. Available at: https://doi.org/10.1128/AEM.01783-09.

Medie, F.M. *et al.* (2012) 'Genome analyses highlight the different biological roles of cellulases', *Nature Reviews Microbiology*, 10, pp. 227–234. Available at: https://doi.org/10.1038/nrmicro2729.

Menzel, P., Ng, K.L. and Krogh, A. (2016) 'Fast and sensitive taxonomic classification for metagenomics with Kaiju', *Nature Communications*, 7, 11257. Available at: https://doi.org/10.1038/ncomms11257.

Mikheenko, A., Saveliev, V. and Gurevich, A. (2016) 'MetaQUAST: Evaluation of metagenome assemblies', *Bioinformatics*, 7, pp. 1088–1090. Available at: https://doi.org/10.1093/bioinformatics/btv697.

Min, K. *et al.* (2015) 'A dye-decolorizing peroxidase from *Bacillus subtilis* exhibiting substrate-dependent optimum temperature for dyes and b-ether lignin dimer', *Scientific Reports*, 5. Available at: https://doi.org/10.1038/srep08245.

Mock, T. *et al.* (2016) 'Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes', *Global Change Biology*, pp. 61–75. Available at: https://doi.org/10.1111/gcb.12983.

Montenecourt, B.S. and Eveleigh, D.E. (1977) 'Preparation of mutants of *Trichoderma reesei* with enhanced cellulase production', *Applied and Environmental Microbiology*, 34, pp. 777–782. Available at: https://doi.org/10.1016/j.recesp.2011.09.022.

Morgenstern, I., Powlowski, J. and Tsang, A. (2014) 'Fungal cellulose degradation by oxidative enzymes: from dysfunctional GH61 family to powerful lytic polysaccharide monooxygenase family', *Briefings in Functional Genomics*, 13, pp. 471–481. Available at: https://doi.org/10.1093/bfgp/elu032.

Mueller, S.A., Anderson, J.E. and Wallington, T.J. (2011) 'Impact of biofuel production and other supply and demand factors on food price increases in 2008', *Biomass and Bioenergy*, 35, pp. 1623–1632. Available at: https://doi.org/10.1016/j.biombioe.2011.01.030.

Müller, G. *et al.* (2015) 'Harnessing the potential of LPMO-containing cellulase cocktails poses new demands on processing conditions', *Biotechnology for Biofuels*, 8. Available at: https://doi.org/10.1186/s13068-015-0376-y.

Naas, A.E. *et al.* (2014) 'Do rumen *Bacteroidota* utilise an alternative mechanism for cellulose degradation?', *mBio*, 5. Available at: https://doi.org/10.1128/mBio.01401-14.

Naik, S.N. *et al.* (2010) 'Production of first and second generation biofuels: A comprehensive review', *Renewable and Sustainable Energy Reviews*, pp. 578–597. Available at: https://doi.org/10.1016/j.rser.2009.10.003.

O'Leary, N.A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation'. *Nucleic Acids Research*, 44, pp. 733–45. Available at: https://doi.org/10.1093/nar/gkv1189.

Nelkner, J. *et al.* (2019) 'Effect of Long-Term Farming Practices on Agricultural Soil Microbiome Members Represented by Metagenomically Assembled Genomes (MAGs) and Their Predicted Plant-Beneficial Genes.', *Genes*, 10. Available at: https://doi.org/10.3390/genes10060424.Neufeld, J.D., Vohra, J., *et al.* (2007) 'DNA stable-isotope probing', *Nature Protocols*, 2, pp. 860–866. Available at: https://doi.org/10.1038/nprot.2007.109.

Neufeld, J.D., Dumont, M.G., *et al.* (2007) 'Methodological considerations for the use of stable isotope probing in microbial ecology', in *Microbial Ecology*, pp. 435–442. Available at: https://doi.org/10.1007/s00248-006-9125-x.

Neufeld, J.D., Chen, Y., Dumont, M.D. and Murrell, C.M. (2008) 'Marine methylotrophs revealed by stable-isotopeprobing, multiple displacement amplificationand metagenomics', *Environmental Microbiolgy*, 10, pp. 1526–1535. Available at: https://doi.org/10.1111/j.1462-2920.2008.01568.x.

Neumann, A.P., McCormick, C.A. and Suen, G. (2017) '*Fibrobacter* communities in the gastrointestinal tracts of diverse hindgut-fermenting herbivores are distinct from those of the rumen.', *Environmental microbiology*, 19, pp. 3768–3783. Available at: https://doi.org/10.1111/1462-2920.13878.

Nguyen, S.T.C. *et al.* (2018) 'Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy', *Applied Microbiology and Biotechnology*, 102, pp. 1629–1637. Available at: https://doi.org/10.1007/s00253-018-8778-y.

Nichols, D. *et al.* (2008) 'Short peptide induces an "uncultivable" microorganism to grow in vitro', *Applied and Environmental Microbiology*, 74, pp. 4889–4897. Available at: https://doi.org/10.1128/AEM.00393-08.

Nichols, D. *et al.* (2010) 'Use of iChip for high-throughput in situ cultivation of "uncultivable microbial species', *Applied and Environmental Microbiology*, 76, pp. 2445–2450. Available at: https://doi.org/10.1128/AEM.01754-09.

Nielsen, D.C. *et al.* (2011) 'Fallow effects on soil' in *Soil Management: Building a stable base for agriculture*, Lincoln: USDA-ARS, pp. 287–300. Available at: https://doi.org/10.2136/2011.soilmanagement.c19.

Nimchua, T. *et al.* (2012) 'Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes', *Journal of Microbiology and Biotechnology*, 22, pp. 462–469. Available at: https://doi.org/10.4014/jmb.1108.08037.

Nonaka, H. *et al.* (2006) 'Complete genome sequence of the dehalorespiring bacterium *Desulfitobacterium hafniense* Y51 and comparison with *Dehalococcoides ethenogenes* 195.', *Journal of Bacteriology*, 188, pp. 2262–2274. Available at: https://doi.org/10.1128/JB.188.6.2262-2274.2006.

Nuccio, E.E. *et al.* (2020) 'Niche differentiation is spatially and temporally regulated in the rhizosphere', *ISME Journal*, 14, pp. 999–1014. Available at: https://doi.org/10.1038/s41396-019-0582-x.

Oates, N.C. *et al.* (2021) 'A multi-omics approach to lignocellulolytic enzyme discovery reveals a new ligninase activity from *Parascedosporium putredinis* NO1', *Proceedings of the National Academy of Sciences*, 118, e2008888118. Available at: https://doi.org/10.1073/pnas.2008888118.

Oksanen, J. *et al.* (2008) 'The vegan package', [Software]*,* R package version 2.0.

Olofsson, J. *et al.* (2017) 'Integrating enzyme fermentation in lignocellulosic ethanol production: Life-cycle assessment and techno-economic analysis', *Biotechnology for Biofuels*, 10. Available at: https://doi.org/10.1186/s13068-017-0733-0.

Olomu, I.N. *et al.* (2020) 'Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome.', *BMC Microbiology*, 20, p. 157. Available at: https://doi.org/10.1186/s12866-020-01839-y.

Ostle, N.J. *et al.* (2009) 'UK land use and soil carbon sequestration', *Land Use Policy*, 26, pp. 274–283. Available at: https://doi.org/https://doi.org/10.1016/j.landusepol.2009.08.006.

Ounit, R. *et al.* (2015) 'CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers', *BMC Genomics*, 16, 236. Available at: https://doi.org/10.1186/s12864-015-1419-2.

Pan, X. *et al.* (2005) 'Strategies to Enhance the Enzymatic Hydrolysis of Pretreated Softwood with High Residual Lignin Content', *Applied Biochemistry and Biotechnology*, 124, pp. 1069–1079.

Pandey, K.K. and Pitman, A.J. (2003) 'FTIR studies of the changes in wood chemistry following decay by brown-rot and white-rot fungi', *International Biodeterioration and Biodegradation*, 52, pp. 151–160. Available at: https://doi.org/10.1016/S0964-8305(03)00052-0.

Pandit, P.D. *et al.* (2016) 'Mining of hemicellulose and lignin degrading genes from differentially enriched methane producing microbial community', *Bioresource Technology*, 216, pp. 923–930. Available at: https://doi.org/10.1016/j.biortech.2016.06.021.

Park, D.H. *et al.* (2010) 'Mutations in γ-aminobutyric acid (GABA) transaminase genes in plants or *Pseudomonas syringae* reduce bacterial virulence.', *The Plant Journal: For Cell and Molecular Biology*, 64, pp. 318–330. Available at: https://doi.org/10.1111/j.1365-313X.2010.04327.x.

Parks, D.H. *et al.* (2021) 'GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy', *Nucleic Acids Research*, 50, pp. D785–D794. Available at: https://doi.org/10.1093/nar/gkab776.

Parton, W.J. *et al.* (1998) 'DAYCENT and its land surface submodel: description and testing', *Global and Planetary Change*, 19, pp. 35–48. Available at: https://doi.org/https://doi.org/10.1016/S0921-8181(98)00040-X.

Paterson, E. *et al.* (2011) 'Long-term exclusion of plant-inputs to soil reduces the functional capacity of microbial communities to mineralise recalcitrant root-derived carbon sources', *Soil Biology and Biochemistry*, 43, pp. 1873–1880.

Pelmont, J. *et al.* (1989) 'A new bacterial alcohol dehydrogenase active on degraded lignin and several low molecular weight aromatic compounds', *FEMS Microbiology Letters*, 57, pp. 109–113. Available at: https://doi.org/10.1111/j.1574-6968.1989.tb03230.x.

Pepe-Ranney, C. *et al.* (2016) 'Unearthing the ecology of soil microorganisms using a high resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil', *Frontiers in Microbiology*, 7, 703. Available at: https://doi.org/10.3389/fmicb.2016.00703.

Perchat, N. *et al.* (2018) 'Elucidation of the trigonelline degradation pathway reveals previously undescribed enzymes and metabolites.', *Proceedings of the National Academy of Sciences of the United States of America*, 115, pp. E4358–E4367. Available at: https://doi.org/10.1073/pnas.1722368115.

Pham, V.H.T. and Kim, J. (2012) 'Cultivation of unculturable soil bacteria', *Trends in Biotechnology*, 30, pp. 475–484. Available at: https://doi.org/10.1016/j.tibtech.2012.05.007.

Pinnell, L.J. *et al.* (2014) 'Recovering glycoside hydrolase genes from active tundra cellulolytic bacteria.', *Canadian Journal of Microbiology*, 60, pp. 469–76. Available at: https://doi.org/10.1139/cjm-2014-0193.

Piro, V.C., Matschkowski, M. and Renard, B.Y. (2017) 'MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling', *Microbiome*, 5, 101. Available at: https://doi.org/10.1186/s40168-017-0318-y.

Pold, G. *et al.* (2016) 'Long-term warming alters carbohydrate degradation potential in temperate forest soils', *Applied and Environmental Microbiology*, 82. Available at: https://doi.org/10.1128/AEM.02012-16.

Pollo, S.M.J., Zhaxybayeva, O. and Nesbø, C.L. (2015) 'Insights into thermoadaptation and the evolution of mesophily from the bacterial phylum *Thermotogota*', *Canadian Journal of Microbiology*, 61, pp. 655–670. Available at: https://doi.org/10.1139/cjm-2015-0073.

Premalatha, N. *et al.* (2015) 'Optimization of cellulase production by *Enhydrobacter* sp. ACCA2 and its application in biomass saccharification', *Frontiers in Microbiology*, 6, 1046. Available at: https://doi.org/10.3389/fmicb.2015.01046.

Pu, Y. *et al.* (2013) 'Assessing the molecular structure basis for biomass recalcitrance during dilute acid and hydrothermal pretreatments', *Biotechnology for Biofuels*, 6, p. 15. Available at: https://doi.org/10.1186/1754-6834-6-15.

Qin, X. *et al.* (2018) 'Deciphering lignocellulose deconstruction by the white rot fungus *Irpex lacteus* based on genomic and transcriptomic analyses', *Biotechnology for Biofuels*, 11, p. 58. Available at: https://doi.org/10.1186/s13068-018-1060-9.

Qiu, L. *et al.* (2021) 'Erosion reduces soil microbial diversity, network complexity and multifunctionality', *The ISME Journal*, 15, pp. 2474–2489. Available at: https://doi.org/10.1038/s41396-021-00913-1.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26, pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

R Core Team (2017) 'R: A language and environment for statistical computing.' R Foundation for Statistical Computing, Vienna, Austria.

Ramanjaneyulu, G. and Reddy, B.R. (2019) 'Chapter 21 – Emerging Trends of Microorganism in the Production of Alternative Energy', in V. Buddolla (ed.) *Recent Developments in Applied Microbiology and Biochemistry*. Academic Press, pp. 275–305. Available at: https://doi.org/https://doi.org/10.1016/B978-0-12-816328-3.00021-0.

Ransom-Jones, E. *et al.* (2017) 'Lignocellulose-degrading microbial communities in landfill sites represent a repository of unexplored biomass-degrading diversity', *Applied and Environmental Science*, 2, pp. 1–13.

Ravin, N. V. *et al.* (2018) 'Genome analysis of *Fimbriiglobus ruber* SP5T, a planctomycete with confirmed chitinolytic capability', *Applied and Environmental Microbiology*, 84. Available at: https://doi.org/10.1128/AEM.02645-17.

Rawlence, N.J. *et al.* (2014) 'Using palaeoenvironmental DNA to reconstruct past environments: progress and prospects', *Journal of Quaternary Science*, 29, pp. 610–626. Available at: https://doi.org/https://doi.org/10.1002/jqs.2740.

Remenár, M. *et al.* (2015) 'Isolation of previously uncultivable bacteria from a nickel contaminated soil using a diffusion-chamber-based approach', *Applied Soil Ecology*, 95, pp. 115–127. Available at: https://doi.org/10.1016/j.apsoil.2015.06.013.

Ren, C. *et al.* (2021) 'Altered microbial CAZyme families indicated dead biomass decomposition following afforestation', *Soil Biology and Biochemistry*, 160, p. 108362. Available at: https://doi.org/https://doi.org/10.1016/j.soilbio.2021.108362.

Van Rijssel, S.Q. *et al.* (2022) 'Soil microbial diversity and community composition during conversion from conventional to organic agriculture', *Molecular Ecology*, 31, pp. 4017–4030. Available at: https://doi.org/https://doi.org/10.1111/mec.16571.

Riley, R. *et al.* (2014) 'Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi', *Proceedings of the National Academy of Sciences*, 111, pp. 9923–9928. Available at: https://doi.org/10.1073/pnas.1400592111.

Riva, S. (2006) 'Laccases: blue enzymes for green chemistry', *Trends in Biotechnology*, pp. 219–226. Available at: https://doi.org/10.1016/j.tibtech.2006.03.006.

Romdhane, S. *et al.* (2022) 'Land-use intensification differentially affects bacterial, fungal and protist communities and decreases microbiome network complexity', *Environmental Microbiome*, 17, p. 1. Available at: https://doi.org/10.1186/s40793-021-00396-9.

Romeo, T., Vakulskas, C.A. and Babitzke, P. (2013) 'Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems', *Environmental Microbiology*, 15, pp. 313–324. Available at: https://doi.org/https://doi.org/10.1111/j.1462-2920.2012.02794.x.

Rosario, D. *et al.* (2018) 'Understanding the Representative Gut Microbiota Dysbiosis in Metformin-Treated Type 2 Diabetes Patients Using Genome-Scale Metabolic Modeling', *Frontiers in Physiology*, 9, 775. Available at: https://doi.org/10.3389/fphys.2018.00775.

Rosenberg, E. *et al.* (2013) *The prokaryotes: Prokaryotic physiology and biochemistry*, *The Prokaryotes: Prokaryotic Physiology and Biochemistry*. Available at: https://doi.org/10.1007/978-3-642-30141-4.

Rosewarne, C.P. *et al.* (2014) 'Analysis of the bovine rumen microbiome reveals a diversity of Sus-like polysaccharide utilisation loci from the bacterial phylum *Bacteroidota*', *Journal of Industrial Microbiology and Biotechnology*, 41, pp. 601–606. Available at: https://doi.org/10.1007/s10295-013-1395-y.

Rossi, G, Riccardo S., Geuse, G., Staganelli, G., and Wang, T.K. (1936) 'Direct microscopic and bacteriological investigations of the soil.', *Soil Science*, 41, pp. 53–66.

Rygaard, A. Mac *et al.* (2017) 'Effects of Gelling Agent and Extracellular Signalling Molecules on the Culturability of Marine Bacteria.', *Applied and Environmental Microbiology*, 83(9). Available at: https://doi.org/10.1128/AEM.00243-17.

Saghaï, A. *et al.* (2022) 'Diversity of archaea and niche preferences among putative ammonia-oxidizing *Nitrososphaeria* dominating across European arable soils', *Environmental Microbiology*, 24, pp. 341–356. Available at: https://doi.org/https://doi.org/10.1111/1462-2920.15830.

Sah, S. and Singh, R. (2016) 'Phylogenetical coherence of Pseudomonas in unexplored soils of Himalayan region', *3 Biotech*, 6, 170. Available at: https://doi.org/10.1007/s13205-016-0493-8.

Sait, M., Hugenholtz, P. and Janssen, P.H. (2002) 'Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys', *Environmental Microbiology*, 4, pp. 654–666. Available at: https://doi.org/10.1046/j.1462-2920.2002.00352.x.

San, J.E. *et al.* (2020) 'Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls', *Frontiers in Microbiology*, 10, 3119. Available at: https://doi.org/10.3389/fmicb.2019.03119.

Sangwan, N., Xia, F. and Gilbert, J.A. (2016) 'Recovering complete and draft population genomes from metagenome datasets', *Microbiome*, 4, p. 8. Available at: https://doi.org/10.1186/s40168-016-0154-5.

Scharlemann, J. *et al.* (2014) 'Global soil carbon: understanding and managing the largest terrestrial carbon pool', *Carbon Management*, 5, pp. 81–91. Available at: https://doi.org/10.4155/cmt.13.77.

Schellenberger, S., Kolb, S. and Drake, H.L. (2010) 'Metabolic responses of novel cellulolytic and saccharolytic agricultural soil Bacteria to oxygen', *Environmental Microbiology*, 12, pp. 845–861. Available at: https://doi.org/10.1111/j.1462-2920.2009.02128.x.

Schleper, C., Jurgens, G. and Jonuscheit, M. (2005) 'Genomic studies of uncultivated archaea', *Nature Reviews Microbiology*, 3, pp. 479–488. Available at: https://doi.org/10.1038/nrmicro1159.

Schmieder, R. and Edwards, R. (2011) 'Quality control and preprocessing of metagenomic datasets', *Bioinformatics*, 27, pp. 863–864. Available at: https://doi.org/10.1093/bioinformatics/btr026.

Sharma, G., Khatri, I. and Subramanian, S. (2016) 'Complete Genome of the Starch-Degrading Myxobacteria *Sandaracinus amylolyticus* DSM 53668T', *Genome Biology and Evolution*, 8, pp. 2520–2529. Available at: https://doi.org/10.1093/gbe/evw151.

Sheridan, P.O. *et al.* (2020) 'Gene duplication drives genome expansion in a major lineage of *Thaumarchaeota*', *Nature Communications*, 11. Available at: https://doi.org/10.1038/S41467-020-19132-X.

Simmons, C.W. *et al.* (2014) 'Metatranscriptomic analysis of lignocellulolytic microbial communities involved in high-solids decomposition of rice straw', *Biotechnology for Biofuels*, 7, p. 495. Available at: https://doi.org/10.1186/s13068-014-0180-0.

Singh, R. *et al.* (2016) 'Microbial enzymes: industrial progress in 21st century', *3 Biotech*, 6, 174. Available at: https://doi.org/10.1007/s13205-016-0485-8.

Sinha, S.K. and Datta, S. (2016) 'β-Glucosidase from the hyperthermophilic archaeon *Thermococcus sp.* is a salt-tolerant enzyme that is stabilized by its reaction product glucose', *Applied Microbiology and Biotechnology*, 100, pp. 8399–8409. Available at: https://doi.org/10.1007/s00253-016-7601-x.

Soares Júnior, F.L. *et al.* (2013) 'Endo-and exoglucanase activities in bacteria from mangrove sediment', *Brazilian Journal of Microbiology*, 44, pp. 969–976. Available at: https://doi.org/10.1590/S1517-83822013000300048.

Sobetzko, P., Travers, A. and Muskhelishvili, G. (2012) 'Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle.', *Proceedings of the National Academy of Sciences of the United States of America*, 109, pp. 42-50. Available at: https://doi.org/10.1073/pnas.1108229109.

Sondhi, S. *et al.* (2014) 'Purification and characterization of an extracellular, 184hese184i-alkali- stable, metal tolerant laccase from *Bacillus tequilensis* SN4', *PloS ONE*, 9, 96951. Available at: https://doi.org/10.1371/journal.pone.0096951.

Sonnenschein, N. *et al.* (2009) 'Ranges of control in the transcriptional regulation of *Escherichia coli*.', *BMC Systems Biology*, 3, 119. Available at: https://doi.org/10.1186/1752-0509-3-119.

Soto-Navarro, C. *et al.* (2020) 'Mapping co-benefits for carbon storage and biodiversity to inform conservation policy and action', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375, 20190128. Available at: https://doi.org/10.1098/rstb.2019.0128.

De Souza, P.M. and de Oliveira Magalhães, P. (2010) 'Application of microbial α-amylase in industry – a review', *Brazilian Journal of Microbiology*, 41,pp. 850–861. Available at: https://doi.org/10.1590/S1517-83822010000400004.

Souza, R.C. *et al.* (2015) 'Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes', *Applied Soil Ecology*, 86, pp. 106–112. Available at: https://doi.org/https://doi.org/10.1016/j.apsoil.2014.10.010.

Souza, R.C. *et al.* (2016) 'Shifts in taxonomic and functional microbial diversity with agriculture: How fragile is the Brazilian Cerrado?', *BMC Microbiology*, 16, 42. Available at: https://doi.org/10.1186/s12866-016-0657-z.

Spehn, E.M. *et al.* (2002) 'The role of legumes as a component of biodiversity in a cross-European study of grassland biomass nitrogen', *Oikos*, 98, pp. 205–218. Available at: https://doi.org/https://doi.org/10.1034/j.1600-0706.2002.980203.x.

Speth, C. *et al.* (2019) 'Galactosaminogalactan (GAG) and its multiple roles in *Aspergillus* pathogenesis', *Virulence*, 10, pp. 976–983. Available at: https://doi.org/10.1080/21505594.2019.1568174.

Srivastava, N. *et al.* (2014) 'A Review on Fuel Ethanol Production From Lignocellulosic Biomass', *International Journal of Green Energy*, 12, pp. 949–960. Available at: https://doi.org/10.1080/15435075.2014.890104.

Staley, J. T., and A. Konopka. 1985. 'Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats', *Annual Review of Microbiology*, 39, pp. 321-346.

Stewart, R.D. *et al.* (2018) 'Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen', *Nature Communications*, 9, 870. Available at: https://doi.org/10.1038/s41467-018-03317-6.

Stewart, R.D. *et al.* (2019) 'Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery', *Nature Biotechnology*, 37, pp. 953–961. Available at: https://doi.org/10.1038/s41587-019-0202-3.

Stolyar, S. *et al.* (2007) 'Metabolic 185hese185ing of a mutualistic microbial community', *Molecular Systems Biology*, 3, 92. Available at: https://doi.org/https://doi.org/10.1038/msb4100131.

Suen, G. *et al.* (2011) 'The complete genome sequence of *Fibrobacter succinogenes* s85 reveals a cellulolytic and metabolic specialist', *PloS ONE*, 6, 18814. Available at: https://doi.org/10.1371/journal.pone.0018814.

Suihko, M.L. and Skyttä, E. (2009) 'Characterisation of aerobically grown non-spore-forming bacteria from paper mill pulps containing recycled fibres', *Journal of Industrial Microbiology and Biotechnology*, 36, pp. 53–64. Available at: https://doi.org/10.1007/s10295-008-0472-0.

Sun, H. *et al.* (2016) 'Soil microbial community and microbial residues respond positively to minimum tillage under organic farming in Southern Germany', *Applied Soil Ecology*, 108, pp. 16–24. Available at: https://doi.org/https://doi.org/10.1016/j.apsoil.2016.07.014.

Sünnemann, M. *et al.* (2021) 'Low-intensity land-use enhances soil microbial activity, biomass and fungal-to-bacterial ratio in current and future climates', *Journal of Applied Ecology*, 58, pp. 2614–2625. Available at: https://doi.org/https://doi.org/10.1111/1365-2664.14004.

Susanti, D. *et al.* (2012) 'Complete genome sequence of *Desulfurococcus fermentans*, a hyperthermophilic cellulolytic crenarchaeon isolated from a freshwater hot spring in Kamchatka, Russia', *Journal of Bacteriology*, pp. 5703–5704. Available at: https://doi.org/10.1128/JB.01314-12.

Tabor, P.S. and Neihof, R.A. (1984) 'Direct determination of activities for microorganisms of Chesapeake Bay populations.', *Applied and environmental microbiology*, 48, pp. 1012–9.

Tanaka, T. *et al.* (2014) 'A Hidden Pitfall in the Preparation of Agar Media Undermines Microorganism Cultivability', *Applied and Environmental Microbiology*, 80, pp. 7659–7666. Available at: https://doi.org/10.1128/AEM.02741-14.

Tanaka, Y. *et al.* (2017) 'Isolation of Novel Bacteria Including Rarely Cultivated Phyla, *Acidobacteria* and *Verrucomicrobia*, from the Roots of Emergent Plants by Simple Culturing Method.', *Microbes and Environments*, 32, pp. 288–292. Available at: https://doi.org/10.1264/jsme2.ME17027.

Taylor, C.B. (1951) 'Nature of the Factor in Soil-extract Responsible for Bacterial Growth-stimulation', *Nature*, 168, pp. 115–116. Available at: https://doi.org/10.1038/168115a0.

Taylor, C.R. *et al.* (2012) 'Isolation of bacterial strains able to metabolize lignin from screening of environmental samples', *Journal of Applied Microbiology*, 113, pp. 521–530. Available at: https://doi.org/10.1111/j.1365-2672.2012.05352.x.

Ten, L.N. *et al.* (2006) '*Sphingobacterium composti* sp. nov., a novel Dnase-producing bacterium isolated from compost', *Journal of Microbiology and Biotechnology*, 16, pp. 1728–1733.

Terrapon, N. *et al.* (2015) 'Automatic prediction of polysaccharide utilisation loci in *Bacteroidota* species', *Bioinformatics*, 31, pp. 647–655. Available at: https://doi.org/10.1093/bioinformatics/btu716.

Teufel, F. *et al.* (2022) 'SignalP 6.0 predicts all five types of signal peptides using protein language models', *Nature Biotechnology*, 40, pp. 1023–1025. Available at: https://doi.org/10.1038/s41587-021-01156-3.

Thakur, P.B. *et al.* (2013) 'Characterization of Five ECF Sigma Factors in the Genome of *Pseudomonas syringae pv. Syringae B728a'*, *PLOS ONE*, 8,58846. Available at: https://doi.org/10.1371/journal.pone.0058846.

186hese, M.S., Ronna, B. and Ott, U. (2016) 'P value interpretations and considerations.', *Journal of Thoracic Disease*, 8, pp. 928–931. Available at: https://doi.org/10.21037/jtd.2016.08.16.

Thomas, F. *et al.* (2011) 'Environmental and gut Bacteroidetes: The food connection', *Frontiers in Microbiology*, 2, 93. Available at: https://doi.org/10.3389/fmicb.2011.00093.

Tian, J.H. *et al.* (2014) 'Occurrence of lignin degradation genotypes and phenotypes among prokaryotes', *Applied Microbiology and Biotechnology*, 98, pp. 9527–9544. Available at: https://doi.org/10.1007/s00253-014-6142-4.

Tilman, D. *et al.* (2011) 'Global food demand and the sustainable intensification of agriculture', *Proceedings of the National Academy of Sciences*, 108, pp. 20260–20264. Available at: https://doi.org/10.1073/pnas.1116437108.

Tjerneld, F. *et al.* (1985) 'Enzymatic hydrolysis of cellulose in aqueous two-phase systems. I. partition of cellulases from *Trichoderma reesei*', *Biotechnology and Bioengineering*, 27, pp. 1036–1043. Available at: https://doi.org/10.1002/bit.260270715.

Trias, J., Rosenberg, E.Y. and Nikaido, H. (1988) 'Specificity of the glucose channel formed by protein D1 of *Pseudomonas aeruginosa*.', *Biochimica et Biophysica Acta*, 938, pp. 493–496. Available at: https://doi.org/10.1016/0005-2736(88)90148-4.

Trivedi, C. *et al.* (2019) 'Plant-driven niche differentiation of ammonia-oxidizing bacteria and archaea in global drylands', *The ISME Journal*, 13, pp. 2727–2736. Available at: https://doi.org/10.1038/s41396-019-0465-1.

Tsiafouli, M.A. *et al.* (2015) 'Intensive agriculture reduces soil biodiversity across Europe', *Global Change Biology*, 21, pp. 973–985. Available at: https://doi.org/https://doi.org/10.1111/gcb.12752.

Tuck, S.L. *et al.* (2014) 'Land-use intensity and the effects of organic farming on biodiversity: a hierarchical meta-analysis', *Journal of Applied Ecology*, 51, pp. 746–755. Available at: https://doi.org/https://doi.org/10.1111/1365-2664.12219.

Tveit, A. *et al.* (2013) 'Organic carbon transformations in high-Arctic peat soils: Key functions and microorganisms', *ISME Journal*, 7, pp. 299–311. Available at: https://doi.org/10.1038/ismej.2012.99.

Tveit, A.T., Urich, T. and Svenning, M.M. (2014) 'Metatranscriptomic analysis of arctic peat soil microbiota', *Applied and Environmental Microbiology*, 80, pp. 5761–5772. Available at: https://doi.org/10.1128/AEM.01030-14.

Uffelmann, E. *et al.* (2021) 'Genome-wide association studies', *Nature Reviews Methods Primers*, 1, p. 59. Available at: https://doi.org/10.1038/s43586-021-00056-9.

Utturkar, S.M. *et al.* (2017) 'A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies', *Frontiers in Microbiology*, 8, 1272. Available at: https://doi.org/10.3389/fmicb.2017.01272.

Uzman, A. (2003) 'Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P.', *Biochemistry and Molecular Biology Education*, 31, pp. 212–214. Available at: https://doi.org/https://doi.org/10.1002/bmb.2003.494031049999.

Vaaje-Kolstad, G. *et al.* (2010) 'An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides', *Science*, 330, pp. 219–222. Available at: https://doi.org/10.1126/science.1192231.

Valášková, V. and Baldrian, P. (2006) 'Degradation of cellulose and hemicelluloses by the brown rot fungus *Piptoporus betulinus* - Production of extracellular enzymes and characterization of the major cellulases', *Microbiology*, 152, pp. 3613–3622. Available at: https://doi.org/10.1099/mic.0.29149-0.

Vargas-García, M.C. *et al.* (2007) 'In vitro Studies on lignocellulose degradation by microbial strains isolated from composting processes', *International Biodeterioration and Biodegradation*, 59, pp. 322–328. Available at: https://doi.org/10.1016/j.ibiod.2006.09.008.

Vartoukian, S.R., Palmer, R.M. and Wade, W.G. (2010a) 'Strategies for culture of "unculturable" bacteria', *FEMS Microbiology Letters*, 309, pp. 1–7. Available at: https://doi.org/10.1111/j.1574-6968.2010.02000.x.

Ventura, M. *et al.* (2007) 'Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum' *Microbiol Mol Biol Rev.* 71, pp. 495–548. Available at: https://doi.org/10.1128/MMBR.00005-07.

Verastegui, Y. *et al.* (2014) 'Multisubstrate isotope labeling and metagenomic analysis of active soil bacterial communities', *mBio*, 5. Available at: https://doi.org/10.1128/mBio.01157-14.

Vermaas, J. V. *et al.* (2015) 'Mechanism of lignin inhibition of enzymatic biomass deconstruction', *Biotechnology for Biofuels 2016 8:1*, 8, p. 217. Available at: https://doi.org/10.1186/s13068-015-0379-8.

Vermassen, A., Leroy S., Talon, R., Provot C., Popowska, M., and Desvaux, M. (2019) 'Cell Wall Hydrolases in Bacteria: Insight on the Diversity of Cell Wall Amidases, Glycosidases and Peptidases Toward Peptidoglycan', *Front Microbiol*, 10, 331. Available at: https://doi.org/10.3389/fmicb.2019.00331

Větrovský, T., Steffen, K.T. and Baldrian, P. (2014) 'Potential of cometabolic transformation of polysaccharides and lignin in lignocellulose by soil *Actinobacteria*', *PLoS ONE*, 9, 89108. Available at: https://doi.org/10.1371/journal.pone.0089108.

Vílchez, S. *et al.* (2000) 'Proline catabolism by *Pseudomonas putida*: cloning, characterization, and expression of the put genes in the presence of root exudates.', *Journal of Bacteriology*, 182, pp. 91–99. Available at: https://doi.org/10.1128/JB.182.1.91-99.2000.

Vo, C.-D.-T. *et al.* (2020) 'The O_2-independent pathway of ubiquinone biosynthesis is essential for denitrification in *Pseudomonas aeruginosa*.', *The Journal of Biological Chemistry*, 295, pp. 9021–9032. Available at: https://doi.org/10.1074/jbc.RA120.013748.

De Vries, F.T. *et al.* (2013) 'Soil food web properties explain ecosystem services across European land use systems', *Proceedings of the National Academy of Sciences*, 110, pp. 14296–14301. Available at: https://doi.org/10.1073/pnas.1305198110.

Wagg, C. *et al.* (2019) 'Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning', *Nature Communications*, 10, p. 4841. Available at: https://doi.org/10.1038/s41467-019-12798-y.

Wagner, G.P., Kin, K. and Lynch, V.J. (2012) 'Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples', *Theory in Biosciences*, 131. Available at: https://doi.org/10.1007/s12064-012-0162-3.

Wang, C. *et al.* (2016) 'Metagenomic analysis of microbial consortia enriched from compost: New insights into the role of *Actinobacteria* in lignocellulose decomposition', *Biotechnology for Biofuels*, 9. Available at: https://doi.org/10.1186/s13068-016-0440-2.

Wang, C. *et al.* (2021) 'Miscanthus: A fast-growing crop for environmental remediation and biofuel production', *GCB Bioenergy*, 13, pp. 58–69. Available at: https://doi.org/https://doi.org/10.1111/gcbb.12761.

Wang, L., Littlewood, J. and Murphy, R.J. (2013) 'Environmental sustainability of bioethanol production from wheat straw in the UK', *Renewable and Sustainable Energy Reviews*, pp. 715–725. Available at: https://doi.org/10.1016/j.rser.2013.08.031.

Wang, S. *et al.* (2020) 'Evolutionary Timeline and Genomic Plasticity Underlying the Lifestyle Diversity in *Rhizobiales*', *mSystems*, 5, e00438-20. Available at: https://doi.org/10.1128/mSystems.00438-20.

Wang, W. *et al.* (2011) 'Characterization of a microbial consortium capable of degrading lignocellulose', *Bioresource Technology*, 102, pp. 9321–9324. Available at: https://doi.org/10.1016/j.biortech.2011.07.065.

Wang, W. and Fang, J. (2009) 'Soil respiration and human effects on global grasslands', *Global and Planetary Change*, 67, pp. 20–28. Available at: https://doi.org/https://doi.org/10.1016/j.gloplacha.2008.12.011.

Wang, X. *et al.* (2015) 'Stable-isotope probing identifies uncultured *Planctomycetes* as primary degraders of a complex heteropolysaccharide in soil', *Applied and Environmental Microbiology*, 81, pp. 4607–4615. Available at: https://doi.org/10.1128/AEM.00055-15.

Wang, X. *et al.* (2017) 'A Putative Type II Secretion System Is Involved in Cellulose Utilisation in *Cytophaga hutchinsonii*', *Frontiers in Microbiology*, 8,1482. Available at: https://doi.org/10.3389/fmicb.2017.01482.

Ward, N.L. *et al.* (2009) 'Three Genomes from the Phylum *Acidobacteria* Provide Insight into the Lifestyles of These Microorganisms in Soils', *Applied and Environmental Microbiology*, 75, pp. 2046–2056. Available at: https://doi.org/10.1128/AEM.02294-08.

Watve, M. *et al.* (2000) 'The "K" selected oligophilic bacteria: A key to uncultured diversity?', *Current Science*, 78, pp. 1535–1542.

Weiss, B. *et al.* (2021) 'Unravelling a Lignocellulose-Decomposing Bacterial Consortium from Soil Associated with Dry Sugarcane Straw by Genomic-Centered Metagenomics.', *Microorganisms*, 9. Available at: https://doi.org/10.3390/microorganisms9050995.

Werner, J. *et al.* (2014), Proteogenomics of the halophile *H. tiamatea*. Environ Microbiol, 16: 2525-2537. https://doi.org/10.1111/1462-2920.12393

Westereng, B. *et al.* (2015) 'Enzymatic cellulose oxidation is linked to lignin by long-range electron transfer', *Scientific Reports*, 5. Available at: https://doi.org/10.1038/srep18561.

Wieder, W.R. *et al.* (2015) 'Representing life in the Earth system with soil microbial functional traits in the MIMICS model', *Geoscientific Model Development*, 8, pp. 1789–1808. Available at: https://doi.org/10.5194/gmd-8-1789-2015.

Wieder, W.R., Bonan, G.B. and Allison, S.D. (2013) 'Global soil carbon projections are improved by modelling microbial processes', *Nature Climate Change*, 3, pp. 909–912. Available at: https://doi.org/10.1038/nclimate1951.

Wilhelm, R.C. *et al.* (2019) 'Bacterial contributions to delignification and lignocellulose degradation in forest soils with metagenomic and quantitative stable isotope probing', *ISME Journal*, 13, pp. 413–429. Available at: https://doi.org/10.1038/s41396-018-0279-6.

Winkler, K. *et al.* (2021) 'Global land use changes are four times greater than previously estimated', *Nature Communications*, 12, p. 2501. Available at: https://doi.org/10.1038/s41467-021-22702-2.

Withers, E. *et al.* (2020) 'Use of untargeted metabolomics for assessing soil quality and microbial function', *Soil Biology and Biochemistry*, 143, 107758.

Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20, 257. Available at: https://doi.org/10.1186/s13059-019-1891-0.

Wood, J.L., Tang, C. and Franks, A.E. (2018) 'Competitive Traits Are More Important than Stress-Tolerance Traits in a Cadmium-Contaminated Rhizosphere: A Role for Trait Theory in Microbial Ecology', *Frontiers in Microbiology*, 9, 121. Available at: https://doi.org/10.3389/fmicb.2018.00121.

Wu, D. *et al.* (2012) 'Structural Basis of Substrate Binding Specificity Revealed by the Crystal Structures of Polyamine Receptors SpuD and SpuE from Pseudomonas aeruginosa', *Journal of Molecular Biology*, 416, pp. 697–712. Available at: https://doi.org/https://doi.org/10.1016/j.jmb.2012.01.010.

Wu, M. *et al.* (2021) 'Chemical composition of soil organic carbon and aggregate stability along an elevation gradient in Helan Mountains, northwest China', *Ecological Indicators*, 131, p. 108228. Available at: https://doi.org/https://doi.org/10.1016/j.ecolind.2021.108228.

Wuaden, C.R. *et al.* (2020) 'Early adoption of no-till mitigates soil organic carbon and nitrogen losses due to land use change', *Soil and Tillage Research*, 204, 104728. Available at: https://doi.org/https://doi.org/10.1016/j.still.2020.104728.

Xia, Y. *et al.* (2013) 'Mining of Novel Thermo-Stable Cellulolytic Genes from a Thermophilic Cellulose-Degrading Consortium by Metagenomics', *PLoS ONE*, 8, 53779. Available at: https://doi.org/10.1371/journal.pone.0053779.

Xia, Y. *et al.* (2014) 'Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis', *Scientific Reports*, 4. Available at: https://doi.org/10.1038/srep06708.

Xia, Y. *et al.* (2016) 'Cellular adhesiveness and cellulolytic capacity in *Anaerolineae* revealed by omics-based genome interpretation', *Biotechnology for Biofuels*, 9(1). Available at: https://doi.org/10.1186/s13068-016-0524-z.

Yang, Chenxian *et al.* (2021) 'A Novel Polyphenol Oxidoreductase OhLac from *Ochrobactrum sp.* J10 for Lignin Degradation', *Frontiers in Microbiology*, 12, 694166. Available at: https://doi.org/10.3389/fmicb.2021.694166.

Yang, Chao *et al.* (2021) 'Soil salinity regulation of soil microbial carbon metabolic function in the Yellow River Delta, China', *Science of The Total Environment*, 790, p. 148258. Available at: https://doi.org/https://doi.org/10.1016/j.scitotenv.2021.148258.

Yeager, C.M. *et al.* (2017) 'Polysaccharide degradation capability of *Actinomycetales* soil isolates from a semiarid grassland of the Colorado Plateau', *Applied and Environmental Microbiology*, 83, pp. 1–19. Available at: https://doi.org/10.1128/AEM.03020-16.

Yuan, Y. *et al.* (2012) 'Fatty Acid Biosynthesis in *Pseudomonas aeruginosa* Is Initiated by the FabY Class of β-Ketoacyl Acyl Carrier Protein Synthases', *Journal of Bacteriology*, 194, pp. 5171–5184. Available at: https://doi.org/10.1128/JB.00792-12.

Zabel, F. *et al.* (2019) 'Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity', *Nature Communications*, 10, p. 2844. Available at: https://doi.org/10.1038/s41467-019-10775-z.

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*, 4, 104. Available at: https://doi.org/10.3389/fmicb.2013.00104.

Zhang, H. *et al.* (2018) 'DbCAN2: A meta server for automated carbohydrate-active enzyme annotation', *Nucleic Acids Research*, 46, pp. 95 – 101. Available at: https://doi.org/10.1093/nar/gky418.

Zhang, H. *et al.* (2020) 'Microbial dynamics and soil physicochemical properties explain large-scale variations in soil organic carbon', *Global Change Biology*, 26, pp. 2668–2685. Available at: https://doi.org/https://doi.org/10.1111/gcb.14994.

Zhang, J. *et al.* (2021) 'High-throughput cultivation and identification of bacteria from the plant root microbiota', *Nature Protocols*, 16, pp. 988–1012. Available at: https://doi.org/10.1038/s41596-020-00444-7.

Zhang, Q. and Bao, J. (2017) 'Industrial cellulase performance in the simultaneous saccharification and co-fermentation (SSCF) of corn stover for high-titer ethanol production', *Bioresources and Bioprocessing*, 4, 17. Available at: https://doi.org/10.1186/s40643-017-0147-7.

Zhang, T. *et al.* (2011) 'Identification of a haloalkaliphilic and thermostable cellulase with improved ionic liquid tolerance', *Green Chemistry*,13, pp. 2083-2090. Available at: https://doi.org/10.1039/C1GC15193B.

Zhao, X., Zhang, L. and Liu, D. (2012) 'Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose', *Biofuels, Bioproducts and Biorefining*, 6, pp. 465–482. Available at: https://doi.org/10.1002/bbb.1331.

Zheng, Q. *et al.* (2019) 'Soil multifunctionality is affected by the soil environment and by microbial community composition and diversity', *Soil Biology and Biochemistry*, 136, 107521. Available at: https://doi.org/https://doi.org/10.1016/j.soilbio.2019.107521.

Zheng, Y. *et al.* (2019) 'Effect of lignin degradation product sinapyl alcohol on laccase catalysis during lignin degradation', *Industrial Crops and Products*, 139, 111544. Available at: https://doi.org/https://doi.org/10.1016/j.indcrop.2019.111544.

Zhu, D. *et al.* (2017) 'Biodegradation of alkaline lignin by *Bacillus ligniniphilus* L1', *Biotechnology for Biofuels*, 10, 44. Available at: https://doi.org/10.1186/s13068-017-0735-y.

Zheng, J. *et al*. (2023) 'dbCAN3: automated carbohydrate-active enzyme and substrate annotation', *Nucleic Acids Research*, gkad323. Available at: https://doi.org/10.1093/nar/gkad328.

Zhu, Y. and McBride, M.J. (2014) 'Deletion of the *Cytophaga hutchinsonii* type IX secretion system gene sprP results in defects in gliding motility and cellulose utilisation', *Applied Microbiology and Biotechnology*, 98, pp. 763–775. Available at: https://doi.org/10.1007/s00253-013-5355-2.

Zhu, Y. and McBride, M.J. (2017) 'The unusual cellulose utilisation system of the aerobic soil bacterium Cytophaga hutchinsonii', *Applied Microbiology and Biotechnology*, 101, pp. 7113–7127. Available at: https://doi.org/10.1007/s00253-017-8467-2.

Žifčáková, L. *et al.* (2017) 'Feed in summer, rest in winter: microbial carbon utilization in forest topsoil', *Microbiome*, 5, 122. Available at: https://doi.org/10.1186/s40168-017-0340-0.

ZoBell, C. E. (1946) '*Marine Microbiology, a Monograph on Hydrobacteriology*', Waltham, Massachusetts: Chronica Botanica Company, 1946.

Zorrilla, F. *et al.* (2021) 'metaGEM: reconstruction of genome scale metabolic models directly from metagenomes', *Nucleic Acids Research*, 49, pp. 126–126. Available at: https://doi.org/10.1093/nar/gkab815.

Zverlov, V. V. and Schwarz, W.H. (2008) 'Bacterial Cellulose Hydrolysis in Anaerobic Environmental Subsystems-*Clostridium thermocellum* and *Clostridium stercorarium*, Thermophilic Plant-fiber Degraders', *Annals of the New York Academy of Sciences*, 1125, pp. 298–307. Available at: https://doi.org/10.1196/annals.1419.008.