ECOLOGY OF THE ELUSIVE: GENOME-INFORMED INVESTIGATION OF SOIL MICROBIAL DIVERSITY

Lauren Alteio
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Terrestrial and Aquatic Ecology Commons

Recommended Citation
https://doi.org/10.7275/hsq9-ph87 https://scholarworks.umass.edu/dissertations_2/1680

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
ECOLOGY OF THE ELUSIVE: GENOME-INFORMED INVESTIGATION OF SOIL MICROBIAL DIVERSITY

A Dissertation Presented

by

LAUREN VICTORIA ALTEIO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2019

Graduate Program in Organismic and Evolutionary Biology
ECOLOGY OF THE ELUSIVE: GENOME-INFORMED INVESTIGATION OF SOIL MICROBIAL DIVERSITY

A Dissertation Presented

by

LAUREN VICTORIA ALTEIO

Approved as to style and content by:

________________________________________
Laura A. Katz, Chair

________________________________________
Courtney C. Babbitt, Member

________________________________________
Kristina Stinson, Member

________________________________________
Tanja Woyke, Member

________________________________________
Paige S. Warren, Graduate Program Director

Graduate Program in Organismic and Evolutionary Biology
DEDICATION

To Jeremy Hayward for your unrelenting passion for science and fungi, and for your belief in me from the start. You were a dear friend and mentor. Not a day goes by that you are not on my mind.
ACKNOWLEDGMENTS

Thank you Laura for supporting me through the roughest period of my graduate career. You advocated for me so strongly, I cannot express how grateful I am to have you as a mentor. I am so proud to be a part of the Katz lab. Thank you for accepting me as your own when you had very little to gain.

Thank you Tanja for your support, enthusiasm and advocacy, and for hosting me at the JGI during my SCGSR internship. I am so glad I had the opportunity to work with you. I admire your passion for science and ability to navigate challenging situations with ease. I hope we have many more years of collaboration to come!

Thank you Courtney and Kristina for your support and helpful discussions about my dissertation research, and your helpful career advice.

Thank you Paige for your diplomacy and advocacy. I would not have made it to the finish line without you. You are doing an incredible job as GPD!

Thank you Stephanie and Dagmar for your support and enthusiastic discussions about soils and Acidobacteria! I appreciate you taking me in and hosting me during my visits to DoME.

I am grateful to Jeff Blanchard for setting me up for a successful scientific career. Thank you to the Blanchard lab for all of their support.

Thank you to William Rodriguez for being a lab-mate and brother to me and a fierce supporter and advocate for my science. You are amazing! Best of luck in everything you do in your future.

I am grateful to Alex Truchon for your dedication to our projects and interesting conversations about our results.

Thank you to the Harvard Forest community. Since 2012 you have become a family for me. You helped cultivate my love of science, and provided a supportive environment for me to grow, learn and lead.

Thank you to all of my friends at UMass, the JGI, Walnut Creek run club, DoME and beyond who supported me both near and far through the troughs and peaks of my graduate experience. I truly would not be here today if it were not for all of you. I am grateful to Trisha Zintel for your support, heart-to-hearts and painting sessions. Thank you Kit Straley for your encouragement and creativity in the face of great challenges. Thank you Emily Fusco for your wise words and for being my hiking and running partner. I am grateful to Carolyn Anderson and Elsa Cousins for all of your cooking, adventuring, fermenting and rallying over the last few years. You are great friends and roommates!

Thank you Kim De Paepe, Nora Grossschmidt and Dimitri Meier for making me feel right at home in Vienna. I look forward to seeing you all soon!
Thank you to the SUNY-ESF community, particularly the Horton Lab, who supported me as an undergraduate and encouraged me to pursue a PhD. I am grateful to Tom Horton for accepting me into your lab and allowing me to chart my own course. Thank you Becka for being a mentor, friend and confidant while I was an undergraduate and beyond.

Thank you, Benni, for your relentless support and positivity through the last part of my PhD candidacy. Du bist afoch leiwand, mein Freund.

Lastly, (but definitely not least) thank you to my family. Thank you Aunt Andrea and Uncle Bruce for always leaving your doors and hearts open for me. Thank you Morgan for standing by me through it all. Thank you Mom and Dad for believing in me no matter what. Words cannot express how much your love and support has meant to me throughout this journey.
ABSTRACT

ECOLOGY OF THE ELUSIVE: GENOME-INFORMED INVESTIGATION OF SOIL MICROBIAL ECOLOGY

SEPTEMBER 2019

LAUREN VICTORIA ALTEIO

B.S., STATE UNIVERSITY OF NEW YORK COLLEGE OF ENVIRONMENTAL SCIENCE AND FORESTRY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Laura A. Katz

Soil is considered one of the most diverse ecosystems on Earth, harboring diversity of organisms across the three domains of life. It is spatially and chemically heterogeneous: properties that intertwine in a complex matrix to support organismal diversity and function across different scales. Soil microorganisms both respond to and drive changes in ecosystems through metabolic activities. A single gram of soil is teeming with millions of cells comprised of thousands of species. Much of this diversity remains uncharacterized due to technical and methodological challenges faced by soil ecologists. Due to the complex physicochemical properties of soil and cross-feeding interactions between organisms, it is difficult to culture microorganisms in isolation. The immense biological diversity of soils also reduces bioinformatic genome assembly efficiency, therefore obscuring the scope of diversity. As one of Earth's main reservoirs of stored carbon, containing roughly two-thirds of carbon globally, terrestrial ecosystems may serve as a carbon source under future climate scenarios and drive further climate change. Despite challenges associated with the study of soil microorganisms, it remains critical to discover and describe diversity of microbial communities in soils if we are to understand resilience of our ecosystems to climate change.

Surveys of microbial diversity and function in soil have been conducted using amplicon sequencing, metagenomics, and metatranscriptomics, however a large knowledge gap persists in
the characterization of diversity and ecological niches of elusive microorganisms. These are organisms that are typically recalcitrant to laboratory culture, and may appear in relatively low abundance in soil communities or exhibit a high degree of population microheterogeneity, thereby resulting in poor representation in genome assemblies. The focus of my dissertation research is the application of complementary genomic techniques in order to uncover more of the previously unknown microbial diversity contained in forest soils, and link this diversity to higher-level ecosystem function. Much of what is known about soil diversity has been contributed through cultivation-independent investigations, however diversity estimates indicate that we are only beginning to scratch the surface of bacterial, archaeal, and viral diversity in forest soils. We are therefore vastly underestimating the roles these organisms play in biogeochemical processes, such as the release of CO$_2$ to the atmosphere through respiration. However, the scope of microbial diversity and their suite of metabolic functions remain challenging to link to ecosystem level processes due to methodological limitations.

For chapter 1 of my dissertation, I worked in collaboration with researchers at the University of Vienna using extensive literature searches to explore the different spatial scales at which we study microbial diversity and function with the goal of linking microorganisms and their role as drivers of higher level processes. This work suggests that the level at which microorganisms interact, termed the 'microbial consortium', is a key scale which provides insights into microbial diversity, function, and enables scaling up from the single cell to the ecosystem. In chapter 2, I applied complementary metagenomic techniques to the discovery of soil biological diversity, including bulk metagenomics and a pooled, cell-sorting approach coupled to high-throughput sequencing, termed mini-metagenomics. In combination, these approaches uncover the genetic diversity of elusive microorganisms at the Harvard Forest Long-Term Ecological Research (LTER) site. Together, these approaches have generated some of the highest quality metagenome
assembled genomes (MAGs) to date from this LTER experimental site, and have revealed a swath of diversity beyond the organisms typically found in high abundance in the soil. I demonstrate how complementary metagenomic techniques facilitate the discovery of biological diversity by highlighting the expanded knowledge of potential intracellular bacteria in the phylum *Bacteroidetes*. In chapter 3, I characterize the metabolism of representatives in the phylum *Acidobacteria* subdivision 2, which are abundant in forest soils but have yet to be described as there are no available genome sequences in this taxonomic group. Finally, chapter 4 describes sixteen novel giant viruses which have been discovered in Harvard Forest soil for the first time in collaboration with researchers at the Joint Genome Institute. These expand knowledge of phylogenetic diversity of the nucleocytoplasmic large DNA viruses (NCLDV) by 21%, and further demonstrate the utility of complementary metagenomic approaches in uncovering diversity of elusive viral entities in addition to microbial life.

Observed changes at Prospect Hill, the longest-running soil warming experimental site at Harvard Forest, reveal increases in soil microbial respiration, increases in nitrogen mineralization, decreases in soil organic matter and decreases in the overall microbial biomass of these soils in response to warming. Based on these findings, we can expect similar changes to occur at the Barre Woods warming experiment, which was established at the Harvard Forest LTER site in 2002. Additionally, we may anticipate similar changes in temperate forest soils as the Earth's climate changes and surface temperatures continue to rise. With these changes, the microbial community must change and adapt to shifting nutrient and substrate availability, moisture conditions and changing soil structure. This dissertation work supports our understanding of the expansion of niches for soil microorganisms with oligotrophic growth strategies and flexible metabolism. These traits will enable soil organisms to cope with a nutrient-limited environment that is predicted to occur in response to long-term climate change.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... v

ABSTRACT .............................................................................................................................. vii

LIST OF TABLES .................................................................................................................... xvi

LIST OF FIGURES ................................................................................................................ xvi

1. THE 'MICROBIAL CONSORTIUM': THE KEY SCALE FOR INVESTIGATING MICROBIALY-
 MEDIATED PROCESSES ACROSS SOIL SCALES .......................................................... 1
   1.1. Abstract ......................................................................................................................... 1
   1.2 Introduction .................................................................................................................. 2
   1.3 Advances in soil ecology at higher-level scales ......................................................... 4
      1.3.1 'Soil core/profile scale' - controls presented by plant roots, mycorrhizal fungi and soil fauna ........................................................... 5
      1.3.2 Methodological approaches and challenges at the ‘soil core/profile’ scale ......................... 7
      1.3.3 The millimeter-scale – including soil micro- and macroaggregates ......................... 9
      1.3.4 Methodological approaches and challenges at the aggregate scale ....................... 12
   1.4 The 'microbial consortium': The key scale for investigating microbially-mediated processes in soil .......................................................... 13
      1.4.1 Simplifying soils in the lab with microfluidics and ‘soil-on-a-chip’ ............................. 16
      1.4.2 Mini-metagenomics for capturing diversity at the 'microbial consortium' scale ................................. 17
      1.4.3 Single-cell investigations of the ‘microbial consortium’ ...................................... 19
      1.4.4 Computational approaches to the ‘microbial consortium’ .................................. 21
   1.5 Scaling up towards the ecosystem scale .................................................................... 22
   1.6 Synthesis ....................................................................................................................... 25
2. COMPLEMENTARY METAGENOMIC APPROACHES REVEAL BACTERIAL AND ARCHAEAL DIVERSITY IN FOREST SOIL ................................................................. 29
  
  2.1 Abstract ........................................................................................................ 29
  
  2.2 Introduction .................................................................................................. 30
  
  2.3 Results and Discussion ................................................................................ 34
    2.3.1 Improved assembly and binning from mini-metagenomes .......... 34
    2.3.2 Expansion of phylogenetic diversity ..................................................... 36
    2.3.3 Caveats of mini-metagenomics .............................................................. 39
    2.3.4 Representation of sorted-MAGs and MAGs Across Terrestrial Soil Metagenomes ................................................................. 40
    2.3.5 Biological Insights into Carbon Metabolism in Soil Bacteroidetes ................................................................. 42
    2.3.6 Reduced Carbon Metabolism in an Unclassified Bacteroidetes Clade ................................................................. 45
  
  2.4 Conclusions .................................................................................................. 47
  
  2.5 Methods ........................................................................................................ 47
    2.5.1 Sample Collection and Incubation ........................................................ 47
    2.4.2 Sample Preparation and Cell Sorting .................................................... 48
    2.4.3 Mini-Metagenomes ............................................................................... 49
    2.4.4 Bulk Metagenomes ............................................................................... 49
    2.4.6 Phylogenetic Tree Construction and Phylogenetic Diversity ......... 50
    2.4.5 Genome Binning and Quality Assessment ........................................ 50
    2.4.7 Gene Recruitment ............................................................................... 52
    2.4.8 Metabolic Insights ............................................................................... 53
  
  3. GENOME-INFORMED ECOPHYSIOLOGICAL CHARACTERIZATION OF NOVEL, UNCULTIVATED SUBDIVISION 2 ACIDOBACTERIA ........................................ 60
  
  3.1 Abstract ........................................................................................................ 60
  
  3.2 Introduction .................................................................................................. 61
3.3 Results and Discussion ........................................................................................................... 63

3.3.1 Taxonomic distribution across bulk- and mini-metagenomes reveals abundant acidobacteria in Harvard Forest soils .......... 63

3.3.2 Phylogenomic distribution of acidobacterial MAGs in forest soil expands known diversity of subdivision 2 acidobacteria .......... 65

3.3.3 Comparison of genomic similarity across acidobacteria supports discovery of novel species ............................................. 66

3.3.4 Genomic potential reflects functional similarities driven by acidobacterial subdivision ......................................................................................... 68

3.3.5 Carbohydrate degradation potential indicates overlap in metabolism and niche specialization across Acidobacteria ...... 70

3.3.6 Gene expression analysis suggests metabolic flexibility of subdivision 2 acidobacteria ......................................................................................... 74

  3.3.6.1 Evidence of assimilatory sulfur metabolism .................. 75

  3.3.6.2 Assimilatory nitrate reduction in subdivision 2 ............ 76

  3.3.6.3 Terminal oxidases enable respiration in low oxygen environments ................................................................................................. 77

  3.3.6.4 Hydrogenases enable hydrogen scavenging and utilization ......................................................................................... 78

3.3.7 Metabolic strategies across Acidobacteria reflect environmental selection and niche specialization ............................................. 79

3.8 Synthesis ................................................................................................................................. 81

3.9 Methods ................................................................................................................................. 82

  3.9.1 Sample Collection and Processing ............................................................. 82

  3.9.2 Mini-Metagenomics ...................................................................................... 82

  3.9.3 Bulk Metagenomics ...................................................................................... 83

  3.9.4 Metatranscriptomics ...................................................................................... 83

  3.9.5 Genome Binning and Quality Assessment ............................................... 84

  3.9.6 Distribution of taxa in bulk and sorted-MAGs ........................................ 85
3.9.7 Taxonomic distribution of extracted 16S rRNA gene sequences ... 85
3.9.8 Phylogenomic analysis of MAGs and Sorted-MAGs ................. 86
3.9.9 Average Nucleotide Identity and Amino Acid Identity ............... 86
3.9.10 Metabolic Insights from MAGs ........................................ 86
3.9.11 Estimation of gene expression in Acidobacteria ..................... 87
3.9.12 Detecting hydrogenases in acidobacterial MAGs ..................... 88

4. HIDDEN DIVERSITY OF SOIL GIANT VIRUSES ........................................ 101
   4.1 Abstract .................................................................................. 101
   4.2 Introduction ............................................................................ 101
   4.3 Results .................................................................................... 103
      4.3.1 Mini-metagenomics facilitated the discovery of giant viruses ... 103
      4.3.2 Cell-sorted viral particles expand known diversity of NCLDV ...... 105
      4.3.3 Genomic features of soil giant viruses .................................... 106
      4.3.4 Genome novelty of soil giant viruses ..................................... 107
      4.3.5 True diversity of giant viruses in forest soil ....................... 108
   4.4 Discussion ............................................................................... 109
   4.5 Methods ................................................................................ 111
      4.5.1 Sampling and sample preparation .......................................... 111
      4.5.2 Mini-Metagenomes .............................................................. 112
      4.5.3 Bulk Metagenomes .............................................................. 112
      4.5.4 Metatranscriptomes ............................................................. 113
      4.5.5 Metagenome Binning .......................................................... 113
      4.5.6 Screening for giant viruses .................................................. 113
      4.5.7 Annotation and quality control of viral genome bins ............ 114
      4.5.8 Experimental benchmarking of the mini-metagenomics
           approach ............................................................................... 114
4.5.9 Computational benchmarking of giant virus metagenomic binning ................................................................. 115

4.5.10 Phylogenomics ................................................................................. 115

4.5.11 Major capsid protein analysis ............................................................ 117

4.5.12 Gene sharing network .................................................................... 117

APPENDICES

A. SUPPLEMENTARY TABLES ................................................................. 123

B. SUPPLEMENTARY FIGURES ............................................................... 125

BIBLIOGRAPHY ...................................................................................... 138
LIST OF TABLES

Table 3.1. Number of reads assigned to Acidobacteria from the mini- and bulk metagenome datasets using RDP Classifier ................................................................. 87

Table 3.2. CheckM quality assessment of bulk and sorted-MAGs classified to the phylum Acidobacteria ........................................................................................................... 88

Table 3.3. Number of expressed KEGG genes in subdivision 2 acidobacteria draft genomes. ......................................................................................................................... 89

Table 3.4. Number of expressed CAZy genes in each subdivision 2 acidobacteria genome ......................................................................................................................... 90

Table S1. RDP Classifier results for full-length 16S sequences extracted from acidobacteria MAGs ................................................................................................. 122
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. Conceptual image depicting the different scales at which processes occur in soil and exemplifying major process controls at each scale</td>
<td>27</td>
</tr>
<tr>
<td>1.2. Diversity contributes to the complexity of soil ecosystems</td>
<td>28</td>
</tr>
<tr>
<td>2.1. Overview of mini-metagenome and bulk metagenome approaches used in this study</td>
<td>54</td>
</tr>
<tr>
<td>2.2. Assessment of sorted-MAG and MAG quality</td>
<td>55</td>
</tr>
<tr>
<td>2.3. Phylogenetic diversity of soil taxa identified in this study</td>
<td>56</td>
</tr>
<tr>
<td>2.4. Comparison of MAGs from this study with published data from terrestrial metagenomes</td>
<td>57</td>
</tr>
<tr>
<td>2.5. Insights into carbon metabolism within the phylum Bacteroidetes</td>
<td>58</td>
</tr>
<tr>
<td>3.1. Taxonomic distribution in mini- and bulk metagenomes</td>
<td>91</td>
</tr>
<tr>
<td>3.2. Distribution of acidobacterial clades across the bulk metagenome and mini-metagenome samples</td>
<td>92</td>
</tr>
<tr>
<td>3.3. Concatenated marker gene tree of acidobacterial MAGs and references</td>
<td>93</td>
</tr>
<tr>
<td>3.4. Quality assessment of MAGs and reference MAGs from GTDB</td>
<td>94</td>
</tr>
<tr>
<td>3.5 Maximum likelihood tree of acidobacterial subdivisions 1, 2, 3, 4, 6 and 8 based on full-length 16S rRNA gene.</td>
<td>95</td>
</tr>
<tr>
<td>3.6. PCoA of COG/NOG genes across Acidobacteria subdivisions</td>
<td>96</td>
</tr>
<tr>
<td>3.7. PCoA of CAZy gene families across Acidobacteria subdivisions</td>
<td>97</td>
</tr>
<tr>
<td>3.8. Metabolic diagram of the 5 subdivision 2 acidobacteria draft genomes</td>
<td>98</td>
</tr>
<tr>
<td>4.1. Discovery pipeline for soil giant viruses</td>
<td>117</td>
</tr>
<tr>
<td>4.2. Expansion of NCLDV diversity by soil giant viruses</td>
<td>118</td>
</tr>
<tr>
<td>4.3. Genome novelty of soil giant viruses</td>
<td>119</td>
</tr>
<tr>
<td>4.4. Hidden diversity of giant viruses in bulk metagenomes</td>
<td>120</td>
</tr>
</tbody>
</table>
S1. Full concatenated marker gene tree of acidobacterial MAGs and references. .................. 124

S2. Genomic similarity across the 5 subdivision 2 acidobacteria characterized in this study ........................................................................................................................................................................... 125

S3. Genomic similarity across subdivision 1 acidobacteria ................................................................................................................................. 126

S4. PcoA clustering COG/NOG categories across acidobacteria MAGs and reference genomes........................................................................................................................................................................................................... 127

S5. Heatmap of EggNOG genes across Acidobacteria subdivisions ........................................................................................................... 128

S6. PcoA clustering CAZy categories across acidobacteria MAGs and reference genomes ...... 129

S7. Heatmap of CAZy gene families across acidobacterial subdivisions ............................................. 130

S8. Relative abundance of CAZy gene families across acidobacteria MAGs and reference genome sequences ...................................................................................................................................................................................................... 131

S9. Phylogenetic tree of group 1h/5 [NiFe] hydrogenases .............................................................................. 132

S10. Glycoside hydrolases expressed across subdivision 2 acidobacteria MAGs .................. 133

S11. Relative abundance of CAZy gene families expressed across subdivision 2 MAGs ....... 134

S12. Distribution of acidobacterial classes in the mineral and organic soil horizons ............. 135
CHAPTER 1
THE ‘MICROBIAL CONSORTIUM’: THE KEY SCALE FOR INVESTIGATING MICROBIALLY-MEDIATED PROCESSES ACROSS SOIL SCALES

1.1. Abstract

Microorganisms are abundant and diverse in soils, and constitute the main drivers of biogeochemical processes. They govern processes across different scales – from the single-cell (µm) to the ecosystem (Km). Yet, across these scales, processes are controlled by different factors - from the genetic-physiological capabilities at the single-cell level to land-use and vegetation type at the ecosystem scale. To fully appreciate ecosystem function, we need to understand the composition of microbial assemblages, the activities and interactions of microorganisms, and the controls governing them across these scales. In this perspective, we discuss different spatial scales that should be considered when investigating soil microbial processes along with their controls. These scales include (1) the ‘soil core/profile’ scale at which plant roots and mycorrhizal fungi interact with bacteria; (2) soil macroaggregates and microaggregates; and (4) the scale at which single cells interact within ‘microbial consortia’. Until now, most investigations have been made at discrete scales, from the ‘soil core/profile’ to the ecosystem-scale, with more recent advances at the single-cell and soil aggregate levels. We propose that the ‘microbial consortium’ is a key scale that influences processes across scales, due to emergent properties captured neither at the single cell nor ecosystem scale. Due to technical challenges this scale remains underexplored, and we propose more refined methods and tools for these investigations.

1 Alteio LV, Eichorst SA, Kaiser C, Katz LA, Richter A, Wanek W, Woebken D. and Woyke, T. Toward the understanding of spatial scale: Microbial consortia as the key scale for investigating soil processes. In prep for submission to FEMS. Authors listed in alphabetical order.
1.2 Introduction

Soils are among the most diverse and complex habitats for microbial life on the planet with estimates of $10^9$ to $10^{10}$ cells per gram of soil, harboring 1,000 to 1,000,000 species within this gram (e.g. R. Amann & Rosselló-Móra, 2016; Gans, Wolinsky, & Dunbar, 2005; Jay T Lennon & Locey, 2016; Locey & Lennon, 2016) As regulators of ecosystem processes, understanding taxonomic diversity and functional redundancy within soil microbial communities is critical for predicting habitat resilience under conditions of environmental change (Allison & Martiny, 2008; Delgado-Baquerizo et al., 2016). However, the extensive microbial diversity and the function carried out by members of the microbial community remains largely uncharacterized due to methodological challenges. The complex physicochemical properties of soil in addition to high levels of spatial heterogeneity make accurate estimates of diversity and functional potential difficult to attain (Nemerugt et al., 2013).

The main driver of microbial diversity in soil communities is presumed to be spatial heterogeneity of soil as this provides ecological opportunities for various microbial species, populations and communities (Nagy, Ábrahám, Keymer, & Galajda, 2018; Nemerugt et al., 2013). However, studies of the spatial properties of soil are limited (as reviewed in Baveye et al., 2018), thereby restricting scientific understanding of this key property in shaping microbial community assemblages, and driving higher-level processes. Soil spatial heterogeneity plays a role in the diversity and function of microorganisms across scales - from the µm scale at which single cells interact, to the meter scale where large-scale ecosystem processes occur (Štursová, Bártá, Šantrůčková, & Baldrian, 2016; Vos, Wolf, Jennings, & Kowalchuk, 2013). Yet the controls across these scales vary considerably. For example, the edaphic properties stemming from soil
heterogeneity result in patchy distribution of nutrients, generating ‘hot spots’ of microbial activity and diversification (Rillig, Muller, & Lehmann, 2017).

Prior studies have implicated microorganisms as drivers of biogeochemical cycles, and cultivation and community omics have further expanded knowledge of taxonomic and functional diversity. Numerous field and laboratory incubation experiments have been established that encompass soils across biomes (e.g. Bond-Lamberty et al., 2016; Hultman et al., 2015; Y. Luo, Wan, Hui, & Wallace, 2001; Melillo et al., 2017; Noh et al., 2016; Tveit, Urich, & Svenning, 2014). These studies use a broad range of technical approaches to characterize microbial diversity and link this to ecosystem level process measurements. Research at the ecosystem scale includes broad measurements of nutrient and gaseous fluxes (e.g. Carey et al., 2016; Frey, Lee, Melillo, & Six, 2013; Jerry M Melillo et al., 2011; Tucker, Bell, Pendall, & Ogle, 2013), as well as biomass (e.g. S. D. Frey, Drijber, Smith, & Melillo, 2008; Spohn et al., 2016) and phospholipid fatty acid (PLFA) estimates of microbial community structure (Watzinger, 2015).

Recently, technical challenges that result from the expansive microbial diversity and spatial heterogeneity have been addressed using more refined approaches to studying the diversity of soil microorganisms. These approaches include cell-sorting coupled to high-throughput sequencing, microfluidics, and bioinformatic investigations of microorganisms in co-occurrence networks (Berry & Widder, 2014; Ofaim et al., 2017). Although studying microbial communities using these approaches elucidates diversity and activity at discrete scales, it does not reveal the emergent properties of soil ecosystems which become apparent through measurements of interactions within and across scales. Investigations of microbial assemblages across scales remain key in order to link microbial diversity and function to higher-level ecosystem processes.
In this perspective, we address different soil scales that should be considered when investigating microbial processes along with their controls, namely (1) soil cores/profile scale; (2) the soil aggregate scale including micro- and macroaggregates; and (3) the ‘microbial consortium’ scale at which single cells interact (Figure 1.1). We further discuss potential methods to investigate the controls of microbial community structure and functions across the different scales. We propose that the underexplored ‘microbial consortium’ is a key scale that can have significant influences on many larger scale biogeochemical processes. It is the aim of this perspective to encourage investigators to take more refined approaches in their studies of soil communities, particularly with a focus on the ‘microbial consortium’.

1.3 Advances in soil ecology at higher-level scales

1.3.1 ‘Soil core/profile scale’ - controls presented by plant roots, mycorrhizal fungi and soil fauna

The ‘soil core/profile’ scale has been extensively investigated across soil environments. The classical approach of collecting and homogenizing soil cores is found in countless studies (reviewed in Fierer, 2017); however the ‘soil core’ represents a rather artificial scale, reflecting the needs of the experimenters and available analytical methods. Nonetheless, this is the scale from which estimates of $10^9$ to $10^{10}$ microbial cells and 1,000 to 1,000,000 species per gram of soil have been derived (R. Amann & Rosselló-Móra, 2016; Gans et al., 2005; Jay T Lennon & Locey, 2016; Locey & Lennon, 2016a). This microbial diversity is hypothesized to be a control of ecosystem level processes even though the details of these processes occur at much smaller scales that are not accounted for due to homogenization of soil within cores (Bach, Williams, Hargreaves, Yang, & Hofmockel, 2018). Perhaps this is one of the reasons that the exact influence of microbial diversity and community structure on these processes has generally remained elusive (Prosser et al., 2007).
Studies at the ‘soil core/profile’ scale indicate that the mineral properties of soils, as well as plant roots and mycorrhizal associations are the main controls at this level. Plants and mycorrhizal fungi physically modify soil structure via fragmentation and formation of aggregates (Angers & Caron, 1998) which stabilize nutrients and generate microhabitats for bacteria (Bach et al., 2018; Six & Paustian, 2014). Plant roots introduce pulses of nutrients to the soil in the form of labile exudates, a process termed the ‘rhizosphere priming effect’ (Bengtson, Barker, & Grayston, 2012; Cheng, 2009). In addition to contributing to soil aggregation and stimulating microbial decomposition of recalcitrant soil organic matter, this ‘priming effect’ also promotes root colonization of mycorrhizal fungi through the release of strigolactones and labile carbon exudates (Bengtson et al., 2012; Bonfante & Anca, 2009; Bouwmeester, Roux, Antonio Lopez-Raez, & Card, 2007; Cheng, 2009; Kuzyakov, 2010). Mycorrhizal fungi uptake, release, and transport distant nutrient sources (Gorka et al., 2019), and as such they also serve as ‘hyphal highways’ for bacterial dispersal and access to nutrients throughout the soil, as well as promoting microenvironment connectivity (Roux et al., 2017; Simon et al., 2015; Warmink, Nazir, Corten, & van Elsas, 2011).

Other soil biota, including micro-, meso- and macrofauna, act as biological controls on the ‘soil core/profile’ scale by modifying soil structure and maintaining bacterial population sizes through trophic interactions (Geisen et al., 2018; Siddiky, Kohler, Cosme, & Rillig, 2012). Microfauna include protists and nematodes (Bardgett & Van Der Putten, 2014; Stork & Eggleton, 1992) which are found in cell densities of $10^4$ to $10^7$ individuals per m$^2$ of soil (Bardgett & Van Der Putten, 2014; Vinciguerra, 2009). These microfauna interact with bacteria and fungi through predation, thereby controlling population densities and making nutrients available to other trophic levels (Bardgett & Van Der Putten, 2014; Geisen et al., 2018; Vinciguerra, 2009). Mesofauna include invertebrate species such as mites (Acari), springtails (Collembola) and
millipedes, that strongly influence the fungal populations through mycophagy (Stork & Eggleton, 1992) and facilitate litter fragmentation (Lavelle, 1997). Macrofauna, including earthworms, termites and ants, are commonly referred to as ‘ecosystem engineers’ as they mechanically alter the physical structure and chemical composition of their soil surroundings (Stork & Eggleton, 1992). Together, soil fauna control bacterial population dynamics, re-distribute and prime substrates for further decomposition, and generate ‘hot-spots’ of microbial activity across soil ecosystems (Kuzyakov & Blagodatskaya, 2015).

1.3.2 Methodological approaches and challenges at the ‘soil core/profile’ scale

Across the ‘soil core/profile’ scale, numerous techniques have been applied to understand the microbial community. Historically studied through laboratory cultivation, only a minority (0.5% to 1% of the total diversity) of soil bacteria can grow in the lab (e.g. Fierer, Bradford, & Jackson, 2007; Lloyd, Steen, Ladau, Yin, & Crosby, 2018; Rappé & Giovannoni, 2003; Torsvik & Øvreås, 2002) Alternative approaches to strain isolation aim to preserve some of the cross-feeding interactions between bacteria by focusing on communities (D’Souza et al., 2018) These include the isolation chip (ichip; Nichols et al., 2010) and microfluidic-based cultivation (Burmeister et al., 2019), both of which maintain conditions more similar to in situ soil environments.

Challenges associated with laboratory cultivation have been circumvented using cultivation-independent sequencing approaches. Amplicon surveys have been applied in numerous studies at the ‘soil core scale’, revealing shifts in soil community structure under changing environmental conditions (e.g. Bond-Lamberty et al., 2016; Lladó, Žifčáková, Větrovský, Eichlerová, & Baldrian, 2016). However, PCR-based approaches are dependent on primers, which may not adequately amplify all organisms in a community, resulting in bias regarding organismal
abundance (Eloe-Fadrosh, Ivanova, Woyke, & Kyripides, 2016). High-throughput sequencing technologies and improved nucleic acid extraction kits have made community sequencing more accessible and efficient (Van Dijk, Lè Ne Auger, Jaszczyszyn, & Thermes, 2014). Metagenomic and metatranscriptomic approaches target nucleic acids (DNA or RNA) from the total community, unraveling the microbial community structure and gene expression without primer amplification bias (Eloe-Fadrosh et al., 2016). Insights into soil community structure and metabolic potential have been achieved using these techniques; however, significant challenges persist in high-throughput sequence data at the ‘soil core/profile’ scale including lack of available bacterial reference genome sequences and poor bioinformatic sequence assembly (Figure 1.2).

Stable isotope probing (SIP) experiments are more targeted approaches to accessing a subset of the microbial community (e.g. Angel et al., 2018; Dumont & Murrell, 2005; Eichorst & Kuske, 2012; Radajewski, McDonald, & Murrell, 2003). In SIP approaches, grams of soil are incubated with an isotopically labeled substrate (e.g. using $^{15}$N-, or $^{13}$C-labeled compounds) or water (e.g. H$_2^{18}$O, D$_2$O). These isotopes become incorporated into microbial biomarkers (e.g. DNA, RNA, PLFA or proteins), allowing investigators to distinguish active from inactive microorganisms by analyzing stable isotope enrichment of the respective biomarker (Murrell & Whiteley, 2011). When coupled with genomics-based methods, investigators can target specific functional groups within the soil community and elucidate the influence of microorganisms on a given process (e.g. Dumont & Murrell, 2005; Murrell & Whiteley, 2011; Radajewski, Ineson, Parekh, & Murrell, 2000). Recently, a higher-throughput method has been developed, that combines high-density phylogenetic microarrays (‘chips’) with SIP (CHIP-SIP; Mayali et al., 2012). This highly sensitive approach enables quantification of stable isotope incorporation (such as $^{13}$C or $^{15}$N) using secondary ion mass spectrometry (NanoSIMS), coupled to taxonomic identification of ribosomal
RNA using microarrays (Mayali et al., 2012). This approach has been applied to estuarine (Mayali et al., 2012) and marine samples (Mayali, Stewart, Mabery, & Weber, 2016; Mayali, Weber, Mabery, & Pett-Ridge, 2014), but still awaits its application on soil samples.

1.3.3 The millimeter-scale – including soil micro- and macroaggregates

Soil aggregate surfaces and the pore spaces between aggregates are considered ‘hot spots’ of microbial activity (Kuzyakov & Blagodatskaya, 2015), and therefore comprise a highly relevant scale for investigating microbial processes. Soil aggregates are hierarchical constructs containing building blocks of different sizes: microaggregates are defined as <250 µm in diameter and comprise organo-mineral complexes <53 µm (also termed “silt and clay-sized aggregate fraction”), whereas macroaggregates are defined as >250 µm aggregates (also termed “sand-sized aggregate fraction”) (Six & Paustian, 2014; Totsche et al., 2018). Microbial process controls at the microaggregate scale include organo-mineral interactions, aggregate stability and pore size distribution (Jastrow, Miller, & Lussenhop, 1998). Process controls at the macroaggregate scale include soil micro-architecture, as well as fungal-bacterial and plant-microbe interactions (Poirier, Angers, & Whalen, 2014) (Figure 1.1).

Progress in understanding the biological component of soil microenvironments at the aggregate scale, has remained slow relative to our understanding of the physical and biogeochemical properties of soils (Baveye et al., 2018). Studies at the soil aggregate scale have revealed that the abundance of specific microbial groups differs across aggregate size classes: for example, fungi are predominantly associated with coarse-scale fractions or macroaggregates, whereas bacteria and archaea are found to a greater extent in the smallest aggregate size fractions (Helfrich, Ludwig, Thoms, Gleixner, & Flessa, 2015; Hemkemeyer, Christensen, Martens, & Tebbe,
This distribution of microorganisms across aggregate size fractions can be attributed to the higher surface area of clay-sized particles in microaggregates compared to macroaggregates, which facilitate the colonization and growth of bacteria through extracellular polysaccharide matrix formation (Sessitsch, Weilharter, Gerzabek, Kirchmann, & Kandeler, 2001). In contrast, larger aggregates comprised of coarser particles provide habitat for fungal hyphae (Gupta, Bhandari, & Naushad, 2012; Kögel-Knabner et al., 2008). Smaller aggregate size fractions may provide refuge for bacteria from grazers, reduce competition for substrates, and serve as a mechanism for the physical protection of carbon in soils (Bach & Hofmockel, 2016).

Soil aggregates are considered hotspots of microbial activity. Some studies found that carbon-degrading enzymes showed higher activity in microaggregates compared to macroaggregates, whereas nitrogen-degrading enzymes (N-acetylglucosaminidase) were higher in macroaggregates (Bach & Hofmockel, 2016; Nie, Pendall, Bell, & Wallenstein, 2014). Recent investigations illustrated that the ratio of bacteria to fungi and of Gram-positive to Gram-negative bacteria strongly correlated with aggregate size and turnover rates of soil organic carbon (SOC) in microaggregates, whereas the presence of bacterivorous nematodes promoted the turnover of SOC in macroaggregates (Jiang et al., 2018). Yet in other investigations, no significant differences in diversity were observed between the microbial communities of aggregates grouped into classes of high and low beta-glucosidase activity, suggesting no relationship between microbial community structure and ecosystem function (Bailey, Fansler, Stegen, & McCue, 2013).

More recently soil aggregates have been proposed to be ‘incubators’ for microbial evolution (Rillig et al., 2017). At the time of formation of an aggregate, a select collection of microorganisms is present, and during the subsequent steps of aggregate stabilization these
communities can evolve and are later released into the soil upon disintegration of the aggregates (Rillig et al., 2017). This suggests microorganisms inhabiting soil aggregates are responding not only to changes in the surrounding soil aggregate environment, but also to interactions with their microbial neighbors. Given the varying effect(s) of aggregate class size on processes, perhaps the variation is due to the composition or genetic potential of the evolving soil aggregate consortium. Taken together, it is plausible that this consortium could not only influence processes within soil aggregates, but also bulk scale processes upon aggregate disintegration illustrating the power of microbial interactions on function across scales.

1.3.4 Methodological approaches and challenges at the aggregate scale

There have been recent advances in individual-based modeling of microorganisms in soil aggregates to predict biogeochemical fluxes based on anaerobic versus aerobic conditions (Ebrahimi & Or, 2016). However, this approach is facing challenges in incorporating additional factors (e.g. the influence of soil depth on soil organic content and quality, microbial community composition, preferential flow and root distribution and activity) into these models, and in scaling up to the ecosystem scale. Other methods are available to explore the interactions and/or structure of soil aggregates. For example, the pore structure of microaggregates can be characterized down to 1 µm resolution using micro-computed tomography (µCT) (Vos et al., 2013). This approach can be coupled to microfluidics, targeted investigations at the micrometer-scale via the manipulation and exchange of fluids, (Aleklett et al., 2018; Stanley, Grossmann, Casadevall i Solvas, & deMello, 2016; Wessel, Hmelo, Parsek, & Whiteley, 2013)(Aleklett et al., 2018; Stanley, Grossmann, Casadevall i Solvas, & deMello, 2016; Wessel, Hmelo, Parsek, & Whiteley, 2013) for downstream single-cell analysis (discussed further in later section).
The use of fluorogenic (e.g. methylumbelliferol, MUF) or fluorescence-resonance energy transfer (FRET) substrates may help finding the locations of extracellular enzyme activity in soil aggregates. This approach combines fluorescence and confocal microscopy at microscales (Kovarik & Allbrutton, 2011). Although challenging, the three-dimensional structure of soil aggregates can be preserved with resin embedding and sectioning for individual cell visualization at the micrometer scale (Vos et al., 2013). When this approach is combined with the use of stable isotope-labeled substrates, single-cell visualization and isotope composition analysis (such as FISH and NanoSIMS – more detailed discussion in subsequent section) for targeted process investigations, it holds tremendous potential for understanding the distribution of microorganisms, their function and interaction at this scale.

1.4 The 'microbial consortium': The key scale for investigating microbially-mediated processes in soil

In the environment, microorganisms live within social networks with other microorganisms, termed ‘microbial consortia’. Microorganisms in consortia are distributed within direct contact distance, promoting interaction with organisms of the same species, as well as other species and the surrounding environment. Examples of ‘microbial consortia’ include a population of a bacterium in a batch culture or a community of interacting microorganisms in a soil aggregate. ‘Microbial consortia’ can be comprised of any number of combinations of bacteria, archaea and single-cell eukaryotes all of which interact through the production and exchange of metabolites, competition for substrates and/or suppression by antagonistic compounds (D'Souza et al., 2018; Pande & Kost, 2017).

There are two main mechanisms of microbial interactions in soil: (i) interspecies metabolic cross-feeding that occurs ubiquitously in microbial environments (D’Souza et al., 2018; Hug & Co,
2018; Zelezniak et al., 2015), and (ii) chemical signaling or microbial cell-to-cell communication (DeAngelis, 2016). Metabolic cross-feeding, i.e., if one microorganism uses metabolites produced by another as energy or nutrient sources, allows for diverse ecological interactions including competition, mutualism and cheating (D’Souza et al., 2018; Estrela, Trisos, & Brown, 2012; Morris, 2015). Interspecific bacterial communication can stem from the release of quorum sensing compounds, such as N-acyl-homoserine lactones in Gram-negative microorganisms or Y-butyrolactones in Gram-positive microorganisms (Teplitski, Robinson, & Bauer, 2000). The number of interactions between individuals as well as the distance of the interactions is limited by the ‘calling distance’ of these compounds. Bacteria cells are estimated to interact within 12.5 to 20 µm distances in the soil, enabling interactions with an average of 120 individuals and <100 other species (Raynaud & Nunan, 2014).

Although soil microorganisms are frequently referred to as being functionally redundant or generalists (Allison & Martiny, 2008), we hypothesize that members of a microbial consortium are instead functional specialists, investing energy in a specific task for economical use of substrates. In support of this working hypothesis, studies have found that the type and amount of nutrients are heterogeneously distributed at the nanoscale (Lehmann et al., 2008), giving rise to resource specialization potential even at the smallest scale (Vos et al., 2013). Instead of microorganisms competing for the same resource, the presence of nutrients in different forms generates additional niches leading to a diversification of resource specialists (MacLean, 2005). As such, this strategic resource allocation reduces or abolishes competition for extracellular resources and can minimize the accumulation of intermediates, which could generate negative feedback effects (Lindemann et al., 2016).
The idea of specialists within a microbial consortium reflects the emergent properties of an ecological system as described by (Nielsen & Müller, 2000), where subunits (here the microbial cells) cannot exist in isolation, and the interactions of the subunits emerge at a higher level as new properties (here functions). Microbial consortia for a given process have been documented previously in other systems to increase the efficiency and/or rate of a targeted process (Brenner, You, & Arnold, 2008). As such, we propose that the consortium of interacting microorganisms is critical and thus the key scale, having significant influences on community metabolism, and thereby on soil processes and ecosystem functions. Our proposed ‘microbial consortium’ is the basis of microbial assemblages at higher spatial scales, including soil aggregates, and communities associated with plant roots and mycorrhizal fungi.

Although investigations at the ‘microbial consortium’ scale are still in their infancy due to technical limitations, the results could be groundbreaking for a better understanding of these microbial interactions and their subsequent effect on processes. We propose that these micro-scale microbial interactions can be explored by (1) microfluidics, such as the recently described ‘soil-on-a-chip’, (2) sequencing mini-metagenomes, (3) single-cell level investigations, and (4) computational investigations, such as co-occurrence networks and modeling. Taken together, these approaches provide insight into the genetic and metabolic diversity of microbial communities with finer resolution than soil community studies at the soil core and aggregate scales.

1.4.1 Simplifying soils in the lab with microfluidics and ‘soil-on-a-chip’

The complexity of microbial interactions within ‘microbial consortia’ warrants the application of more refined approaches to studying the microbial members and their activity in the native environment. Much of what is known about microbial diversity has come from high-
throughput sequencing approaches at the ‘soil core/profile’ scale, which ignore potential interactions between organisms and their environment. Microfluidic systems including ‘soil-on-a-chip’ were developed to investigate soil microorganisms and their interactions in situ (Aleklett et al., 2018; Stanley et al., 2016). Microfluidics permits precise spatiotemporal control and high-resolution imaging to examine interactions from consortium-scale down to the single-cell level (Stanley et al., 2016).

The use of microfluidics in soils has recently been proposed to address fundamental questions in the field of microbial ecology, namely in the understanding of physical heterogeneity, chemical gradients, microbial interactions, rhizosphere interactions and cultivating the ‘yet-to-be-cultivated’ (Aleklett et al., 2018; Massalha, Korenblum, Malitsky, Shapiro, & Aharoni, 2017; Nagy et al., 2018). The construction of artificial soil habitats enables investigation of ecological and evolutionary processes on a chip (Nagy et al., 2018). For example, the spatial heterogeneity of soil and the resulting patchy distribution of nutrients can be simulated using this approach (Aleklett et al., 2018; Nagy et al., 2018).

1.4.2 Mini-metagenomics for capturing diversity at the 'microbial consortium' scale

The combination of genomics and cell sorting provides excellent resolution for investigating the activity of cells in microbial consortia, and may facilitate investigation of microbial intra- and interspecies interactions that play important roles at higher-level scales. Cell-sorting and sequencing has given rise to single-cell genomics, a more recent addition to the toolkit for studying the genetic makeup of uncultivated microbial cells from environmental samples (Blainey, 2013a; Stepanauskas, 2012; Woyke, Doud, & Schulz, 2017). In this approach individual cells are isolated from consortia using fluorescence-activated cell sorting (FACS) or microfluidic sorting, and their
genomes are subjected to whole genome amplification after cell lysis, followed by sequencing to generate single amplified genomes (SAGs) (Stepanauskas et al., 2017; Woyke, Doud, & Schulz, 2017b). While this approach provides genomic information with strain-level resolution, the resulting SAG assemblies are often highly fragmented and incomplete, and the overall process is prone to biases and contamination. These challenges are further conflated by the high genetic diversity in soils in conjunction with the low-throughput of single cell workflows. Furthermore, these single cell approaches may not be adequate to capture the entire collection of organisms at the ‘microbial consortium’ scale.

One alternative approach, termed ‘mini-metagenomics’, sequences a small pool of sorted cells (<100 cells), thus allowing an investigator to characterize the genetic potential and associated taxonomy of microorganisms from heterogeneous environments. Mini-metagenomics was first used in combination with shotgun sequencing to study hospital sink biofilm communities, which are comprised of highly diverse and potentially pathogenic organisms (McLean et al., 2013a). More recently, this approach was used in combination with microfluidics technology to explore microbial consortia in hot spring samples (Berghuis et al., 2019; Yu et al., 2017). For the first time, this mini-metagenomic approach was applied to soil from the Harvard Forest Long-Term Ecological Research (LTER) site and facilitated the expansion of bacterial, archaeal, and giant virus diversity (Alteio et al., in revision; Schulz et al., 2018). Sequencing a smaller pool of organisms within the highly diverse soil microbial community may result in better coverage of genomes, and the potential for improved genome sequence assembly (Figure 1.2; Yu et al., 2017). However, this mini-metagenomic approach does not preserve the overall structure and chemical composition of soil environments, thereby still representing a selective view of the microbial consortium.
The mini-metagenome approach is particularly attractive for studying microbial “dark matter” (Rinke et al., 2013), organisms which have not been isolated using culture. Improved sequence assembly capabilities may provide opportunities to capture lower abundance community members that are otherwise overlooked in bulk metagenome studies. In addition to contributing potentially novel diversity to genome sequence databases, reduction of soil community diversity and complexity by sampling subpopulations could refine genome-level resolution to the species- or even strain-levels (Yu et al., 2017). Accessing microbial diversity at this high resolution may enable quantification of selection in microbial populations using comparative genomics, resulting in improved understanding of how microorganisms change in response to environmental pressures.

Although mini-metagenomics does not preserve the spatial distribution or in situ soil conditions as in a microfluidic system, pooled cell sorting may enable investigation of close cell-cell interactions and symbioses. For example, a recent study using fluorescence activated cell sorting resulted in the co-sorting of Nanoarchaeota and their putative hosts, and enabled the identification of genes involved in maintaining this symbiotic relationship (Jarett et al., 2018). Similarly, cell sorting approaches of single- or pooled cells could decrease the complexity of entire ‘microbial consortia’ in soils, and presents opportunities for a closer look into how cells metabolize, interact and diversify in the soil environment.

1.4.3 Single-cell investigations of the ‘microbial consortium’

‘Microbial consortia’ and interactions therein can also be investigated in situ at the single-cell level with the use of stable isotopes (such as $^{13}$C, $^{15}$N- or $^{18}$O-labeled substrates) coupled with high-resolution techniques, such as secondary ion mass spectrometry (NanoSIMS) and Raman microspectroscopy. Similar to SIP approaches applied at the ‘soil core/profile’ scale, single-cell SIP
studies enable tracing of isotopes from labelled substrates into biomarkers within active cells. Secondary Ion Mass Spectrometry (SIMS) is a surface analysis technique (Benninghoven, Rüdenauer, & Werner, 1987) that can be used to detect and quantify the abundance of rare isotopes ($^{13}$C, $^{15}$N or $^{18}$O) in metabolically active cells relative to that of the common isotope ($^{12}$C, $^{14}$N or $^{16}$O). Embedding samples in a resin and sectioning prior to SIMS can preserve information about the soil sample and associated ‘microbial consortia’ (Vidal et al., 2018). Another approach to SIP studies is Raman microspectroscopy, a rapid, non-destructive vibrational spectroscopic method which provides information on the molecular composition of a sample (S. Eichorst & Woebken, 2014; W. E. Huang et al., 2007; Jarvis & Goodacre, 2004). The incorporation of stable isotopes in the cellular components gives rise to changes in Raman spectral signatures, allowing investigators to detect and quantify isotope tracer incorporation (S. Eichorst & Woebken, 2014), as shown for naphthalene degradation (W. E. Huang et al., 2007), carbon dioxide fixation (M. Li et al., 2012) and general microbial activity (Berry et al., 2015; S. A. Eichorst et al., 2015).

Fluorescence approaches also allow investigation of microbial consortia. Translationally active microorganisms have been identified in aquatic systems using an approach called bioorthogonal non-canonical amino acid tagging (BONCAT). This method involves the incubation of samples with a methionine amino acid analog. After performing click-chemistry, these non-canonical amino acids become fluorescently labeled and can be sorted using FACS. Although this approach has not yet been applied in soils, it will enable detection of metabolically active cells (Hatzenpichler et al., 2016). The combination of NanoSIMS with cell identification via fluorescence in situ hybridization (FISH) (Rudolf Amann & Fuchs, 2008) has proven very useful in linking the identity of a microorganism with its function in environmental samples (Clode et al., 2009; Musat
et al., 2008; Orphan, House, Hinrichs, McKeegan, & DeLong, 2001; Ploug et al., 2011; Woebken et al., 2012).

FISH-NanoSIMS has been increasingly applied to aquatic samples, and with recently developed sample processing steps can also be applied to soil samples (S. A. Eichorst et al., 2015). The combined application of these single-cell techniques to soil samples allows one to not only maintain the sample’s structural information, but also reveal the identity of microorganisms, their activities and interactions within a ‘microbial consortium’. These approaches, although applied at the scale of the single cell within a consortium, can provide valuable information about the taxonomic identity and metabolic activity of organisms within the more complex ‘microbial consortium’. Deconstructing these complex interaction networks will enable improved understanding of microbial diversity, metabolism and interactions with higher resolution than approaches at the ‘soil core’ or aggregate scales, and presents opportunities for scaling from the single cell to higher-level ecosystem process.

1.4.4 Computational approaches to the ‘microbial consortium’

Exploration of microbial interactions have been conducted using bioinformatic analyses of amplicon and metagenomic datasets (Berry & Widder, 2014), coined co-occurrence networks, although rarely at the ‘microbial consortium’ level. These computational analyses provide a comprehensive, albeit indirect, assessment of potential microbial interactions within a community (Barberán, Bates, Casamayor, & Fierer, 2012). Furthermore, microbial co-occurrence patterns can also be applied at higher scales of biological organization, for instance within and across ecosystems, thereby capturing emergent properties of ecological systems (Baveye et al., 2018; Williams, Howe, & Hofmockel, 2014).
More recently, the use of spatially-explicit, individual-based computer models has emerged to better elucidate spatial interactions between individuals (O’Donnell, Young, Rushton, Shirley, & Crawford, 2007). These models incorporate aspects of soil architecture, chemistry and physics to dynamically determine the distribution and interactions between microorganisms and the environment (Baveye et al., 2018; O’Donnell et al., 2007). In turn, models can couple these interactions to metabolic functions, thereby describing the biogeochemical functioning of communities under different environmental conditions. For example, these computer-based models have shown that self-organizing and self-regulating features can emerge from these first-principle interactions in simulations of microbial decomposer communities (Allison, 2005; Kaiser, Franklin, Dieckmann, & Richter, 2014; Kaiser et al., 2015; Momeni, Waite, & Shou, 2013). Moreover, changing environmental conditions has been shown to be a potential control for ecosystem-scale processes (Allison, 2005; Kaiser et al., 2014, 2015). The ability of computer models to accurately predict community function and dynamics is ultimately dependent on their parameterization. Given this, improved model-experiment integration is needed in the field of microbial ecology when applying these models to ensure a more robust view on an investigated system (Baveye et al., 2018; Widder et al., 2016).

1.5 Scaling up towards the ecosystem scale

Integrating across scales to the ecosystem level is challenging, as ecosystem scale processes are based on a complex network of metabolic processes and involve emerging scale-dependent controls. To facilitate the link between the microbial scale and ecosystem functions, processes have been classified as ‘broad processes’ and ‘narrow processes’. Broad processes encompass activities carried out by diverse microbial communities, with many species contributing to the same process (e.g. microbial respiration or organic nitrogen mineralization).
In contrast, narrow processes are more defined in their genetic and physiological specificity (e.g. nitrification or CH$_4$ oxidation). Yet it remains unclear as to whether the controls of ecosystem level processes - for both broad and narrow processes - are the same at the micro- and macroscale.

Measurements of ‘broad processes’ are relatively diffuse in regards to genetic and enzymatic targets, making high-resolution investigations challenging. Broad process measurements of microbial activity have been recorded using soil respiration chambers (Bowden, Davidson, Savage, Arabia, & Steudler, 2004), organic nitrogen mineralization (S. D. Frey et al., 2008; Pisani, Frey, Simpson, & Simpson, 2015; Schmidt et al., 2011) and enzymatic assays (Widmer, Flieûbach, Laczko Â, Schulze-Aurich, & Zeyer, 2000) in complement to broad characterization of microbial community structure, including biomass and phospholipid fatty acid analysis (PLFAs; Watzinger, 2015). These approaches provide insight into processes performed by many if not most microorganisms, but are unable to assign function to specific organisms (Nesme et al., 2016).

In contrast, narrow processes (such as nitrification or CH$_4$ oxidation) are more defined in regards of their genetic and/or enzymatic targets and associated physiology (Schimel & Schaeffer, 2012). Therefore, understanding the influence of controls (such as labile C, Mo and V availability for N$_2$ fixation) is hypothesized to be more direct, scalable and strongly linked to the active microbial consortia mediating these processes (Bakken, Bergaust, Liu, & Frostegård, 2012; Isobe, Koba, Otsuka, & Senoo, 2011; Schimel & Gulledge, 1998).

Challenges exist in integrating across scales primarily due to microbe-microbe interactions. Although processes such as microbial respiration is based on the activity of individual microorganisms, microbial community metabolism is not simply the sum of these single microbial
activities. For example, chemical interactions among cells and populations can have differing outcomes (Ponomarova & Patil, 2015; Stubbendieck, Vargas-Bautista, & Straight, 2016) and in addition several unrelated microbial groups can (simultaneously) influence the same process (Röling, van Breukelen, Bruggeman, & Westerhoff, 2007). Given these non-predictable variables, determining the controls at one scale and then up or down scaling remains a major challenge and may not be possible.

To better understand the controls and determine if they are similar at both the micro- and macroscale, multi-disciplinary investigations with techniques and tools stemming from both the macroscale (such as enzymatic assays and/or measurements of CO₂ or NH₄⁺ production), combined with microbial community structure (via amplicon sequencing) and functional diversity (via metagenome sequencing) investigations, and microscale approaches (such as microfluidics, FISH, NanoSIMS, Raman microspectrosopy) are warranted. Typically, specific processes are investigated at one scale with its commonly applied set of techniques, which makes it challenging to extrapolate to the opposing scale where other processes are measured with other techniques. This is further exacerbated by challenges the -omics field is facing, since much of the genomic, proteomic and metabolic data are only available for select model organisms across the archaea, bacteria and microbial eukaryotes (e.g. (Choi et al., 2017), and typically >50% of their genomic information lacks functional annotation (Delmont et al., 2015). More refined methodologies across scales are necessary to improve understanding of microbial community composition, metabolism, interactions, as well as how these factors are controlled and drive processes across scales in soil ecosystems.
1.6 Synthesis

Scales are an important aspect of soil microbial processes, which ultimately affect overall ecosystem function. In this perspective we sought to raise awareness about spatial scales in soils and reinforce the need to make more refined soil investigations across scales - with particular emphasis on the soil ‘microbial consortium’ scale. We propose that the ‘microbial consortium’ is the key scale that drives processes at higher spatial scales (from aggregate to ecosystem), demanding further consideration and investigation. Mini-metagenomics, microfluidics, stable isotope single-cell level investigations along with computational modeling and co-occurrence networks have the potential to provide the resolution to investigate the active participants in the microbial consortium. In light of our new view of the tree of life (Castelle & Banfield, 2018; Hug et al., 2016) our knowledge regarding the extent of diversity is expanding. The complex properties of soil in addition to the expanse of unexplored microbial diversity contribute to the lack of understanding of microbial communities, and how processes at the single cell scale are related to processes at the ecosystem level. Study of soil communities and processes at the scale of the microbial consortium allow for more targeted measurements of microbial metabolism and interactions that result in feedbacks at higher levels. We postulate that the ‘microbial consortium’ scale could change our understanding of microbial community structure and function in soil(s), resulting in stronger linkages between microorganisms and their roles as ecosystem process drivers.
Figure 1.1. Conceptual image depicting the different scales at which processes occur in soil and exemplifying major process controls at each scale.
Figure 1.2. Diversity contributes to the complexity of soil ecosystems. Soil provides habitats for organisms including virophages, giant viruses, bacteria, archaea, protists, fungi and arthropods. However, this diversity presents challenges in the study of soil ecology. Approaches have been taken at higher level scales (including the soil core and millimeter scales) to capture the diversity and physiology to decrease complexity, including microbial biomass, phospholipid fatty acid analysis, and laboratory culture of microorganisms. To further dissect the diversity and potential function of organisms within microbial consortia, molecular approaches have been applied across scales including bulk metagenomics, mini-metagenomics, and single-cell genomics. These approaches reduce complexity of consortia in situ and enable improved genome assembly and binning.
CHAPTER 2
COMPLEMENTARY METAGENOMIC APPROACHES REVEAL BACTERIAL AND ARCHAEAL DIVERSITY IN FOREST SOIL

2.1 Abstract

Soil ecosystems harbor diverse microorganisms yet remain poorly characterized as neither single-cell sequencing nor whole community sequencing offers a complete picture of these spatially complex communities. Thus, the genetic and metabolic potential of this ‘uncultivated majority’ remains underexplored. To circumvent these challenges, we applied a cell sorting based mini-metagenomics approach and compared the results to bulk metagenomics. Informatic binning of these data produced 200 mini-metagenome assembled genomes (sorted-MAGs) and 29 bulk metagenome assembled genomes (MAGs). These sorted and bulk MAGs increased known phylogenetic diversity of soil taxa and increased total soil tree branch length by 7.2%. Additionally, sorted-MAGs expanded the rare biosphere not captured through MAGs from bulk sequences, exemplified through phylogenetic and functional analyses of the phylum Bacteroidetes. Analysis of 66 Bacteroidetes sorted-MAGs showed conserved patterns of carbon metabolism across four clades. These results indicate that mini-metagenomics enables genome-resolved investigation of predicted metabolism and demonstrates the utility of combining metagenomics methods to tap into the diversity of heterogeneous microbial assemblages.

2.2 Introduction

Soil is considered one of the most biologically diverse ecosystem types, yet much of its microbial diversity remains poorly characterized (e.g. O’Donnell et al., 2007). Each gram of soil is estimated to harbor 1,000 to 1,000,000 different bacterial species (e.g. McLean et al., 2013). Investigating soil microorganisms in situ is challenging due to the heterogeneous nature of the soil environment (e.g. Yu et al., 2017). As a result, terrestrial habitats remain immense reservoirs of untapped genetic and metabolic diversity (Nesme et al., 2014; Torsvik & Øvreås, 2002) encoded within microbial communities that drive important ecosystem-level processes, including nitrogen cycling and carbon dioxide flux (Delgado-Baquerizo et al., 2016; Graham et al., 2016; Lladó, López-Mondéjar, & Baldrian, 2017). Soils are regarded as critical for global health, as they contain 3000 Pg of carbon, with the potential to act as a positive feedback to further climatic shift (Nesme et al., 2014; Torsvik & Øvreås, 2002). It is therefore essential to characterize soil microbial diversity to better understand ecosystem function and resilience in the face of rapid environmental change.

Historically, microbial diversity has been studied using laboratory cultivation techniques (Nesme et al., 2014; Torsvik & Øvreås, 2002) with only a minute fraction of estimated bacterial diversity being successfully cultivated (Delgado-Baquerizo et al., 2016; Graham et al., 2016; Lladó, López-Mondéjar, & Baldrian, 2017). Substantial efforts are being made to develop innovative cultivation techniques, including the ichip and droplet-based sorting coupled with laboratory cultivation (Hicks Pries, Castanha, Porras, & Torn, 2017). Yet, these approaches have only contributed to the expansion of novel families, whereas microbial taxa are predicted to be more phylogenetically divergent, belonging to potentially novel phyla (Overmann, Abt, & Sikorski, 2017; Pham & Kim, 2012). Thus, challenges associated with direct study of soil microorganisms have yielded a large knowledge gap regarding terrestrial microbial diversity. In addition to cultivation
limitations, there is a lack of representative reference genomes for soil microbes (Staley & Konopka, 1985; Urich et al., 2008). From the publicly available Integrated Microbial Genomics (IMG/M) database, we were able to curate a collection of 3,024 isolate genomes, single amplified genomes (SAGs) and metagenome assembled genomes (MAGs) from previous soil studies (Nichols et al., 2010; Overmann et al., 2017). However, with estimates of diversity on the order of millions of species per gram (Lloyd et al., 2018), these references represent only a small percentage of soil microbes.

High-throughput sequencing technologies combined with novel metagenome binning algorithms (Choi et al., 2017) enable genome-resolved metagenomics and have greatly expanded the availability of reference genomes from uncultured taxa by circumventing challenges associated with cultivation (Chen et al., 2016). The more recent applications of directly sequencing DNA from soil microbial communities allows one to obtain a broader perspective on the taxonomic and functional potential of soil microorganisms. However, metagenomics in highly diverse environments may capture only the most abundant, and therefore best-assembling representatives from the total community (Urich et al., 2008), and population heterogeneity can hamper the assembly efficiency, even of abundant microorganisms (Albertsen et al., 2013; Wrighton et al., 2012).

Traditional MAGs combine genomes from similar organisms within populations (Castelle & Banfield, 2018; Hug et al., 2016; Nesme et al., 2016). Depending on the binning parameters used, MAGs may collapse contigs from a highly diverse sample into a single genome, further complicating data interpretation. Soils are typically dominated by a small set of highly abundant taxa (Delmont et al., 2015; Nayfach & Pollard, 2016; Parks et al., 2017), and therefore the rare biosphere may be overlooked in metagenomic studies despite playing an important role in soil
biogeochemical processes (Sczyrba et al., 2017). Lastly, bulk metagenomics can also include extracellular DNA from dead microorganisms, which may be abundant in the environment. This exogenous DNA has the potential to inflate estimates of diversity and genomic potential (Sczyrba et al., 2017), and further reduce our ability to assemble sequences from rare taxa. Decoupling intracellular and exogenous DNA during sequencing may provide a more accurate estimate of microbial diversity (Delgado-Baquerizo et al., 2016).

Challenges associated with bulk metagenomics may be mitigated by reducing community complexity. The most extreme example involves the application of fluorescence-activated cell sorting (FACS) for separating communities into single cells for single-cell genomics, which provides genomic information with strain-level resolution (Jousset et al., 2017). However, the resulting SAG assemblies are often highly fragmented and incomplete, and the overall process is prone to biases and contamination.

In order to circumvent some of the challenges associated with bulk metagenomics and single-cell genomics, we applied a pooled cell sorting approach, termed mini-metagenomics, on forest soils collected from the Barre Woods soil warming experiment at the Harvard Forest Long-Term Ecological Research (LTER) site. Prior to this study, mini-metagenomics of microorganisms had been applied only in aqueous environments, including hot springs, hospital sink biofilms and activated sludge (Carini et al., 2016; Jay T Lennon, Muscarella, Placella, & Lehmkuhl, 2017; Nagler, Insam, Pietramellara, & Ascher-Jenull, 2018). Mini-metagenomics has higher throughput than single-cell genomics, providing the opportunity to capture more diversity than is possible with single-cell sequencing. Mini-metagenomics may enable investigation of different components of the soil community in comparison to bulk metagenomics, including cells that can be dissociated from particles, and cells with varying lysis susceptibility. The use of two overlapping metagenomic
methods may allow us to capture a broader taxonomic diversity than if only one approach were applied on its own. Additionally, cell sorting using FACS requires cells to be intact in order to be sorted, thereby minimizing challenges introduced by extracellular DNA in bulk soil samples. Using mini-metagenomics to reduce the number of cells relative to bulk metagenomics may decrease the number of genomes collapsed into a single MAG (J T Lennon, Muscarella, Placella, Lehmkuhl, & Kellogg, 2018). Hence, we evaluated this method as a tool to complement bulk metagenomics in uncovering the ‘microbial dark matter’ in soil.

Here we demonstrate the utility of combining mini-metagenomics and bulk metagenomics for biological discovery in a highly diverse forest soil microbial community. Separation of intact cells from soil via FACS enabled mini-metagenomic sequencing, while bulk metagenomics provided total community context for benchmarking. Our approach generated 200 sorted-MAGs and 29 bulk metagenome MAGs of medium quality, expanding phylogenetic diversity of known soil clades. Our data suggest that the sorted-MAGs represent some of the diversity of previously un-sequenced organisms that are challenging to access using bulk approaches, offering insights into the functional potential of soil dark matter.

2.3 Results and Discussion

2.3.1 Improved assembly and binning from mini-metagenomes

Our application of mini-metagenomics combines microbial cell sorting and metagenome sequencing in order to divide a complex soil community into many smaller, less complex subsets. We performed FACS on pools of cells from four soil samples collected from the Barre Woods experimental warming plots at the Harvard Forest Long-Term Ecological Research (LTER) site. From each of the four samples we sequenced 90 replicate pools of 100 cells for a total of 359 mini-
metagenomes (one mini-metagenome failed quality control standards). In conjunction with mini-metagenomic sequencing, we performed bulk metagenomics on these four soils generating a total of 1.2 Gbp and 1.3 Gbp, respectively (Figure 2.1).

Binning of assembled contigs produced 1793 mini-metagenome assembled genomes (sorted-MAGs) and 275 bulk metagenome MAGs (Figure 2.2 and 2.3). Following CheckM quality assessment (Blainey, 2013b; Stepanauskas, 2012; Woyke et al., 2017b), 200 sorted-MAGs and 29 bulk MAGs surpassed a completeness threshold of ≥50% complete, ≤10% contamination and ≤10% strain heterogeneity. We considered MAGs with less than 50% completeness as ‘low quality’ based on MIMAG standards (Berghuis et al., 2019; McLean et al., 2013a; Yu et al., 2017), and excluded them from additional analyses (Figure 2.2 and Figure 2.3). Overall, quality filtering removed lower quality sorted-MAGs on the basis of completeness, whereas bulk MAGs were removed due to a higher degree of contamination. Assessment of MAG quality using CheckM showed an average percentage completeness of 81.5% in medium quality bulk metagenome MAGs (n=29) relative to 61.9% in the medium quality sorted-MAGs (n=200; p=3.29X10⁻⁷; Figure 2.2 and 2.4). When assessed for marker gene contamination, bulk metagenome MAGs revealed an average estimated level of contamination of 1.92%, similar to the 0.98% average contamination in the sorted-MAGs (p=0.01117; Figure 2.2 and Figure 2.4).

As one measure to compare mini-metagenomics and bulk metagenomics methods, we assessed GC content and found an average of 49.2% GC and 60.5% GC in sorted-MAGs and MAGs, respectively (Figure 2.2, Figure 2.5). Variation in GC content can be attributed to known biases in the single cell workflow such as susceptibility of cells to sorting and lysis (Yu et al., 2017), as well as amplification bias introduced during MDA (Parks, Imelfort, Skennerton, Hugenholtz, & Tyson, 2015). The cell isolation method used in mini-metagenomics reduces inflation of community
diversity as a result of exogenous DNA and enables sequencing of organisms that are typically attached to soil particles through the application of a mild surfactant. Additionally, the difference in DNA extraction procedures between mini-metagenomics and bulk metagenomics represents an opportunity to capture an expanded diversity of microorganisms, as each approach may access a different component of the community. Taken together, mini-metagenomics and bulk metagenomics generated a large number of quality MAGs that can be used as complementary datasets in genome-resolved studies to investigate broad microbial diversity.

2.3.2 Expansion of phylogenetic diversity

As one aim of our study was to provide reference genomes that represent soil microbiome diversity, we evaluated the contribution of both sorted-MAGs and bulk MAGs to phylogenetic diversity in the context of previously published genomes of soil bacteria and archaea. We inferred the phylogenetic relationships using concatenated marker genes from the 200 sorted-MAGs, the 29 bulk MAGs and 3,024 soil microbe reference genomes from the IMG/M (Figure 2.6) (Bowers et al., 2017). For this analysis, we clustered sequences at 95% average nucleotide identity (ANI) to estimate distinct species-level lineages, resulting in 170 sorted-MAGs, 25 bulk MAGs and 2,341 reference sequences from IMG/M (Figure 2.6 and Figure 2.7). This small decrease in the number of MAGs as a result of clustering indicates very little redundancy between previous MAGs and available reference sequences. Sorted and bulk MAGs from this study contributed genome diversity across numerous soil clades, including Alphaproteobacteria (16 sorted-MAGs, 2 bulk MAGs), Acidobacteria (11 sorted-MAGs, 14 bulk MAGs) and Planctomycetes (2 sorted-MAGs, 1 bulk MAG). Sorted and bulk MAGs also contributed diversity to less abundant soil taxa including TM6 (6 sorted-MAGs, 1 bulk MAG) and Betaproteobacteria (3 sorted-MAGs, 1 bulk MAG).
Comparison of MAGs recovered through mini and bulk metagenomics revealed a broad diversity of soil bacteria and archaea, and demonstrated the complementarity of these approaches for biological discovery. The sorted-MAGs expanded taxonomic diversity of abundant groups including *Bacteroidetes* (48), *Verrucomicrobia* (8), and taxa with lower abundances including *Thaumarchaeota* (4), *Omitrophica* (3), *Ignavibacteria* (2), *Melainabacteria* (1) and *Firestonebacteria* (1). Interestingly, numerous sorted-MAGs belonged to phyla typically comprised of pathogens and endosymbionts such as the *Chlamydiae* (31) and *Gammaproteobacteria* (30), specifically within *Legionellales*, as well as *TM6* (7) (Blainey, 2013b; Clingenpeel, Clum, Schwientek, Rinke, & Woyke, 2014) (Figure 2.3). The phyla identified by sorted-MAGs represented abundant taxa found in previous soil community studies (Stepanauskas et al., 2017) in addition to the rare biosphere, demonstrating the utility of mini-metagenomics for expanding diversity beyond abundant soil taxa (Figure 2.3). As for the bulk MAGs, some of these belonged to rare taxa not recovered through mini-metagenomics, including *WPS-2* (3), *Euryarchaeota* (1) and *Saccharibacteria* (1).

We assessed the phylogenetic diversity (PD), the total amount of branch length contributed by sequences of interest within a phylogenetic tree, from the sorted-MAGs to determine the contribution of this single study to known microbial diversity. Calculation of phylogenetic diversity revealed a 7.2% increase in total branch length contributed by the sorted-MAGs in relation to the soil reference sequences from IMG/M (Figure 2.3). Mini-metagenomes not only expanded phylogenetic diversity within clades of known soil bacteria and archaea, but also candidate phyla and low abundance taxa typically found in forest soils. More specifically, the sorted-MAGs increased the branch lengths of well-studied bacterial groups, including *Bacteroidetes* (33.6%) and *Alphaproteobacteria* (19.4%), along with groups notoriously recalcitrant to laboratory cultivation, such as *Verrucomicrobia* (62.1%), *Acidobacteria* (28.0%) and...
TM6 (11.3%) (Chen et al., 2016). Most notable was the PD increase in the *Chlamydiae* (72.5%), a taxonomic group which is typically overlooked in soil metagenomic studies due to their low abundance and likely dependence on eukaryotic host cells (Deeg, Zimmer, et al., 2018; Horn, 2008; Pagnier et al., 2015; Schulz et al., 2015). We hypothesize that the application of mild detergent and syringe filtration during sample processing may have lysed the microbial eukaryotes that serve as hosts for bacterial endosymbionts, making these bacteria more accessible for FACS. A similar phenomenon was suggested for the detection of 16 novel giant viruses from these same samples (DeAngelis et al., 2015; Delgado-Baquerizo et al., 2018; Fierer, 2017), as these viruses are most often associated with eukaryotic host cells (S. A. Eichorst et al., 2018; McLean et al., 2013a).

The sorted-MAGs demonstrated the potential for mini-metagenomics to increase our knowledge of diversity beyond what can be achieved using MAGs from bulk metagenome studies alone. The bulk MAGs contributed to the phylogenetic diversity of many of the same clades of soil bacteria as the sorted-MAGs, including *Acidobacteria* (10.5%), *TM6* (7.3%) and *Alphaproteobacteria* (2.6%). However, even in clades where more bulk-derived genomes were added relative to sorted-MAGs, such as in *Acidobacteria*, the sorted-MAGs were phylogenetically more diverse.

### 2.3.3 Caveats of mini-metagenomics

Although the mini-metagenomics approach produced a greater number of medium quality genome bins relative to bulk metagenomics, this approach is not without challenges. In comparison to bulk metagenomics, mini-metagenomics may be prohibitive as it involves equipment and expertise that may not be easily accessible. In addition to logistical obstacles, methodological challenges including cell isolation, differential membrane lysis efficiency, and GC-based genome amplification skew likely introduce bias during sample processing. This may be
reflected in our data, where organisms which are typically abundant in forest soils such as *Actinobacteria, Chloroflexi* and *Firmicutes* (Lagkouvardos et al., 2014), were present in low numbers using mini-metagenomics as compared to traditional bulk metagenomics (Figure 2.3). Though these taxa might have been missed due to the aforementioned biases, it is also possible that sequences from these organisms were not binned or were placed in a lower quality bin based on our filtering threshold. For example, bacteria in the phylum *Spirochaetes* were represented by 47 distinct sorted-MAGs; however, none of these passed quality filtering standards and were therefore excluded (Figure 2.3). An alternative DNA amplification method has been developed, termed WGA-X, which improves cell lysis and amplification of high GC-content organisms over MDA (Schulz et al., 2018). With this improved method of DNA amplification, more representative mini-metagenomic sampling might be possible. Additionally, the methods for sample processing prior to FACS may be modified to achieve a more targeted cell or particle fraction, thereby further expanding the utility of mini-metagenomics to detect dark matter microorganisms.

### 2.3.4 Representation of sorted-MAGs and MAGs Across Terrestrial Soil Metagenomes

To assess the representation of our newly generated soil reference genomes across other terrestrial ecosystems, we searched for protein coding sequences from our collection of sorted-MAGs and MAGs across publicly available soil metagenomes from 80 terrestrial metagenome studies. For this analysis, we dereplicated the 200 sorted-MAGs and 29 bulk MAGs from this study by clustering at 95% average nucleotide identity without reference sequences, resulting in 173 sorted-MAGs and 28 bulk MAGs as cluster representatives (Figure 2.4). We assessed these mini- and bulk MAGs in the context of broader terrestrial community studies by comparing them against 2,210 metagenomes from the 80 terrestrial studies using LAST (Aherfi, Colson, La Scola, & Raoult, 2016) (Figure 2.4). We defined highly represented sorted-MAGs and MAGs as those with at least
200 protein coding sequences with hits to metagenome samples at ≥95% amino acid identity over 70% alignment length (Delgado-Baquerizo et al., 2018).

Some of our sorted-MAGs and MAGs detected in previous metagenomic soil investigations were members of the phylum Acidobacteria (10 sorted-MAGs, and 15 MAGs; Figure 2.4). Five bulk MAGs in the phylum Proteobacteria were detected in metagenomes from forest, agricultural, arctic, grassland, and vadose zone soils, whereas two bulk MAGs in candidate division WPS-2 were detected in metagenomes from Harvard Forest and other forest soil metagenomes, as well as arctic and surface soils. Interestingly, one MAG in the Planctomycetes was detected only in metagenome sequences from the Harvard Forest, indicating that this may represent a unique MAG which has not been found in previous terrestrial metagenome studies.

The phylum Bacteroidetes was highly represented by the sorted-MAGs (55.5%) as compared to the bulk metagenome MAGs (0.1%) and unbinned metagenome data (3.8%). In contrast, Acidobacteria was the most abundant phylum in the bulk MAGs (77%) and unbinned metagenome data (32%), as compared to the sorted-MAGs (8.5%). Taxonomic annotation of unassembled reads from Harvard Forest bulk metagenome revealed that Acidobacteria was among the most abundant phyla, comprising 32.0% relative abundance of annotated reads from the total community.

The sorted-MAGs in the phylum Bacteroidetes appeared to be novel as they had relatively poor matches to protein coding sequences from publicly available soil metagenomes. This presumed novelty could also contribute to computation challenges associated with sequence assembly, as only the most abundant taxa are over-represented in public databases (Stepanauskas et al., 2017). Yet, many of these sorted and bulk MAGs were not represented in previous Harvard Forest metagenomes. Taken together, the low representation of our Bacteroidetes sorted-MAGs
across previously published metagenome samples illustrates the expanded biodiversity gained through the use of mini-metagenomes, demonstrating the utility of this approach for accessing the rare taxa within diverse samples.

2.3.5 Biological Insights into Carbon Metabolism in Soil Bacteroidetes

*Bacteroidetes* make up ~10% to the total microbial community in soils (Kiełbasa, Wan, Sato, Horton, & Frith, 2011), yet most of our knowledge about members of this phylum stems from sequenced isolates from vertebrate guts and aquatic habitats (C. Luo, Rodriguez-R, & Konstantinidis, 2014; Seshadri et al., 2018). Furthermore, they are poorly represented in metagenome samples from terrestrial environments (Nayfach & Pollard, 2016). Given the relatively small body of work on soil *Bacteroidetes* and the substantial contribution of 66 putatively novel sorted-MAGs from this study (Figure 2.5 and Figure 2.3), we further explored these sorted-MAGs from *Bacteroidetes* to gain insight into their genomic potential in soils with a particular emphasis on carbon metabolism.

The genome sizes of the sorted-MAGs ranged from 1.6 to 5 Mb, which is somewhat consistent with previously reported *Bacteroidetes* genome sizes that range from 0.9 Mb (*Cardinium* endosymbiont) (Fierer, 2017) to 9.1 Mb (*Chitinophaga pinensis*) (Fernández-Gómez et al., 2013; Kabisch et al., 2014; Thomas, Hehemann, Rebuffet, Czjzek, & Michel, 2011). The smaller genome sizes of the sorted-MAGs was likely due to genome completeness estimates, which ranged from 50% to 80.5% based on analysis of CheckM marker genes (Figure 2.5) (Thomas et al., 2011). The sorted-MAGs were distributed across three distinct families, including *Cytophagaceae*, *Chitinophagaceae*, and *Sphingobacteriaceae*, as well as a clade of unclassified sorted-MAGs (Figure 2.5). Although not much is known about the function of soil Bacteroidetes, they are known
to have a large set of genes that encode enzymes for carbohydrate degradation, including a broad array of glycoside hydrolases, which are phylogenetically conserved across taxa (Penz, Schmitz-Esser, Kelly, Cass, & Müller, 2012). We focused on the glycoside hydrolase, glycosyl transferase and carbohydrate binding module genes, which contribute to the degradation, synthesis and transport of polymeric carbon substrates (Glavina et al., 2010). We hypothesized that the distribution of these genes across the Bacteroidetes mini-MAGs would allow us to infer the ecological roles of these organisms in their native soil environment.

The distribution of CAZy families across the Bacteroidetes exhibited clade-specific abundance patterns of glycoside hydrolases, glycosyl transferases, and carbohydrate binding modules (Figure 2.5) in three families, including the Cytophagaceae, Chitinophagaceae and Sphingobacteriaceae. Members of the Cytophagaceae family appeared to be specialized for polymeric carbon degradation, namely cellulose. Three sorted-MAGs assigned to the Cytophagaceae family contained proteins in glycoside hydrolase family 5, a gene family comprised almost entirely of endocellulases with the potential to degrade polymeric cellulose structures into polysaccharides (Parks et al., 2015), which is consistent with previously sequenced Cytophagaceae genomes (C. Berlemont & Martiny, 2015). In contrast, members of the Chitinophagaceae and Sphingobacteriaceae families appeared to be generalists in carbon utilization. More specifically, the Chitinophagaceae sorted-MAGs harbored the potential to utilize cellulose, hemicellulose and chitin. Seventeen of the twenty-seven sorted-MAGs in the Chitinophagaceae family contained at least one chitinase in glycoside hydrolase family 18 or 19 (Cantarel et al., 2009) along with cellulases in glycoside hydrolase families 5, 8, and 9 and glycoside hydrolases in family 43 that may degrade hemicellulose and pectin (R. Berlemont & Martiny, 2013) (Figure 2.5). In support of this conjecture, the sequenced genome of Chitinophaga pinensis (member of the Chitinophagaceae
family) contains genes to degrade leaf matter and fungal structures, suggesting its ability to
degrade both cellulose and chitin (Taillefer, Arntzen, Henrissat, Pope, & Larsbrink, 2018). Twenty
sorted-MAGs belonged to the family *Sphingobacteriaceae* and typically harbored the potential to
degrade cellulose, xylan, and chitin, with GH families 2, 3, 5, 13, 18 and 20 being the most
abundant across sorted-MAGs in this group. The class *Spingobacteria* are typically found in soils,
with an average abundance of ~4.6% (Hoell, Vaaje-Kolstad, & Eijsink, 2010). Interestingly, one
sorted-MAG (Q3300020668_2) had the highest number of glycoside hydrolase genes within the
*Sphingobacteriaceae* (125 annotated glycoside hydrolases), representing a diverse array of
carbohydrate degradation capabilities and potential metabolic flexibility. This is consistent with
previous investigations describing the family *Sphingobacteriaceae* as capable of degrading diverse
polysaccharides (Mewis, Lenfant, Lombard, & Henrissat, 2016).

Putatively novel *Bacteroidetes* sorted-MAGs stemming from experimental warming plots
at the Harvard Forest Long-Term Ecological Research spanned three different families and
harbored an extensive diversity of CAZymes, such as chitin, cellulose and hemicellulose. The
genomic potential to utilize these labile carbon compounds is consistent with previous
metagenomic investigations in soils of warmed plots (Mckee, Martínez-Abad, Ruthes, Vilaplana,
& Brumer, 2019). Furthermore, previously sequenced genomes of *Bacteroidetes* reveal an
extensive enzymatic repertoire to degrade carbon (Janssen, 2006), presumably making them well-
suited for survival in soils with a diverse collection of carbon compounds from plant material.
Interestingly, the number of identified carbohydrate active enzyme genes increased with genome
size for each of the six CAZy categories (Figure 2.5). This suggests that members of the
*Bacteroidetes* accumulate the capacity to degrade various carbohydrates, presumably expanding
their niche for carbohydrate utilization in the soil.
2.3.6 Reduced Carbon Metabolism in an Unclassified *Bacteroidetes* Clade

Nineteen sorted-MAGs belonged to an unclassified clade of *Bacteroidetes*, which were highly depleted in glycoside hydrolases and carbohydrate binding modules but retained a high number of glycosyl transferases (Figure 2.5). The low abundance of CAZy genes associated with substrate access and degradation may indicate that these organisms are not involved in substrate decomposition. Rather, the relatively higher abundance of glycosyl transferase genes involved in formation of glycoside bonds may indicate that these organisms are responsible for synthesis of higher molecular weight compounds and may depend on living in close association with other organisms.

*Bacteroidetes* such as *Amoebophilus asiaticus* (Shen et al., 2017), *Cardinium sp.* (Y. Luo et al., 2001; Zhou et al., 2011), *Sulcia muelleri* (Thomas et al., 2011) and *Blattabacterium sp.* (Schmitz-Esser et al., 2010) were identified as endosymbionts and symbionts (Figure 2.5). Similar to these known symbionts, the estimated GC contents of unclassified sorted-MAGs in this study were low relative to other Bacteroidetes sequences, with an average of 39.97% GC (Zchori-Fein, Perlman, Kelly, Katzir, & Hunter, 2004). These unclassified *Bacteroidetes* may have limited metabolic capabilities while retaining comparable genome sizes to *Bacteroidetes* previously identified as host-associated. For example, the genome size of *A. asiaticus* was 1.89 Mb, and average assembly size for the unclassified *Bacteroidetes* sorted-MAGs was 2.4 Mb, which are smaller relative to free-living *Bacteroidetes* genomes. Symbionts may undergo the process of reduction in genome size when in contact with the host organism, resulting in a linear relationship between the number of protein-coding genes contained and the size of the genome (Chang et al., 2015). Similarities in the genome structure and relatively low composition of CAZy genes may indicate a symbiotic or host-associated life strategy for some of these sorted-MAGs in the clade of
unclassified *Bacteroidetes*. Completeness estimates of sorted-MAGs within this clade ranged from 51.13% to 80.52%, with an average completeness of 60.5%. Estimated completeness may therefore play a role in the decreased number of annotated CAZy genes.

The abundance of unclassified *Bacteroidetes* within this study may be further evidence of the liberation of symbionts from host cells and vacuoles prior to FACS. Alternatively, the relatively low abundance of glycoside hydrolases in sorted-MAGs within the unclassified clade may be indicative of an opportunistic life strategy (Ló Pez-Sánchez, Neef, Peretó, Patiñ O-Navarrete, & Pignatelli, 2009). The inability to degrade compounds through hydrolysis makes this clade well-suited to living in close association with active degraders.

2.4 Conclusions

This application of mini- and bulk metagenomics has demonstrated the utility of these complementary techniques for biological discovery within the complex soil ecosystem. Using mini-metagenomics to reduce the number of cells prior to sequencing, we have uncovered bacterial and archaeal soil diversity that could not be accessed using bulk metagenomics alone. Mini-metagenomics is a powerful tool for the discovery of rare biosphere organisms and potential endosymbionts, revealing biodiversity in dominant soil groups as well as low abundance taxa. Taken together, mini- and bulk metagenomics allow us to probe deeper into microbial diversity and function within heterogeneous environments beyond soil.

2.5 Methods

2.5.1 Sample Collection and Incubation

Soils were collected on 24 May 2017 from the Barre Woods long-term experimental warming plots located at the Harvard Forest Long Term Ecological Research (LTER) site in
Petersham, MA. Cores were taken from subplots within the larger 30x30 meter plots. Soils were separated into organic (approximately top 5 cm of soil core) and mineral (lower 5 cm of soil core) horizons by visual inspection and were sieved with a 2mm mesh. Approximately 5g of soil was immediately frozen in a dry ice/ethanol bath for DNA extraction, then was transported to the University of Massachusetts Amherst for storage at -80°C. Approximately 15g of soil was transferred to a 50 mL falcon tube for transportation on ice to the Joint Genome Institute (JGI) in Walnut Creek, CA. Samples were further processed as described in (Moran, Mclaughlin, & Sorek, 2009).

2.4.2 Sample Preparation and Cell Sorting

Cells were separated from four incubated soils (heated organic, heated mineral, control organic, and control mineral samples) for FACS through the addition of 0.02% Tween 20 followed by vortexing for 5 minutes. Samples were centrifuged for 5 minutes at 500xg to pellet large soil particles. Following centrifugation, the supernatant was filtered through a 5µm syringe filter to remove the remaining soil particulates. Samples were diluted 1:100 in PBS and stained with SYBR-green. For each of the four soil samples, ninety pools of 100 SYBR+ cells were sorted into microwell plates using a BD Influx cell sorter to perform FACS. Sorted pools underwent cell lysis and whole genome amplification using the Qiagen RepliG Single Cell kit for Multiple Displacement Amplification (MDA). A total of 360 libraries were generated for sequencing with Nextera XT v2 kit (Illumina) with 9 rounds of PCR amplification.

2.4.3 Mini-Metagenomes

Following library preparation, the 360 mini-metagenome libraries were sequenced on the Illumina NextSeq platform at the DOE Joint Genome Institute (JGI, Walnut Creek, CA). Pools of 90
libraries were processed in four sequencing runs with 2x150bp read lengths. Raw Illumina reads were quality filtered to remove contamination and low-quality reads using BBTools (v37.38) (López-Sánchez et al., 2009; McCutcheon & Moran, 2012; Moran et al., 2009), resulting in 359 mini-metagenomes for downstream analysis, as one mini-metagenome did not pass quality filtering standards. Read normalization was performed using BBNorm (R. Berlemont & Martiny, 2013) and error correction was conducted using Tadpole (Bushnell, n.d.). Assembly of filtered, normalized Illumina reads was completed using SPAdes (v3.10.1) (Bushnell, n.d.) with the following options: -phred-offset 33 -t 16 -m 115 --sc -k 25,55,95. All contig ends were trimmed of 200bp and contigs were discarded if the length was <2kb or read coverage was less than 2 using BBMap (Bushnell, n.d.) with the following options: nodisk ambig, filterbycoverage.sh: mincov.

2.4.4 Bulk Metagenomes

Total DNA was extracted from ~0.25g of soil using the DNeasy PowerSoil DNA extraction kit (QIAGEN). Extracted DNA was assessed using the Bioanalyzer and Qubit. Unamplified TruSeq libraries were prepared for 4 DNA samples prior to sequencing on the Illumina HiSeq-2000 platform at the DOE JGI. Raw Illumina reads were trimmed, quality filtered, and corrected using bfc (version r181) with the following options: -1 -s 10g -k 21 -t 10. Following quality filtering, reads were assembled using SPAdes (v3.11.1) (Bushnell, n.d.) with the following options: -m 2000 --only-assembler -k 33,55,77,99,127 --meta -t 32. The entire filtered read set was mapped to the final assembly and coverage information was generated using BBMap (v37.62) (Bankevich et al., 2012) with default parameters except ambiguous=random. The version of the processing pipeline was jgi_mga_meta_rqc.py, 2.1.0. Of the twenty-eight metagenome samples sequenced, only four were selected for inclusion in analysis for this study because they corresponded to those samples sorted using FACS.
2.4.5 Genome Binning and Quality Assessment

Assembled contigs from the mini-metagenomes and bulk metagenomes were binned based on tetranucleotide frequency using MetaBat2 (Bushnell, n.d.). Genome bins were generated for mini-metagenomes without contig coverage patterns due to MDA bias. Binning of mini-metagenomes resulted in 1793 sorted-MAGs, and binning of bulk metagenomes from all samples resulted in 275 bulk metagenome MAGs. Genome bins were assessed for completeness and contamination using a set of marker genes implemented in CheckM (Bankevich et al., 2012). Mini- and bulk metagenome MAGs were filtered to ≥50% completeness, ≤10% contamination, and ≤10% strain heterogeneity, to retain medium quality sorted-MAGs and bulk metagenome MAGs for downstream analysis (Bushnell, n.d.). Following quality filtering, 200 medium quality sorted-MAGs, and 29 medium quality bulk metagenome MAGs remained.

2.4.6 Phylogenetic Tree Construction and Phylogenetic Diversity

A concatenated marker gene phylogenetic tree was constructed for two-hundred medium quality sorted-MAGs, 29 bulk MAGs, and 3,024 reference genomes from soil bacteria and archaea available in the IMG/M database. A set of 56 universal single copy marker proteins (Kang, Froula, Egan, & Wang, 2015) was identified with hmmsearch (v3.1b2) (Parks et al., 2015) and specific HMMs for each of the markers. For every marker protein, alignments were built with MAFFT (v7.294b) (Bowers et al., 2017) and subsequently trimmed with BMGE using BLOSUM30 (Bankevich et al., 2012; Yu et al., 2017). MAGs and reference sequences were clustered at 95% average nucleotide identity with FastANI v1.0 (Eddy, 2011), resulting in 170 sorted-MAGs, 25 bulk MAGs, and 2,341 reference sequences with distinct taxonomic classification. Single protein alignments were then concatenated and a phylogenetic tree inferred with FastTree2 using the
options: -spr 4 -mlacc 2 -slownni -lg (Katoh, Misawa, Kuma, & Miyata, 2002)[89] and was visualized using iTol (Criscuolo & Gribaldo, 2010).

The contribution of sorted-MAGs and bulk MAGs to phylogenetic diversity was determined by calculating the sum of total branch length of contributed genomes relative to reference genomes (Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018). Total Branch length was calculated for a phylogenetic tree containing only 2,341 bacterial and archaeal reference sequences from IMG/M (Price, Dehal, & Arkin, 2010). We then calculated the additional total branch length contributed by sorted-MAGs and bulk MAGs. The percentage increase in total branch length was determined for the complete phylogenetic tree, as well as for clades that included sorted-MAGs.

Taxonomy was assigned to sorted-MAGs, bulk MAGs, and metagenome reads by searching sequences against the NCBI-NR database using DIAMOND (Letunic & Bork, 2016). Blast results were imported into MEGAN6 (Wu et al., 2009) for taxonomic assignment. The relative abundance of each phylum was computed and visualized in R using ggplot2 (Chen et al., 2016).

2.4.7 Gene Recruitment

Two-hundred sorted-MAGs and 29 bulk MAGs were de-replicated by clustering based on 95% average nucleotide identity. Protein coding sequences from the resulting 199 representative sorted-MAGs and MAGs were compared against coding sequences predicted from 2,210 soil metagenome samples from 80 terrestrial metagenome studies stored in the IMG/M database using LAST (Buchfink, Xie, & Huson, 2014)(Figure 2.4). Individual sorted-MAGs and MAGs were designated as a match to metagenome samples if the following criteria were met: a minimum of 200 CDS with hits at ≥ 95% amino acid identity over 70% alignment lengths to CDS of an individual
metagenome. The rationale for choosing the minimum 200 hit count was to ensure that the evidence included more than merely housekeeping genes, which may be more highly conserved. The 90% amino acid identity cutoff was chosen based on Luo et al. 2014 (Huson et al., 2018), who assert that organisms grouped at the 'species' level typically show >85% AAI among themselves. Since our dataset includes divergent sub-lineages, the more conservative threshold of 95% amino acid identity was adopted. The average percentage of CDS with a metagenome hit was calculated for each mini-metagenome (Figure 2.4), and the results were plotted as a multi-bar chart in iTol (Wickham, 2016).

2.4.8 Metabolic Insights

A maximum likelihood tree for Bacteroidetes was constructed using IQTree (Kiełbasa et al., 2011) for the 66 sorted-MAGs and soil Bacteroidetes references from IMG/M. The tree was rooted with Pedosphaera parvula in the phylum Verrucomicrobia. Family level taxonomic classification, genome size, and genome size based on CheckM marker gene assessment (C. Luo et al., 2014) were visualized using iTol (Letunic & Bork, 2016).

Functional annotation for sorted-MAGs was assigned using the Carbohydrate Active Enzyme database (CAZy) (Nguyen, Schmidt, von Haeseler, & Minh, 2015) implemented in dbCAN2 (Parks et al., 2015). The percentage of total annotated genes assigned to each gene family was calculated and is displayed as a multibar chart in iTol (Letunic & Bork, 2016).
Figure 2.1. Overview of mini-metagenome and bulk metagenome approaches used in this study. 

A Mini-Metagenomics performed on four soil samples, including one heated sample from the top organic soil, one heated sample from the lower mineral soil, one control organic sample, and one control mineral sample (n=4). Cells were separated from soil particles using a mild detergent, followed by vortexing, centrifugation and filtration through a 5μm syringe filter. Suspended cells were stained with SYBR-green and sorted into 90 pools of 100 cells each, generating 359 mini-metagenomes. 

B Bulk metagenomic sequencing conducted on the four soils that were used in mini-metagenomics. 

C Following nucleic acid extraction, libraries were prepared, and shotgun sequencing was performed. Sequence data underwent assembly and quality control. Data were binned and assessed for bin quality. Only medium quality genome bins with estimates of 50% completeness, 10% contamination and 10% strain heterogeneity were used in downstream phylogenomic and functional analyses. Further details are provided in the Methods.
Figure 2.2. Assessment of sorted-MAG and MAG quality. Sorted-MAGs (orange, n=1793) and bulk MAGs from the four samples corresponding to those sorted with FACS (blue, n=275). Medium quality sorted-MAGs (dark orange, n=200) and MAGs (dark blue, n=29) are those with ≥50% completeness, ≤10% contamination and ≤10% strain heterogeneity based on analysis of CheckM marker genes (V. Lombard, Golaconda Ramulu, Drula, Coutinho, & Henrissat, 2014). Size of the circle represents the number of 16S rRNA gene copies within each MAG.
Figure 2.3. Phylogenetic diversity of soil taxa identified in this study. A Maximum-likelihood tree of the phylogenetic distribution of medium quality sorted-MAGs and bulk MAGs in the context of previously sequenced soil taxa. Colored branches represent clades that include sorted-MAGs and/or bulk MAGs. Orange branches include only sorted-MAGs, blue include only bulk MAGs, and green include both mini and bulk MAGs. Numbers in orange represent number of contributed sorted-MAGs, blue numbers represent bulk MAGs, and gray numbers represent number of reference sequences in each clade. B Phylogenetic diversity expansion through mini- and bulk MAGs. Gray represents the total branch length contributed by soil reference sequences from the IMG database. Orange bars represent total branch length from sorted-MAGs and blue represents branch length from bulk MAGs. The percentage increase in phylogenetic diversity from this study is shown next to each bar.
Figure 2.4. Comparison of MAGs from this study with published data from terrestrial metagenomes. Innermost is a maximum likelihood tree based on a concatenated alignment of 56 conserved marker proteins from medium quality sorted-MAGs and bulk MAGs recovered in this study. Mini and bulk MAGs were dereplicated by clustering at 95% average nucleotide identity, resulting in 173 sorted-MAGs and 28 bulk MAGs. The clade names are color-coded according to phylum. Individual tracks around the tree depict hits of individual mini- and bulk MAGs by metagenome samples arising from each terrestrial habitat type as specified in the legend. Height of the bar chart indicates the total number of mini- and bulk MAG coding sequences that match metagenome samples. The MAGs were considered matches if they had a minimum of 200 coding sequences with hits at ≥ 95% amino acid identity over 70% alignment lengths to CDS of an individual metagenome. The figure was rendered using iTOL (H. Zhang et al., 2018).
Figure 2.5. Insights into carbon metabolism within the phylum Bacteroidetes. A Concatenated marker gene tree of 66 Bacteroidetes sorted-MAGs and 70 Bacteroidetes reference sequences from the IMG/M database. Sorted-MAGs are distributed across three families of Bacteroidetes, including Cytophagaceae, Chitinophagaceae, Sphingobacteriaceae, and a clade of unclassified sorted-MAGs. B Genome size is shown with darkest color representing the largest genome of 9.1Mb and lightest representing a genome size of 0.6Mb. C Genome completeness is shown as a color gradient, ranging from 50% to 80.5%. D Genome GC content is presented as a gradient presented as a gradient from 21.13% to 61.24%. E, F, G The percentages of genes annotated as glycoside hydrolases (E), glycosyl transferases (F), and carbohydrate binding modules (G) are illustrated as bars with vertical lines denoting 0% and 50% of genes. Bacteroidetes with known symbiotic relationships are indicated with *.
CHAPTER 3

GENOME-INFORMED ECOPHYSIOLOGICAL CHARACTERIZATION OF NOVEL, UNCULTIVATED

SUBDIVISION 2 ACIDOBACTERIA

3.1 Abstract

The bacterial phylum Acidobacteria is found ubiquitously in the environment, representing nearly 20% of global bacterial diversity. In soil, acidobacteria are particularly abundant and are thought to be drivers of carbohydrate degradation as they are slow-growing and contain a large array of genes for carbon metabolism. To date, most of what is known about Acidobacteria has come from bacterial culture or amplicon sequencing studies. These approaches have led to the proposal of 26 subdivisions in this phylum, and have captured a broad scope of diversity in acidobacterial subdivisions 1, 3, 4, and 8. Among the less-well known clades, subdivision 2 acidobacteria are highly abundant in soils, however characterization of this taxonomic clade remains limited and there are no reference genome sequences available. We here analyze Acidobacteria from combined mini-metagenomic and bulk metagenomic characterization of communities in forest soils collected at the Barre Woods warming experiment at the Harvard Forest Long-Term Ecological Research (LTER) site. We generated 26 medium- and high-quality acidobacterial metagenome assembled genomes (MAGs), 13 of which are members of subdivision 2 acidobacteria. This study represents one of the first analyses of the predicted metabolism of subdivision 2 acidobacteria from soils, and describes the ecological niches for these

bacteria. As a major carbon sink, soils are anticipated to become a carbon source under conditions of continued climate change. *Acidobacteria* play a large role in terrestrial carbon cycling, making discovery of acidobacterial diversity and analyses such as these critical to understanding microbial diversity, function and niche specialization as a result of environmental selection under a changing climate.

### 3.2 Introduction

Members of the phylum *Acidobacteria* are found in various terrestrial habitats (Dunbar, Barns, Ticknor, & Kuske, 2002; Hausmann et al., 2018; Kuske, Barns, & Busch, 1997; Männistö, Rawat, Starovoytov, & Häggblom, 2012; Navarrete et al., 2013; Pankratov, 2012; Štursová et al., 2016). While representatives have been identified in freshwater ponds (Zimmermann, Portillo, Serrano, Ludwig, & Gonzalez, 2012), acid mine lakes (Wegner & Liesack, 2017a) and hot springs (Losey et al., 2013), this phylum is of particular interest in soil ecosystems as acidobacteria can comprise around 25 to 40% of the bacterial community (Fierer, 2017). There are 26 recognized subdivisions or clades of the phylum *Acidobacteria* (Barns, Cain, Sommerville, & Kuske, 2007), which can be assigned to 15 class-level units (Dedysh & Yilmaz, 2019). However, this tremendous scope of diversity remains largely underexplored due to methodological challenges (i.e. lack of cultivable lineages, challenges with assembly of low abundance taxa from metagenomes).

Members of taxonomic subdivisions 1, 2, 3, 4 and 6 are typically the most highly abundant in soils (Janssen, 2006; Jones et al., 2009), and data are now emerging on their ecophysiology in the soil environment. Insights have come from whole genome and metagenomic sequencing, yet we currently only have representative genomes available for members of subdivisions 1, 3, 4 and 6 isolated from soils (S. A. Eichorst et al., 2018). *Acidobacteria* harbor a large array of metabolic genes involved in substrate acquisition and degradation, which may make them well-adapted to
fluctuations in nutrient concentration and quality (S. A. Eichorst et al., 2018; Kielak, Matheus, Cipriano, Eiko, & Kuramae, 2016; Naether et al., 2012). The ability to occupy diverse niches in a soil ecosystem, and to adjust to low nutrient conditions may confer an advantage to Acidobacteria under conditions of rapid climatic change.

Although members of subdivision 2 acidobacteria are fairly ubiquitous in soils (Catão et al., 2014; Janssen, 2006), there are currently no published genomes and few metabolic data. Previous studies have revealed the elusive subdivision 2 acidobacteria in high abundance in soils across different biomes using analysis of 16S rRNA (e.g. Catão et al., 2014; DeAngelis et al., 2015; Naether et al., 2012; Navarrete et al., 2013). However, few studies have been able to describe the functional potential of this group of bacteria. Of 11,230 16S rRNA sequences in the SILVA database assigned to subdivision 2, 89% of these were sequenced from soil environments (Quast et al., 2012), which reinforces the importance of studying this group in soils. To date, the metabolism of only one metagenome assembled genome (MAG) classified as a subdivision 2 acidobacterium has been described (Wegner & Liesack, 2017a).

To explore the diversity and genomic potential of Acidobacteria, we applied complementary metagenomic approaches to soil collected at the Barre Woods long-term experimental warming plots, located at the Harvard Forest Long-Term Ecological Research (LTER) site (Alteio et al., in revision; Schulz et al., 2018). Previous studies have shown that these soils harbor an extensive diversity and abundance of Acidobacteria (Alteio et al., in revision; DeAngelis et al., 2015). At these long-term warming sites, Acidobacteria have increased in abundance with warming treatment relative to the control soils, suggesting an important role for these bacteria under a changing climate (DeAngelis et al., 2015). In this study, we uncover a broader diversity of elusive subdivision 2 acidobacteria using metagenome assembled genomes (MAGs) from forest soil and characterize the metabolic potential of these acidobacterial lineages, to investigate
potential ecological niches for these bacteria. Additionally, simulated climate change has resulted in decreases in available soil organic matter and an overall decrease in microbial biomass. Due to these changes, we investigated the metabolism of subdivision 2 acidobacteria to determine if they express traits that make them particularly well-adapted to low-nutrient soil conditions, like those observed in the heated plots at the Barre Woods experimental site.

3.3 Results and Discussion

3.3.1 Taxonomic distribution across bulk- and mini-metagenomes reveals abundant acidobacteria in Harvard Forest soils

The bulk metagenomes from forest soil collected at the Barre Woods long-term warming experiment at the Harvard Forest selected primarily acidobacteria while mini-metagenomic (i.e. pooled cell-sorting) approach yielded a more even distribution of bacterial taxa (Figure 3.1; Alteio et al., in revision). Acidobacteria were the most prevalent in the bulk metagenomes comprising ca. 75% of all reads, in comparison to the mini-metagenomes where they comprise 5% of the total reads (Figure 3.1). The acidobacterial reads represent the major acidobacterial classes detected in previous studies of soil diversity (e.g. Dedysh & Yilmaz, 2019) and are distributed across the following classes: Acidobacteriia, Blastoctocatella, Holophagae, Solibacteres, subdivision 6 and unclassified (Figure 3.1).

We extracted the 16S rRNA genes from mini-metagenome and metagenome reads using PhyloFlash (Gruber-Vodicka, Seah, & Pruesse, 2019) to further explore taxonomy. Hierarchical classification of these extracted 16S rRNA sequences using the RDP Classifier (Q. Wang, Garrity, Tiedje, & Cole, 2007) resulted in 155 and 152 reads identified as acidobacteria from the total set of 1,374 metagenome and 759 mini-metagenome reads, respectively. We found members of subdivisions 1-7, 10 and 13 in this forest soil, with members of subdivisions 1, 2 and 3 being the
most prevalent (Figure 3.2). More specifically, members of subdivision 1 represented ~37% of the acidobacterial reads (26 metagenomic, 88 mini-metagenomic reads), subdivision 2 represented ca. 30% of the acidobacterial reads (48 metagenomic, 46 mini-metagenomic reads) and subdivision 3 represented ca. 6% of the acidobacterial reads (11 metagenomic, 8 mini-metagenomic reads) (Table 3.1).

### 3.3.2 Phylogenomic distribution of acidobacterial MAGs in forest soil expands known diversity of subdivision 2 acidobacteria

Genome binning from these combined approaches resulted in a total of 26 metagenome assembled genomes (MAGs) assigned to the phylum *Acidobacteria*. The estimated completeness of these MAGs ranged from 50.02% to 99.95%, while the estimated contamination (i.e. the presence of lineage-specific marker genes from multiple lineages in a single genome bin as assessed by CheckM) ranged from 0% to 7.22% (Table 3.2; Parks et al., 2015). The estimated genome size ranged from 2.1 to 6.8 Mb, and GC content ranged from 54.6% to 60.9% (Table 3.2), which are smaller in size and lower in GC content when compared to sequenced isolates and MAGs generated from peatland soils (Hausmann et al., 2018).

Using a set of 120 bacterial marker genes from in the Genome Taxonomy Database (GTDB; Parks et al., 2018), we assessed the phylogenomic placement of these 26 MAGs. Twelve MAGs clustered within subdivision 1, and one MAG clustered within subdivision 3 (Figures 3.3, S1). Additionally, 13 MAGs clustered within a clade of acidobacteria reference MAGs that were previously identified as subdivision 2 acidobacteria (Figure 3.3). Compared to the nine available subdivision 2 MAGs in the GTDB database, the unclassified MAGs from this study are higher in completeness and lower in contamination, making them some of the highest quality subdivision 2 MAGs to-date (Figure 3.4).
We screened the 13 unclassified acidobacterial MAGs for the presence of a 16S rRNA gene for better taxonomic assignment using CMsearch (Cui, Lu, Wang, Jing-Yan Wang, & Gao, 2016). Four contained a nearly full-length 16S rRNA gene (1427 to 1534 bp; Figure 3.3). Based on RDP classification, these sequences belong to subdivision 2 with 100% confidence (Table S1). This taxonomic assignment was further confirmed through the generation of a maximum likelihood phylogenetic tree with sequence representatives of subdivision 2 (Figure 3.5).

The generation of numerous high-quality subdivision 2 MAGs in this study presents the opportunity to expand our existing knowledge of diversity within this abundant clade of Acidobacteria. Previously only mentioned in amplicon sequencing studies due to challenges in isolation and assembly using high-throughput sequencing (Jones et al., 2009), MAGs enable exploration of the metabolic diversity of this clade that was previously not possible. The investigation of this group using high-quality draft genomes allows us to elucidate ecological roles and adaptations of subdivision 2 acidobacteria that may explain their abundance and ubiquity across soil ecosystems.

### 3.3.3 Comparison of genomic similarity across acidobacteria supports discovery of novel species

We compared the average nucleotide identity (ANI) across the MAGs and acidobacterial genomes stemming from isolates. To do this, we first refined our acidobacterial dataset using a more stringent quality filtering threshold of ≥80% completeness and ≤2.5% contamination as the isolated genomes have nearly 100% completeness and <5% contamination (S. A. Eichorst et al., 2018). Applying this quality filtering threshold, the resulting dataset included 5 MAGs in subdivision 1, and 5 MAGs in subdivision 2. The ANI values ranged from 74.53% to 80.79% across members of subdivision 2 (Figure S2), suggesting that they could represent unique species based
on the proposed species threshold of 95% ANI (e.g. (Goris et al., 2007; Konstantinidis & Tiedje, 2005; L. Rodriguez-R & Konstantinidis, 2016). The ANI values ranged from 73.57% to 78.39% across members of subdivision 1 (Figure S3), again suggesting unique species. We compared these subdivision 1 MAGs to their closest, cultured relative based on the phylogenomic tree and found that subdivision 1 MAG HF_1_M5 had the highest percent identity (73.53%) to “Candidatus Koribacter versatilis Ellin345” and MAG HF_1_M2 (70.19%) to Acidobacterium capsulatum (Figure S3). This low percent identity based on nucleotide BLAST analyses indicates the substantial diversity we captured within the well-studied subdivision 1 clade of acidobacteria. Clearly, even with extensive available sequence data and cultured isolates, we are only beginning to scratch the surface of this understudied yet ubiquitous group of bacteria.

As the ANI values were relatively low (e.g. 70-78%), we compared average amino acid identity (AAI) across the subdivision 2 acidobacterial genomes to further assess genomic similarity. The AAI values across subdivision 2 acidobacteria ranged from 55.54% to 80.99% (Figure 3.2). The AAI values across subdivision 1 ranged from 50.46% to 71.08% (Figure S3). The low AAI percentages across these acidobacterial subdivisions again suggest that these are distinct species, and perhaps potentially novel genera (Goris et al., 2007; L. M. Rodriguez-R & Konstantinidis, n.d.).

Using these combined approaches – ANI, AAI, as well as phylogenomic and 16S rRNA gene trees – presents opportunities to taxonomically classify uncultivated acidobacteria with high confidence. We recognize that taxonomic identification can only be assigned to bacteria if a cultured isolate is submitted to culture collections, and a host of physiological characteristics are described (Konstantinidis, Rosselló-Móra, & Amann, 2017). However, recent efforts in the community of microbial researchers has opened the possibility for taxonomic classification of uncultivated and ‘yet-to-be cultivated’ microorganisms. According to standards proposed in Konstantinidis et al 2017, the MAGs analyzed in this study are of high enough quality for ANI/AAI
comparisons of genomic discreteness (≥80% completeness, ≤2.5% contamination), and we provide a thorough description of the ecological data surrounding the collection of samples. Additionally, a few of the high quality MAGs contain complete or nearly complete 16S rRNA genes (Konstantinidis et al., 2017). Based on these characteristics and the description of key metabolic functions and gene expression of these MAGs (discussed in later sections), we can confidently define taxonomy for these acidobacterial MAGs.

3.3.4 Genomic potential reflects functional similarities driven by acidobacterial subdivision

The high quality of the draft genomes generated through this study presents the opportunity to characterize some of the potential metabolic diversity of uncultured Acidobacteria through analyses of gene content. We annotated functional genes in the high-quality subdivision 1 and 2 MAGs from this study, as well as select reference genomes, using EggNOG (Huerta-Cepas et al., 2019) in order to describe the functional potential of these organisms and estimate potential niches in the diverse soil ecosystem.

Previous acidobacterial comparative genomics revealed a clear grouping by subdivisions and environments based on the analysis of the COG/NOG categories (Eichorst et al., 2018). We sought to expand upon these observations with our high-quality MAGs (≥80% completeness, ≤2.5% contamination) and performed a principal coordinate analysis of the COG/NOG categories. The PCo1 axis explains 22% of the variability and appears to separate largely based on taxonomic subdivision, while PCo2 axis explains 13% of the variability (Figures 3.6, S4, S5). Acidobacteria in subdivision 1 appear in two clusters, with one appearing more similar to subdivision 3. The second cluster of subdivision 1 acidobacteria falls closer to “Candidatus Koribacter versatilis Ellin345” (referred to as Ellin345 from here forward) and the subdivision 2 acidobacteria, indicating potential similarity in metabolic function.
The clustering of subdivision 1 acidobacteria into two distinct groups along with Ellin345 aligns with the taxonomic features of this particular genome. Phylogenomic analyses place the genome of Ellin345 in a distinct clade from the rest of the subdivision 1 acidobacteria (Figure 3.3), illustrating the taxonomic distinctiveness of this particular genome. The high quality subdivision 1 MAGs from this study including HF_1_M3, HF_1_M4, HF_1_M5 and HF_1_M6 group closely with Ellin345 on the phylogenomic tree (Figure 3.3), and also cluster closely with Ellin345 in the PCoA analyses (Figures 3.6, S4).

The previously-noted unusual metabolic genes found in Ellin345 may be driving the clustering pattern of subdivision 1 and 2 MAGs (Figure 3.6), indicating that subdivision 2 acidobacterial MAGs may have functions in common with this cultivated soil bacterium. A comparative study of three acidobacteria genome sequences in subdivisions 1 and 3 revealed that approximately 25% of coding sequences in Ellin345 were unique to that genome sequence, with the remainder of genes being mainly housekeeping genes (Ward et al., 2009). The clustering pattern including Ellin345 and our additional MAGs demonstrates some functional similarity between these sequences, indicating the potential for some overlap in ecological niches for these acidobacteria.

3.3.5 Carbohydrate degradation potential indicates overlap in metabolism and niche specialization across Acidobacteria

In addition to the broad functional categories assessed using EggNOG (Figure 3.6), we detected a diverse array of genes involved in the transport and metabolism of carbohydrate substrates in subdivision 2 MAGs. This is consistent with previous descriptions of acidobacteria as versatile heterotrophs with relatively slow growth and oligotrophic metabolic strategies (Kielak, Barreto, Kowalchuk, van Veen, & Kuramae, 2016; Lladó et al., 2016; Männistö, Kurhela, Tiirila, & Häggblom, 2013). We explored the carbohydrate degradation potential of the 5 high-quality
subdivision 2 MAGs using the CAZy database (V. Lombard et al., 2014; H. Zhang et al., 2018). To determine if these subdivision 2 MAGs are functionally distinct from reference genomes from other subdivisions (i.e. 1, 3, 4, 6, 8 and 23), we assessed these genomes and draft genomes for patterns of clustering using principal coordinate analysis.

The clustering pattern of the carbohydrate degradation genes are distinct from the COG/NOG pattern (Figures 3.6, S4, S5). The PCo1 axis explained over a third of the variability (37%) and separates based on subdivisions (Figure 3.7). In general, members of subdivisions 1 and 3 are distinct from members of subdivision 2, 4, 8 and 23. This demonstrates differences in the genomic potential for carbohydrate metabolism in members of subdivision 2 in comparison to members of subdivisions 1 and 3 (Figure 3.7, S6, S7). Yet there does appear to be much overlap amongst our MAGs. Select subdivision 1 and 2 MAGs from our forest soil appear to be more similar in terms of carbohydrate degradation potential as compared to previously published isolate genomes (Figure 3.7). Specifically, subdivision 1 MAGs including HF_1_M3, HF_1_M4 and HF_1_M6 cluster within a larger grouping of subdivision 2 MAGs (Figure 3.7).

To determine the genes that might be driving the overlap of subdivisions 1, 2 and 3 we compared the array of genes related to carbohydrate metabolism in MAGs and reference genomes across these subdivisions (Figures S7, S8). Glycoside hydrolase genes comprise 48% of annotated genes in subdivision 1 on average, an average of 33.8% in subdivision 2, and an average of 39.5% in subdivision 3 indicating that hydrolysis of glycosidic bonds is an important metabolic strategy in soil Acidobacteria (Figure S8). Across the three subdivisions, the acidobacteria contain genes in glycoside hydrolase family 13, a broad group containing numerous subfamilies involved in starch and glycogen degradation (C. Berlemont & Martiny, 2015; R. Berlemont & Martiny, 2016). This family was further divided into subfamilies that show improved substrate specificity than
when observed at the broader gene family level (V. Lombard et al., 2014). The genomes in all three subdivisions contain subfamilies 9, 11 and 13, that are involved in the conversion of amylose to amylopectin through the transfer of a glucan group (Labes et al., 2008), isoamylase activity, as well as pullulanase and dextrinase activity, respectively (C. Berlemont & Martiny, 2015). These findings demonstrate that the ability to degrade starch and glycogen compounds is ubiquitous throughout phyla of soil *Acidobacteria* and could be driving the clustering pattern between these genomes.

The pattern of clustering, where subdivision 2 MAGs overlap with subdivisions 1 and 3, indicates that environmental selection may play a role in determining ecological niches for these acidobacteria, rather than function being driven solely by taxonomic clade. The close grouping of subdivision 2 and 3 acidobacteria, as well as some of the subdivision 1 acidobacteria, indicates some functional overlap of these taxa in terms of carbohydrate degradation potential. This demonstrates that the acidobacteria share close ecological niches in the soil environment, and is consistent with the pattern observed in Catao et al (2014) whereby increases in the abundance of subdivision 2 acidobacteria resulted in a corresponding decrease in the abundance of subdivision 1 acidobacteria. The overlap in carbohydrate degradation potential exhibited here indicates similar metabolic strategies which drives direct competition for substrates, and results in environmental selection for particular *Acidobacteria* taxa.

In addition to starch degradation genes, acidobacteria in subdivisions 1 and 3 contain a broader array of glycoside hydrolases across different gene families. These acidobacteria contain glycoside hydrolases in families 36 and 42 with galactose and arabinopyranoside hydrolytic activity (Figure S7). Additionally, genomes in subdivision 1 and contain genes in GH families 79 with glucuronidase and heparanase activity, families 92 and 125 with mannose hydrolytic activity, as well as families 95 and 141 which break down fucose and xylan (Figure S7). However, these genes
are absent across the subdivision 2 MAGs, demonstrating that genomes in subdivisions 1 and 3 have broader array of genes with the potential to degrade a wider range of carbohydrate substrates. Overall, carbohydrate metabolism may be reduced in subdivision 2 when compared to acidobacteria in subdivisions 1 and 3.

Although the relatively small number of glycoside hydrolases found in subdivision 2 sequences could be attributed to the incomplete nature of these draft genomes relative to isolate genomes, the absence of these genes may also reflect an alternative strategy of these subdivision 2 acidobacteria for carbohydrate metabolism. Many of the CAZy genes found in subdivision 2 draft genomes are associated with hydrolysis of relatively simple carbohydrate molecules, the abundance of which suggests that the putative subdivision 2 acidobacteria are not involved in degradation of complex carbon substrates, but rather, metabolize more readily oxidizable substrates. The ability of subdivision 2 acidobacteria to degrade simple compounds including glucose, xylose and trehalose in addition to starch, is distinct from the more diverse array of carbohydrate genes in subdivisions 1 and 3.

### 3.3.6 Gene expression analysis suggests metabolic flexibility of subdivision 2 acidobacteria

In addition to functional potential of coding sequences in the five subdivision 2 MAGs, we mapped assembled contigs from four Barre Woods soil metatranscriptomes to the MAGs to estimate gene expression. We assigned functional classifications to these mapped contigs using the EggNOG and KEGG databases (Huerta-Cepas et al., 2019; M. Kanehisa & Goto, 2000). For central carbon metabolism, we find evidence of expression of the complete Embden-Meyerhof, tricarboxylic acid cycle, and the pentose phosphate pathway, as well as evidence of the partial Entner-Doudoroff pathway (Figure 3.8). This is consistent with the obligate aerobic strategy of most acidobacteria genera, with the exception of Acidobacterium and Telmatobacter sp. that are
facultatively anaerobic (Campbell, 2014).

In addition to central carbon metabolism, we evaluated the expression of genes involved in accessing nutrients and substrates as these are critical for bacterial survival. The five subdivision 2 acidobacteria MAGs express diverse ABC transporters including transporters for uptake of phosphate, D-xylose, glucose, mannose, sorbitol and mannitol (Figure 3.8). Additionally, the genomes express two-component systems including CusS and CusR for the uptake of copper ions, as well as PhoBRP, SenX3, RegX3, and PstS genes that are indicative of phosphate uptake under limiting conditions (Figure 3.8). Expression of these genes involved in substrate uptake are indicative of versatile heterotrophy that is characteristic of bacteria in the phylum Acidobacteria (S. A. Eichorst et al., 2018; Kielak, Barreto, et al., 2016).

We also evaluated the expression of genes involved in motility. Functional annotations of the MAGs and their respective transcripts reveal evidence of a non-motile life strategy with genes for motility and flagellar assembly not expressed across all five genomes (Table 3.3). This lack of expression of flagellar assembly is in contrast to genomes from cultured isolates in subdivision 1, particularly Ellin345 and Acidobacterium capsulatum (Ward et al., 2009); however, this trait is variable across the phylum Acidobacteria (Campbell, 2014). The absence of expression of these genes indicates that these bacteria were likely sessile at time of collection.

3.3.6.1 Evidence of assimilatory sulfur metabolism

In previous studies of Acidobacteria from peatland soils and acid mine drainage, sulfur cycling is an essential process for survival (Hausmann et al., 2018; Wegner & Liesack, 2017a). We assessed the expression for sulfur metabolism in our subdivision 2 acidobacteria MAGs. Transcripts mapped to the five subdivision 2 acidobacteria MAGs and provide evidence of
assimilatory sulfate reduction in forest soils. Additionally, the five MAGs express Sox genes involved in thiosulfate oxidation, including SoxC, SoxD, and SoxY (Figure 3.8). In contrast to previously studied acidobacteria from peatland soils in subdivisions 1 and 3, we detect no evidence of dissimilatory sulfate reduction (including dsrABC or aprBA) in the subdivision 2 MAGs. Environments such as acid mine drainage and peatlands contain sulfur compounds in high to even toxic concentrations, compared to temperate forest soils from which these acidobacteria MAGs were generated that have relatively low sulfur. Therefore, cycling of sulfur compounds may be important but not crucial for survival in temperate forest soils in comparison to highly sulfuric environments.

3.3.6.2 Assimilatory nitrate reduction in subdivision 2

Nitrogen is an essential nutrient, the cycling of which is mediated largely by microorganisms in terrestrial environments (Lladó et al., 2017; Jerry M Melillo et al., 2011; C. Wang et al., 2018). All five subdivision 2 acidobacteria MAGs have evidence of assimilatory nitrate reduction, particularly the expression of NarB, NR, and NasAB (Figure 3.8). Additionally, the five genomes express most of the genes involved in dissimilatory nitrate reduction to ammonia (Figure 3.8). The reduction of nitrate to nitrite is encoded for by NarGHI and NapAB genes across four of the five subdivision 2 MAGs. Further reduction of nitrite to ammonia is encoded for by the expression of NirBD and NrfAH, both of which are present in four of the five subdivision 2 genomes, whereas only NirBD is expressed in HF_2_M10 (Figure 3.8).

In studies of nitrogen cycling dynamics in combination with simulated climate change, warming has induced increases in nitrogen mineralization which is then used by vegetation (Butler et al., 2012; Rustad et al., 2001). At the Harvard Forest LTER site, increased nitrogen mineralization has been observed in the warmed soil plots as compared to the control (Butler et al., 2012; Serita
The subdivision 2 acidobacteria MAGs in this study express genes involved in assimilatory and dissimilatory nitrogen reduction, as well as the conversion of nitrite to ammonia. This indicates that acidobacteria contribute to nitrogen mineralization, making inorganic nitrogen more available to plants, which in turn is consistent with measured increases in nitrogen mineralization in response to warming (Butler et al., 2012).

### 3.3.6.3 Terminal oxidases enable respiration in low oxygen environments

Using the results from EggNOG and KEGG (Huerta-Cepas et al., 2019; M. Kanehisa & Goto, 2000), we further explored the metabolic potential of subdivision 2 acidobacteria to determine strategies that make them well-adapted to soil ecosystems. Evidence of obligately aerobic respiration is supported by the lack of alternative electron donors within the subdivision 2 acidobacteria MAGs (Campbell, 2014; Wegner & Liesack, 2017b). However, these genomes have transcripts for high- and low-affinity terminal oxidases: all five genomes contain evidence of Cox gene expression, which is part of heme-cytochrome oxidase class A (aa3-type) (S. A. Eichorst et al., 2018). Additionally, the five genomes express high oxygen affinity terminal oxidases (bd-type) in heme-cytochrome oxidase class C (S. A. Eichorst et al., 2018).

The presence of both low- and high-affinity terminal oxidases contributes to the metabolic flexibility of these acidobacteria by enabling respiration across oxygen gradients (Pereira, Santana, & Teixeira, 2001; Sousa et al., 2012), even in microaerobic conditions. Given the heterogeneity of nutrients and oxygen in soils, (Hansel, Fendorf, Jardine, & Francis, 2008; Rillig et al., 2017), nutrient gradients within the heterogeneous soil structure may also lead to ‘hot-spots’ of microbial activity, and additionally result in anoxic microenvironments (Borer, Tecon, & Or, 2018; Kuzyakov & Blagodatskaya, 2015; Rillig et al., 2017). Soil structure as well as biological processes that mediate
the distribution of nutrients and oxygen necessitate flexible metabolic strategies, such as the expression of low- and high-affinity terminal oxidases that support the survival of subdivision 2 acidobacteria.

### 3.3.6.4 Hydrogenases enable hydrogen scavenging and utilization

We assessed expression of 1h/5 [Ni-Fe] hydrogenases that are involved in hydrogen scavenging and uptake from the environment. Transcripts mapped to the subdivision 2 MAGs reveal that all five genomes express [NiFe] hydrogenases, including the genes *HupS* and *HupL* (Figure 3.8). These genes code for the expression of the small and large catalytic subunits of [NiFe] hydrogenases, respectively. The distribution of hydrogenases is scattered throughout the bacterial domain, and is consistent with frequent horizontal gene transfer (Figure S9; S. A. Eichorst et al., 2018; Greening et al., 2015).

The expression of large and small subunit [NiFe] hydrogenases among our MAGs indicates that hydrogen scavenging and uptake is a metabolic strategy across acidobacteria in soils. Hydrogen presents a valuable growth substrate in environments where carbon substrates have become more limiting (Conrad, 1996; Greening et al., 2015). In particular, for acidobacteria that exhibit a slow-growing, oligotrophic strategy, hydrogen scavenging supports dormancy and metabolism in nutrient poor environments (Greening et al., 2015). The ability to use hydrogen as a substrate is consistent with observed decreases in readily oxidizable organic matter present in the warming plots at the Harvard Forest long-term warming simulation experiments (Pold, Grandy, Melillo, & Deangelis, 2017). With reduced availability of carbohydrate substrates which acidobacteria are known to readily degrade, hydrogen utilization offers a survival strategy in low nutrient environments (Greening et al., 2016, 2015), such as soil under long-term warming exposure.
3.3.7 Metabolic strategies across Acidobacteria reflect environmental selection and niche specialization

In addition to the presence of carbohydrate metabolism genes in subdivision 2 acidobacteria, we assessed the expression of genes involved in carbohydrate degradation by assigning functional classification to mapped metatranscriptomes using the carbohydrate active enzyme database (CAZy; V. Lombard et al., 2014) (Figure S10, Table 3.4). Subdivision 2 acidobacteria express the starch degradation genes from the GH13 subfamilies 9, 11 and 13 (C. Berlemont & Martiny, 2015). Additionally, all five of the subdivision 2 MAGs express genes in glycoside hydrolase families GH3, 23 and 109. Genes in family GH3 act broadly on monomers containing glucosides, arabinofuranosides, xylopyranosides and N-acetyl-glucosamines (López-Mondéjar, Zühlke, Becher, Riedel, & Baldrian, 2016). Families 23 and 109 degrade peptidoglycan, and N-galactosamine, respectively (López-Mondéjar et al., 2016). Additionally, genes in glycoside hydrolase families GH15, 16, 29 and 128 are expressed across at least 3 subdivision 2 MAGs. These genes break down glucans, fucose, galactose, xylose, dextrin, and trehalose (López-Mondéjar et al., 2016). Glycoside hydrolase family 18 is expressed as well, and exhibits chitinase and N-acetylglucosaminidase activity (Belova et al., 2018).

Taken together, the expression of these gene families provides further support for flexible carbon substrate metabolism exhibited by subdivision 2 acidobacteria. The abundance of genes in the glycoside hydrolase gene family, representing an average of 40.8% of expressed genes annotated with CAZy across the five subdivision 2 genomes (Figure S11), reveals the versatile heterotrophic strategy of subdivision 2 acidobacteria. This metabolic flexibility could support the abundance of subdivision 2 acidobacteria in nutrient-depleted soils, and enable them to adapt to changing environmental conditions.
Climate change is expected to induce shifts in microbially-mediated processes in soil by changing the quality and quantity of carbon substrates (Bond-Lamberty et al., 2016; Jerry M Melillo et al., 2011; Pold et al., 2017). In addition to containing a small suite of glycoside hydrolase genes, the reduced ability for subdivision 2 acidobacteria to degrade complex carbon substrates is further supported by changes in the available substrate pool measured at the Harvard Forest experimental warming plots after extensive warming (Pold et al., 2017). At the longest-running Harvard Forest warming experiment at Prospect Hill, warming has induced a decrease in recalcitrant carbon substrates, providing more readily oxidizable carbon for microorganisms (Pold et al., 2017). This reduction in structurally complex substrates and subsequent increase in more easily degraded compounds supports an expanded niche for subdivision 2 acidobacteria under soil warming conditions.

In accordance with this, subdivision 2 acidobacteria from samples collected at the sister warming site at Barre Woods are found in higher abundance in the lower mineral horizon soils compared to the upper organic soils (Figure S12). Mineral soils contain fewer inputs of vegetation and complex organic matter when compared to organic soils, which may select for organisms that can survive in low-nutrient conditions. The reduced expression of carbon degradation genes in subdivision 2 acidobacteria, and their increased abundance in low nutrient habitats suggests that they are well-equipped to survive in these conditions. These findings are consistent with a microbial community with more specific and relatively reduced carbon metabolism as measured in previous studies at the Harvard Forest (Bradford et al., 2008; S. D. Frey et al., 2008; Pold et al., 2017), indicating an important role for subdivision 2 acidobacteria under a changing climate.

3.8 Synthesis

*Acidobacteria* are ubiquitous across environments, but have traits that may make them
particularly well-adapted to survival in soils. The generation of the highest-quality subdivision 2 MAGs to-date in this study has enabled genome-informed investigation of this previously undescribed group of bacteria. In combination with metatranscriptomics, we were able to characterize the taxonomic and metabolic diversity of subdivision 2 acidobacteria. Additionally, we were able to hypothesize niche specialization of this subdivision based on expression of genes that make them well-adapted to low-nutrient soils.

The abundance of genes and transcripts allocated to carbohydrate metabolism in these genomes, as well as the expression of terminal oxidases and hydrogenases indicate that the subdivision 2 acidobacteria, in particular, are versatile heterotrophs capable of survival in rapidly changing conditions present in soil microenvironments. Climate change in the form of soil warming is expected to alter the composition of available carbon in soils through microbially-mediated activity. Changes in soil moisture, as well as substrate and nutrient availability induced by climate change may select for more slow-growing, oligotrophic taxa such as acidobacteria. The high abundance of subdivision 2 acidobacteria in soil ecosystems make them particularly interesting for understanding how microbially-mediated processes might shift in response to climate change.

3.9 Methods

3.9.1 Sample Collection and Processing

A set of fourteen soil cores were collected on 24 May 2017 from the Barre Woods plot of the Harvard Forest Long-Term Ecological Research (LTER) site (Petersham, MA, United States). Soil samples were collected using a tulip bulk corer and cores were separated into organic (top layer) and mineral horizon (lower layer) based on visual inspection, resulting in 28 soil samples. Samples were flash-frozen in the field using an ethanol and dry ice bath. Frozen samples were transported
to the University of Massachusetts Amherst (Amherst, MA, United States) for long-term storage, and to the Joint Genome Institute (JGI; Walnut Creek, CA, United States) for further processing and nucleic acid extraction.

3.9.2 Mini-Metagenomics

From the original set of 28 soil samples, one representative sample from each treatment and soil horizon (heated organic, heated mineral, control organic and control mineral) was selected for downstream processing using cell sorting coupled to high-throughput sequencing. Cells and large soil particles were separated through the application of tween 20 and vortexing, followed by centrifugation to pellet smaller soil particles. Intact cells were stained using SYBR-green, and 100 SYBR+ cells were sorted into 90 pools using fluorescence activated cell sorting for each of the four selected soils. These pools of cells underwent multiple displacement amplification (MDA) and downstream nextera library preparation prior to sequencing on the Illumina NextSeq platform. See Schulz et al 2018 and Alteio et al in revision for further details on sample preparation (Alteio et al., in revision; Schulz et al., 2018).

3.9.3 Bulk Metagenomics

The complete set of 28 soil samples was extracted for total DNA using the DNeasy Soil DNA extraction kit (QIAGEN). The 28 unamplified TruSeq libraries were sequenced on the Illumina HiSeq platform at the Joint Genome Institute (JGI; Walnut Creek, CA, United States). Following standard quality filtering procedures, reads were assembled using SpAdes (Bankevich et al., 2012). From the original 28 samples, the 4 samples corresponding to those that were processed via fluorescence activated cell sorting were selected for analysis in this study.
3.9.4 Metatranscriptomics

Total RNA was extracted from the complete set of soil samples collected at the Barre Woods site using the RNeasy Soil RNA extraction kit (QIAGEN). Libraries were prepared and were sequenced on the Illumina NextSeq platform at the Joint Genome Institute (JGI; Walnut Creek, CA, United States). Quality cleaned reads were assembled using the MEGAHIT assembler (D. Li, Liu, Luo, Sadakane, & Lam, 2015).

3.9.5 Genome Binning and Quality Assessment

Genome bins were generated from the mini-metagenome and bulk metagenome datasets using MetaBat2 (Kang et al., 2015). Using this approach 1793 mini-MAGs and 1776 bulk MAGs were generated. Of these, 11 sorted-MAGs and 14 bulk MAGs were classified to the phylum Acidobacteria. Following genome binning, bins were assessed for quality using marker genes implemented in CheckM (Parks et al., 2015). Sorted-MAGs and MAGs were retained for downstream analyses if they surpassed a quality filtering threshold of ≥50% estimated completeness, ≤10% estimated marker gene contamination, and ≤10% strain heterogeneity. Following quality filtering 11 mini-MAGs and 14 bulk MAGs within the phylum Acidobacteria remained. This dataset was further refined to contain draft genomes with ≥80% estimated completeness and ≤2.5% estimated marker gene contamination in order to better compare with sequenced genomes from isolates in the phylum Acidobacteria. Following this refined quality threshold, 5 MAGs from Subdivision 1 and 5 MAGs from Subdivision 2 were retained, respectively.

3.9.6 Distribution of taxa in bulk and sorted-MAGs

Taxonomy was assigned to contigs from the sorted-MAGs and bulk MAGs by comparing sequence identity against the NCBI-NR database using DIAMOND (Buchfink et al., 2014). Blast
results were imported into MEGAN6 (Huson et al., 2018) for taxonomic classification. The relative abundance of each phylum was computed and visualized in R using ggplot2 (Wickham, 2016).

3.9.7 Taxonomic distribution of extracted 16S rRNA gene sequences

To assess the distribution of taxa across the mini-metagenome and bulk metagenome samples, 16S rRNA gene sequences were extracted using PhyloFlash (Gruber-Vodicka et al., 2019) from the set of 759 reads and 1,374 reads from the mini- and bulk metagenome sequences, respectively. These reads were classified to taxonomy using the RDP classifier (Q. Wang et al., 2007), and relative abundance of taxonomic distribution was visualized using the R package ggplot2 (Wickham, 2016).

In addition to extracting 16S rRNA sequences from the unassembled read data, 16S rRNA gene sequences were identified in the metagenome assembled genomes (MAGs) and sorted-MAGs using CMsearch (Cui et al., 2016). This approach revealed 4 of the 13 putative subdivision 2 acidobacterial MAGs contained a full-length 16S rRNA gene (between 1427 to 1534 bp). These extracted 16S rRNA gene sequences were used to generate a maximum-likelihood tree of subdivisions 1, 2, 3, 4, 6 and 8, including 16S rRNA genes from SILVA as reference sequences. The species Acanthopleuribacter pedis (AB303221) was used as an outgroup for the tree. The tree was bootstrapped 1000 times, and bootstrap values are displayed on the tree as ≥95% (●) and ≥70% (●).

3.9.8 Phylogenomic analysis of MAGs and Sorted-MAGs

The GTDB-Toolkit (GTDB-TK) was used to generate a phylogenomic tree in order to determine taxonomic distribution for MAGs derived from mini-metagenomes and bulk
metagenomes (Parks et al., 2018). This tool assessed the MAGs for the presence of 120 marker genes, and generated a concatenated marker gene tree using the GTDB as a reference dataset. The resulting tree was visualized using ARB (Ludwig et al., 2004). A subtree containing reference sequences and MAGs in the phylum *Acidobacteria* was extracted and visualized.

### 3.9.9 Average Nucleotide Identity and Amino Acid Identity

To assess genome similarity across acidobacterial subdivision 2 MAGs (≥80% completeness and ≤2.56% contamination), we used the ANI/AAI matrix tool to calculate average nucleotide identity (ANI) and amino acid identity (AAI) (L. Rodriguez-R & Konstantinidis, 2016). This was repeated for subdivision 1 MAGs and selected subdivision 1 isolate genomes using the ANI/AAI matrix tool. The resulting ANI and AAI values were visualized as heatmaps using the ggplot2 package in R (Wickham, 2016).

### 3.9.10 Metabolic Insights from MAGs

Functional classifications from the KEGG database were assigned to MAGs derived from bulk and mini-metagenome samples using GhostKoala (Minoru Kanehisa, Sato, Morishima, & Sternberg, 2016). Additionally EggNOGmapper (Huerta-Cepas et al., 2019) was used to assign COG/NOG categories to MAGs and sorted-MAGs (Huerta-Cepas et al., 2019). Gene families involved in carbon substrate degradation were identified in MAGs and sorted-MAGs using the Carbohydrate Active Enzyme database (CAZY; V. Lombard et al., 2014) implemented in DBcan2 (H. Zhang et al., 2018).

We assessed the functional similarity between Acidobacteria MAGs from this study, published MAGs and genomes from Acidobacteria isolates in a principal coordinates analysis by
calculating Bray-Curtis Similarity between EggNOG annotated genes across MAGs and isolate genomes. Clustering patterns were visualized using a PCoA plot generated in ggplot2 (Wickham, 2016). Genes that may play a role in the visualized patterns of clustering were identified in a heatmap generated using the superheat package in R (Barter & Yu, 2017). Similarly, patterns of clustering and characterization of carbohydrate metabolism were identified using PCoA plot and heatmap of the CAZy categories.

3.9.11 Estimation of gene expression in Acidobacteria

Assembled contigs from the four metatranscriptomes were aligned to high quality MAGs classified to Acidobacteria subdivision 2 using bbsplit, part of the bbtools suite (Bushnell, n.d.). Reads that mapped to acidobacterial MAGs were assigned functional annotation using GhostKoala to assign KEGG categories (M. Kanehisa & Goto, 2000; Minoru Kanehisa et al., 2016), EggNOGmapper to assign COG/NOGs from the EggNOG database (Huerta-Cepas et al., 2019) and the Carbohydrate Active Enzyme database (CAZy) implemented in DBcan2 (V. Lombard et al., 2014; H. Zhang et al., 2018).

3.9.12 Detecting hydrogenases in acidobacterial MAGs

Sequences from genomes known to contain group 1h/5 [NiFe]-hydrogenases were aligned using MUSCLE (Edgar, 2004). A maximum likelihood tree of these reference sequences and MAGs was generated in IQTree, ModelFinder (Nguyen et al., 2015) and was bootstrapped 1000 times. The tree is rooted with Desulfovibrio vulgaris (WP_010939204.1).
Table 3.1. Number of reads assigned to Acidobacteria from the mini- and bulk metagenome datasets using RDP Classifier.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sequencing Type</th>
<th>Total #16S rRNA Reads</th>
<th>#16S rRNA reads in Acidobacteria</th>
<th>#Subdivision 1</th>
<th>#Subdivision 2</th>
<th>#Subdivision 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BWMinControl</td>
<td>Mini-Metagenome</td>
<td>190</td>
<td>47</td>
<td>19</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>BWMinHeated</td>
<td>Mini-Metagenome</td>
<td>202</td>
<td>40</td>
<td>18</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>BWOrgControl</td>
<td>Mini-Metagenome</td>
<td>226</td>
<td>35</td>
<td>26</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>BWOrgHeated</td>
<td>Mini-Metagenome</td>
<td>141</td>
<td>30</td>
<td>25</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>BW_C_12_M</td>
<td>Metagenome</td>
<td>53</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>BW_C_12_O</td>
<td>Metagenome</td>
<td>47</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>BW_H_2_M</td>
<td>Metagenome</td>
<td>62</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>BW_H_2_O</td>
<td>Metagenome</td>
<td>48</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3.2 CheckM quality assessment of bulk and sorted-MAGs classified to the phylum Acidobacteria.

<table>
<thead>
<tr>
<th>Bin Name</th>
<th>Sample Location</th>
<th>Completeness</th>
<th>Contamination</th>
<th>Strain heterogeneity</th>
<th>Genome size (Mbp)</th>
<th>GC%</th>
<th>#16S rRNA Genes</th>
<th>Subdivision</th>
<th>Treatment</th>
<th>Horizon</th>
<th>Seq Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF_1_M1</td>
<td>Inc-BW-H-2-M</td>
<td>79.39</td>
<td>2.62</td>
<td>0</td>
<td>4.90</td>
<td>56.4</td>
<td>0</td>
<td>1</td>
<td>Heated</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_M2</td>
<td>Inc-BW-C-12-O</td>
<td>90.88</td>
<td>2.14</td>
<td>0</td>
<td>5.05</td>
<td>58.6</td>
<td>0</td>
<td>1</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_M3</td>
<td>Inc-BW-C-12-M</td>
<td>99.95</td>
<td>0.85</td>
<td>0</td>
<td>4.67</td>
<td>57.8</td>
<td>0</td>
<td>1</td>
<td>Heated</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_1_M4</td>
<td>Inc-BW-C-12-M</td>
<td>92.31</td>
<td>0.85</td>
<td>0</td>
<td>4.81</td>
<td>56.1</td>
<td>1</td>
<td>1</td>
<td>Control</td>
<td>Mineral</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_1_M5</td>
<td>Inc-BW-C-12-M</td>
<td>98.93</td>
<td>2.56</td>
<td>0</td>
<td>6.85</td>
<td>58.5</td>
<td>0</td>
<td>1</td>
<td>Heated</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_S1</td>
<td>BW_H_2_O</td>
<td>50.02</td>
<td>5.17</td>
<td>0</td>
<td>3.48</td>
<td>57.5</td>
<td>0</td>
<td>1</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_S2</td>
<td>BW_C_12_M</td>
<td>53.51</td>
<td>0</td>
<td>0</td>
<td>3.53</td>
<td>56.1</td>
<td>1</td>
<td>1</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_S3</td>
<td>BW_H_2_O</td>
<td>53.32</td>
<td>1.71</td>
<td>0</td>
<td>3.98</td>
<td>56</td>
<td>3</td>
<td>1</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_S4</td>
<td>BW_C_12_M</td>
<td>54.94</td>
<td>4.27</td>
<td>0</td>
<td>6.52</td>
<td>57.5</td>
<td>0</td>
<td>1</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_1_S5</td>
<td>BW_C_12_M</td>
<td>56.32</td>
<td>0</td>
<td>0</td>
<td>4.47</td>
<td>55.5</td>
<td>0</td>
<td>1</td>
<td>Heated</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_1_S6</td>
<td>BW_C_12_O</td>
<td>52.19</td>
<td>9.4</td>
<td>0</td>
<td>3.67</td>
<td>56.5</td>
<td>0</td>
<td>1</td>
<td>Control</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_2_M1</td>
<td>Inc-BW-H-2-M</td>
<td>87.92</td>
<td>2.56</td>
<td>0</td>
<td>4.24</td>
<td>58.4</td>
<td>2</td>
<td>2</td>
<td>Control</td>
<td>Organic</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_M2</td>
<td>Inc-BW-C-12-O</td>
<td>95.44</td>
<td>2.56</td>
<td>0</td>
<td>6.51</td>
<td>58</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_M3</td>
<td>Inc-BW-C-12-O</td>
<td>93.59</td>
<td>2.99</td>
<td>0</td>
<td>5.35</td>
<td>57.9</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_M4</td>
<td>Inc-BW-C-12-M</td>
<td>88.03</td>
<td>2.56</td>
<td>0</td>
<td>4.53</td>
<td>58.2</td>
<td>1</td>
<td>2</td>
<td>Heated</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_2_M5</td>
<td>Inc-BW-C-12-M</td>
<td>92.09</td>
<td>1.71</td>
<td>0</td>
<td>4.53</td>
<td>57.9</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Mineral</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_2_M6</td>
<td>Inc-BW-C-12-M</td>
<td>90.46</td>
<td>1.71</td>
<td>0</td>
<td>4.94</td>
<td>58.8</td>
<td>2</td>
<td>2</td>
<td>Control</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_2_M7</td>
<td>Inc-BW-C-12-M</td>
<td>90.43</td>
<td>3.06</td>
<td>10</td>
<td>4.52</td>
<td>58.8</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_2_M8</td>
<td>Inc-BW-C-12-M</td>
<td>76.23</td>
<td>0.85</td>
<td>0</td>
<td>4.52</td>
<td>58.4</td>
<td>2</td>
<td>2</td>
<td>Control</td>
<td>Organic</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_M9</td>
<td>Inc-BW-H-2-O</td>
<td>83.17</td>
<td>7.03</td>
<td>7.69</td>
<td>2.11</td>
<td>60.9</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Organic</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_M10</td>
<td>Inc-BW-H-2-O</td>
<td>95.51</td>
<td>0.85</td>
<td>0</td>
<td>4.73</td>
<td>57.5</td>
<td>2</td>
<td>2</td>
<td>Control</td>
<td>Mineral</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_S1</td>
<td>BW_C_12_M</td>
<td>71.08</td>
<td>3.42</td>
<td>0</td>
<td>4.38</td>
<td>54.6</td>
<td>1</td>
<td>2</td>
<td>Heated</td>
<td>Organic</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_S2</td>
<td>BW_C_12_M</td>
<td>52.14</td>
<td>0.85</td>
<td>0</td>
<td>3.08</td>
<td>59.3</td>
<td>1</td>
<td>2</td>
<td>Heated</td>
<td>Organic</td>
<td>Bulk MAG</td>
</tr>
<tr>
<td>HF_2_S3</td>
<td>BW_C_12_O</td>
<td>67.85</td>
<td>0</td>
<td>0</td>
<td>3.01</td>
<td>57.6</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Mineral</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_2_S4</td>
<td>BW_H_2_O</td>
<td>50.14</td>
<td>0</td>
<td>0</td>
<td>2.19</td>
<td>59.9</td>
<td>1</td>
<td>2</td>
<td>Control</td>
<td>Mineral</td>
<td>Mini-MAG</td>
</tr>
<tr>
<td>HF_3_S1</td>
<td>BW_C_12_O</td>
<td>59.35</td>
<td>7.22</td>
<td>0</td>
<td>4.38</td>
<td>58.1</td>
<td>1</td>
<td>3</td>
<td>Control</td>
<td>Organic</td>
<td>Mini-MAG</td>
</tr>
</tbody>
</table>

77
Table 3.3. Number of expressed KEGG genes in subdivision 2 acidobacteria draft genomes.

<table>
<thead>
<tr>
<th>Tax #</th>
<th>Gene Abbreviation</th>
<th>Gene Name</th>
<th>sf_1</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
<th>M7</th>
<th>M8</th>
<th>M9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1045</td>
<td>hgpF</td>
<td>hydrogenase expression/foldon protein HgpF</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10485</td>
<td>hgpB</td>
<td>hydrogenase nickel incorporation protein HgpB</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2045</td>
<td>hgpC</td>
<td>hydrogenase expression/foldon protein HgpC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10454</td>
<td>hgpD</td>
<td>hydrogenase expression/foldon protein HgpD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10453</td>
<td>hgpA</td>
<td>hydrogenase nickel incorporation protein HgpA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10455</td>
<td>hgpL</td>
<td>hydrogenase large subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10459</td>
<td>hgpH</td>
<td>uptake hydrogenase large subunit</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hgpA</td>
<td>hydrogenase small subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109274</td>
<td>hpsA</td>
<td>cytosol c o2 oxidase substrate 1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsB</td>
<td>cytosol c o2 oxidase c-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109245</td>
<td>hpsD</td>
<td>cytosol c o2 oxidase d-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109246</td>
<td>hpsE</td>
<td>cytosol c o2 oxidase e-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsF</td>
<td>cytosol c o2 oxidase f-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsG</td>
<td>cytosol c o2 oxidase g-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsH</td>
<td>cytosol c o2 oxidase h-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsI</td>
<td>cytosol c o2 oxidase i-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsJ</td>
<td>cytosol c o2 oxidase j-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsK</td>
<td>cytosol c o2 oxidase k-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsL</td>
<td>cytosol c o2 oxidase l-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsM</td>
<td>cytosol c o2 oxidase m-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsN</td>
<td>cytosol c o2 oxidase n-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsO</td>
<td>cytosol c o2 oxidase o-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsP</td>
<td>cytosol c o2 oxidase p-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsQ</td>
<td>cytosol c o2 oxidase q-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsR</td>
<td>cytosol c o2 oxidase r-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsS</td>
<td>cytosol c o2 oxidase s-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsT</td>
<td>cytosol c o2 oxidase t-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsU</td>
<td>cytosol c o2 oxidase u-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsV</td>
<td>cytosol c o2 oxidase v-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsW</td>
<td>cytosol c o2 oxidase w-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsX</td>
<td>cytosol c o2 oxidase x-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsY</td>
<td>cytosol c o2 oxidase y-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>109282</td>
<td>hpsZ</td>
<td>cytosol c o2 oxidase z-type subunit</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

78
Table 3.4. Number of expressed CAZy genes in each subdivision 2 acidobacteria genome.

<table>
<thead>
<tr>
<th>CAZy Gene Family</th>
<th>HF_2_M10</th>
<th>HF_2_M21</th>
<th>HF_2_M3</th>
<th>HF_2_M4</th>
<th>HF_2_M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA_3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA_5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA_6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>AA_9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CBM13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CBM22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CBM6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CBM48</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>CBM56</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CBM67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CBM69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CE_1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CE_10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CE_11</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>CE_14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CE_15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CE_4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>CE_5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CE_8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CE_9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GH_109</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GH_110</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GH_11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_128</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_130</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GH_13_11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH_13_15</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GH_13_16</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH_13_21</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GH_13_67</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GH_13_82</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GH_13_90</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_13_100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_16</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_18</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GH_20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GH_22</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>GH_25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_26</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_32</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_320</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_3207</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GH_321</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_39_14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GH_51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GH_56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_17</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GT_21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GT_28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GT_38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>GT_39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GT_4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GT_40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_54</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_60</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GT_9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PA_3_4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 3.1. Taxonomic distribution in mini- and bulk metagenomes. Taxonomy was assigned to contigs by comparing to the NCBI-nr database using DIAMOND blastp (Buchfink et al., 2014). Taxa counts were determined using MEGAN6 (Huson et al., 2018). The most abundant phyla are displayed, with taxa representing less than 1% of the total community grouped into “Other”.
Figure 3.2. Distribution of acidobacterial clades across the bulk metagenome and mini-metagenome samples. Taxonomic classification was completed using the RDP Classifier on 16S rRNA genes extracted from metagenome and mini-metagenome reads using PhyloFlash. The relative abundance of acidobacterial classes is plotted for each of the four metagenome samples and the four mini-metagenome samples. Each sequencing type (metagenome and mini-metagenome) represents one of the soil horizons (organic or mineral) and one of the temperature treatments (control or heated).
Figure 3.3. Concatenated marker gene tree of acidobacterial MAGs and references. We assessed the MAGs and references from the Gene Taxonomy Database (GTDB) for the presence of 120 marker genes using the GTDB-Toolkit (Parks et al 2018). Acidobacterial subdivisions are displayed next to each clade. The MAGs from this study are colored by soil treatment, with heated samples are in orange text and MAGs from control temperature soils are in blue. The clade of subdivision 2 acidobacteria is highlighted in green. The MAGs from which a full-length 16S rRNA gene could be extracted are marked with an asterisk (*). Genomes from cultured acidobacteria are in bold. Bootstrap support values ≥95% (●) and ≥70% (●).
Figure 3.4. Quality assessment of MAGs and reference MAGs from GTDB. Completeness and contamination was assessed for each MAG and GTDB reference genome (Parks et al 2018) using CheckM marker genes (Parks et al., 2015). All MAGs displayed have a contamination of ≤10%. Sample type is represented by shape where bulk metagenome MAGs are circles, GTDB reference MAGs are triangles, and sorted-MAGs are squares. Points are colored by acidobacterial subdivision, including subdivision 1 in yellow, subdivision 2 in green and subdivision 3 in purple. Names are displayed for the ten high-quality MAGs in subdivisions 1 and 3.
Figure 3.5. Maximum likelihood tree of acidobacterial subdivisions 1,2,3,4,6 and 8 based on full-length 16S rRNA gene. Subdivision is indicated next to grouping. Accession numbers of the 16S rRNA genes in the SILVA database are given in parentheses. Acanthopleuribacter pedis (AB303221) was used as the outgroup. The 16S rRNA genes extracted from Subdivision 2 MAGs are depicted in dark blue. The subdivision 2 clade, including previously sequenced SD2 Acidobacteria, is highlighted in green. Internal nodes with bootstrap values (1000 interactions) of ≥95% (●) and ≥70% (●) are displayed. The scale bar indicates 0.10 changes per nucleotide.
Figure 3.6. PCoA of COG/NOG genes across Acidobacteria subdivisions. The points are colored by acidobacterial subdivision (SDs) where subdivision 1 is presented in yellow, subdivision 2 is green, subdivision 3 is purple and all other subdivisions (4, 6, 8 and 23) are shown in blue. PCoA axis 1 explains 22% of variability and PCoA axis 2 explains 13% of variability.
Figure 3.7 PCoA of CAZy gene families across Acidobacteria subdivisions. Points are colored by acidobacterial subdivisions (SDs) where subdivision 1 is presented in yellow, subdivision 2 is green, subdivision 3 is purple, and sequences from all other subdivisions (4, 6, 8 and 23) are shown in blue. PCoA axis 1 explains 37% of variability, and PCoA axis 2 explains 11% of the variability.
Figure 3.8 Metabolic diagram of the 5 subdivision 2 acidobacteria draft genomes. Metatranscriptome reads mapped to these genomes revealed expression of the Embden-Meyerhof pathway, the pentose phosphate pathway and the tricarboxylic acid cycle of central carbon metabolism. Additionally, MAGs demonstrate expression of dissimilatory and assimilatory nitrate reduction, assimilatory sulfate reduction, GalNAc degradation. The five draft genomes express group 1h/5 [NiFe]-hydrogenases and both high- and low-affinity terminal oxidases involved in oxidative phosphorylation.
CHAPTER 4

HIDDEN DIVERSITY OF SOIL GIANT VIRUSES

4.1 Abstract

Known giant virus diversity is heavily skewed towards viruses isolated from aquatic environments and cultivated in the laboratory. Here, we employed cultivation-independent metagenomics and mini-metagenomics on soils from the Harvard Forest. This lead to the discovery of 16 novel giant viruses, almost exclusively recovered by mini-metagenomics. The new viruses expanded phylogenetic diversity of all giant viruses by more than 20% and represented completely novel lineages or were affiliated with Klosneuviruses, Cafeteria roenbergensis virus or Tupanviruses. One virus had a genome size of 2.4 Mb, the largest ever recorded in the Mimiviridae, while others had genomes encoding up to 80% orphan genes. In addition, more than 240 major capsid proteins were found encoded on unbinned metagenome fragments, further indicating that giant viruses are greatly underexplored in soil ecosystems. The fact that most of these novel viruses evaded detection in the bulk metagenomes suggests mini-metagenomics could be a potent tool to unearth viral giants.

4.2 Introduction

Viruses larger than some cellular organisms and with genomes up to several megabases in size have been discovered in diverse environmental niches across the globe, primarily from aquatic systems, such as fresh-, sea- and wastewater (Letunic & Bork, 2016), with only few from

---

terrestrial environments (Parks et al., 2015) including permafrost (Letunic & Bork, 2016). These viruses are nucleocytoplasmic large DNA viruses (NCDLV), and they infect a wide range of eukaryotes, in particular protists and algae (Aherfi et al., 2016; Andrade et al., 2018). Besides the algae-infecting Phycodnaviridae (Pagnier et al., 2013; Yoosuf et al., 2014), only a few NCDLV have been recovered with their native hosts, such as Cafeteria roenbergensis virus (CroV) in the marine flagellate Cafeteria roenbergensis (Abergel, Legendre, & Claverie, 2015; Boughalmi et al., 2013) and the Bodo saltans virus (BsV) (Abergel et al., 2015; Abrahão et al., 2018; Fischer, 2016; Legendre et al., 2014). Many of the NCDLV are referred to as “giant viruses” based on their large physical size (Fischer, 2016), although the term has also been applied to viruses with a genome size of at least 200 kb (Wilson, Van Etten, & Allen, 2009). Importantly, for many of these NCDLV genome size and particle diameter do not correlate (Fischer, Allen, Wilson, & Suttle, 2010). In the following, the term “giant virus” refers to any member of the monophyletic group of NCLDV, the proposed order Megavirales (Deeg, Chow, & Suttle, 2018).

Most of our current understanding of giant viruses comes from isolates retrieved in co-cultivation with laboratory strains of Acanthamoeba (Claverie & Abergel, 2016). Only recently have the genomes of giant viruses been recovered by approaches such as bulk shotgun metagenomics (Legendre et al., 2014), flow-cytometric sorting (Aherfi et al., 2016; Pagnier et al., 2013) and isolation using a wider range of protist hosts (Aherfi et al., 2016; Pagnier et al., 2013). Recent large-scale marker gene-based environmental surveys (Andreani et al., 2018; Roux et al., 2017; Schulz et al., 2017; Verneau, Levasseur, Raoult, La Scola, & Colson, 2016; W. Zhang et al., 2015) hinted an immense phylogenetic breadth of giant viruses of which, however, only a small fraction has been isolated to date. Possible reasons are challenges in providing a suitable host during co-
cultivation and the inability to recover the viruses together with their native hosts (Khalil et al., 2016; Martínez, Swan, & Wilson, 2014; Wilson et al., 2009). In addition, a systematic recovery of giant virus genomes from metagenomic datasets is lacking and thus, the true genetic diversity of giant viruses remains underexplored. Here we describe, for the first time, giant virus genomes from a forest soil ecosystem that were recovered using a cultivation-independent approach.

4.3 Results

4.3.1 Mini-metagenomics facilitated the discovery of giant viruses

Soil samples from the Harvard Forest were subjected to standard shotgun sequencing of microbial communities. Four of the twenty-eight samples were also analyzed using a ‘mini-metagenomics’ (Bajrai et al., 2016; Khalil et al., 2016; Reteno et al., 2015) approach, where multiple sets of 100 DNA-stained particles were flow sorted and subjected to whole genome amplification and sequencing (Figure 4.1). Metagenomic binning of assembled contigs produced 15 metagenome assembled genomes (MAGs) from the mini-metagenomes and 1 MAG from the bulk metagenomes that displayed features typically found in most NCLDV genomes (Colson, Aherfi, & La Scola, 2017; Hingamp et al., 2013; Mihara et al., 2018), such as hallmark genes encoding for major capsid protein(s) (MCP), factors for maturation of the viral capsid, and packaging ATPases. Furthermore, we observed on most contigs a uniform distribution of genes of viral, bacterial or eukaryotic origin and many without matches in public databases. In addition these new viruses encoded numerous paralogous genes, a feature common to many NCLDV (Halary, Temmam, Raoult, & Desnues, 2016). Many of the duplicated genes were located on different contigs and often unique to the respective genomes, providing additional evidence that these contigs belong to a single viral MAG. Moreover, presence, absence and copy number of nucleocytoplasmic virus orthologous genes (NCVOGs) (Berghuis et al., 2019; McLean et al., 2013b;
Yu et al., 2017) were comparable to previously described giant viruses, suggesting that the MAGs are made up by single viral genomes and several of them being nearly complete. An independently conducted benchmarking experiment of the mini-metagenomics approach revealed that no chimeric contigs are being created during this workflow which further supports the quality of the genomes derived here. Despite the bulk metagenome approach generating five-fold more reads, it only yielded in a single giant virus genome, whereas mini-metagenomics lead to the recovery of 15 additional bins attributable to NCLDV (Figure 4.1). Bulk metagenome reads only mapped to the MAG recovered from bulk metagenomes (at ~9x coverage) and not to any mini-metagenome MAGs, suggesting most of the discovered viruses were of low abundance in the sampled forest soil (Figure 4.1). This was also reflected in the soil metatranscriptomes in which no or only low transcriptional activity of the giant viruses could be detected (Figure 4.1).

4.3.2 Cell-sorted viral particles expand known diversity of NCLDV

The phylogenetic relationships inferred from the tree built from a concatenated alignment of five core NCVOGs (Iyer, Aravind, & Koonin, 2001; Yutin, Wolf, Raoult, & Koonin, 2009) (Figure 4.2) and the consensus of single protein phylogenies showed that newly discovered viruses from forest soil were affiliated with diverse lineages in the Megavirales. Two of the new viruses, Ca. Solivirus and Ca. Solumvirus, were in sister-position to the Pithoviruses, Cedratviruses and the recently isolated Orpheovirus (Filée, 2013; Suhre, 2005). Ca. Sylvanvirus represented a long branch on its own. Most novel soil NCLDV were positioned within the Mimiviridae, a viral family in the Megavirales comprising the Megamimivirinae, the Klosneuvirinae, the algae-infecting Mesomimivirinae and the genus Cafeteriaivirus (Yutin, Wolf, & Koonin, 2014; Yutin et al., 2009) (Figure 4.2). One of the new viruses, Ca. Faunusvirus, grouped with Cafeteria roenbergensis virus and represents the second viral genome sampled in this clade (Figure 4.2). Another novel virus,
Ca. Satyrvirus, branched as sister lineage to the two recently isolated Tupanviruses, which were derived from deep sea and a soda lake samples (Yutin, Wolf, & Koonin, 2014; Yutin et al., 2009), together forming a monophyletic clade in the Megamimivirinae (Figure 4.2). Notably, none of the new lineages were directly affiliated with any of the three other subgroups of well-studied Megamimivirinae (Andreani et al., 2018). Eight of the new viruses branched within the Klosneuvirinae, currently the largest subfamily in the Mimiviridae based on phylogenetic diversity (PD) (Gallot-Lavallée, Blanc, & Claverie, 2017; Schulz et al., 2017) (Figure 4.2).

Strikingly, the addition of the novel giant viruses to the NCLDV tree lead to a 21% increase of the total PD in the Megavirales (Figure 4.2), expanded the diversity of the Mimiviridae by 77% and nearly tripled the PD of the Klosneuvirinae (Figure 4.2). It is important to note that this expansion of PD was from a single study using cultivation-independent techniques, thereby building upon decades of previous giant virus discovery work (Abergel et al., 2015). The fact that all these newly discovered viruses represent distinct lineages in the NCLDV hints that additional sampling is expected to lead to a further substantial increase in giant virus PD.

4.3.3 Genomic features of soil giant viruses

The novel viral genomes assigned to the Klosneuviruses were among the largest ever found (Figure 4.2). With a genome size of up to 2.4 Mb the Ca. Hyperionvirus holds the new record for genome size in the Mimiviridae, dwarfing Klosneuvirus and Tupanvirus with their ~ 1.5 Mb genomes (Colson, La Scola, Levasseur, Caetano-Anollés, & Raoult, 2017; Rodrigues, Mougari, Colson, La Scola, & Abrahão, 2019). Considering that several of the forest soil MAGs are potentially only partially complete, the true genome size of the new viruses might be even larger. Similar to recently discovered Klosneuviruses and Tupanviruses (La Scola et al., 2003), several of the new
viruses affiliated with the Klosnevirinae encode for expanded sets of aminoacyl tRNA synthetases (aaRS), e.g. Ca. Terrestrivirus with up to 19 different aaRS and up to 50 tRNAs with specificity for all 20 different amino acids, a feature only very recently described in the Tupanviruses (Abrahão et al., 2018; Aherfi et al., 2016; Legendre et al., 2015; Rodrigues et al., 2019). In concert with other viral components of the eukaryotic translation system, such viruses likely override host protein biosynthesis using their own enzymes to ensure efficient production of viral proteins. Being less dependent on the host cell machinery makes it conceivable that Klosneviruses might be able to infect multiple hosts, i.e. fewer proteins are necessary to target and interact with alternative hosts. A broader host range has been experimentally verified for Tupanviruses (Abergel et al., 2015; Schulz et al., 2017). While Tupanviruses were able to infect different protists, viral titer did not necessarily increase in all the cases (Abergel et al., 2015; Schulz et al., 2017), suggesting the importance of not yet understood factors necessary for successful host exploitation.

4.3.4 Genome novelty of soil giant viruses

Complementary to the phylogenetic analysis (Figure 4.2), we inferred a gene sharing network to provide further insights into the relationship of the novel viral genomes to known NCLDV lineages based on shared gene content. In agreement with the species tree, viral lineages such as the Mimiviridae, the Marseilleviridae, the Pitho- and Cedratviruses, the Faustoviruses and the Molli- and Pandoraviruses remained well connected (Figure 4.3). Among the novel viruses with the lowest percentage of genes shared with other NCLDV were Ca. Solumvirus and Ca. Solivirus, with Ca. Solivirus being only connected to Orpheovirus and Marseilleviridae and Solumvirus to the Cedratviruses. In contrast to the phylogenetic tree in which Soli- and Solumvirus are affiliated to each other, there was no particular linkage between them in the network. This suggests limited
taxon sampling and we would expect that with discovery of additional giant virus genomes, the phylogenetic position of these viruses will be better resolved.

Another of the soil giant viruses denoted as Ca. Sylvanvirus featured a genome completely disconnected from all other NCLDV (Figure 4.3). With a size of almost 1 Mb it represents one of the largest viral genomes outside Pandoraviruses and the Mimiviridae (Figure 4.3) (Abergel et al., 2015). With the presence of 11 ancestral NCLDV genes, a number similar to several other NCLDV, the Ca. Sylvanvirus genome can be considered near complete. Intriguingly, the vast majority (~80%) of its proteins had neither matches in the NCBI non-redundant (nr) database (Figure 4.3). From the proteins with database hits, 57% had matches to eukaryotes and 27% to bacteria but only 13% to other viruses (Figure 4.3). Importantly, there was no trend in taxonomic affiliation of the hits (Figure 4.3), again emphasizing the lack of any affiliation to known viruses and organisms. Among the identifiable genes were 18 potential kinases, 5 ubiquitin ligases and a histone, all potentially playing important roles in interaction with a currently unknown host.

4.3.5 True diversity of giant viruses in forest soil

The MCPs in the soil bulk metagenomes revealed that the 16 novel viruses represent just the tip of the iceberg in terms of soil giant virus diversity (Figure 4.3). In total, 245 different MCP genes were detected, of which 99% were part of the unbinned metagenome fraction. Most of these MCPs were located on short contigs with a read coverage of below 2, indicating an extremely low abundance of corresponding NCLDV in the respective samples (Figure 4.3). Importantly, none of the bulk-metagenome MCPs matched MCPs from the mini-metagenome-derived MAGs, further underlining the much greater diversity of giant viruses in these samples. MCPs can be heavily duplicated but usually branch together in lineage specific clades enabling taxonomic classification based on their nearest neighbors in the tree (Abergel et al., 2015). Based on identified
phylogenetic relationships it was possible to assign taxonomy to several of the bulk metagenome MCPs, of which most could be attributed to the Klosneuviruses (Figure 4.3). A hint of the true dimension of the NCLDV diversity is revealed when considering that the total number of nearly 300 MCPs discovered in this study, which includes MCPs from all the MAGs, is far greater than the 226 MCPs identified in previously published NCLDV genomes.

4.4 Discussion

Our results illustrate that employing cultivation independent methods on a minute sample from forest soil, a habitat in which giant viruses have rarely been found previously (Abergel et al., 2015), can lead to major discoveries. Recovery of Ca. Solum-, Soli- and Sylvanvirus, three potentially genus, subfamily or even family level NCLDV lineages together with 13 other novel giant virus genomes vastly expands the phylogenetic diversity of the NCLDV and provides novel insights into their genetic makeup. The fact that only a single giant virus MAG was recovered in the bulk metagenomes suggests extremely low abundance of these viruses compared to microbial community members in forest soil. However, mini-metagenomics has proven most effective in recovering these viruses, yet without any detectable traces of host sequences. It is noteworthy that oftentimes the average read coverage of the giant virus MAGs was the highest or among the highest compared to other microbial MAGs derived from the same mini-metagenomes pool of 100 DNA-stained particles. The high coverage and completeness of giant virus genomes is consistent with having several copies of the same viral genome in the same mini-metagenome pool, but the overall low abundance of giant viruses in the system makes it unlikely that several identical viral particles were sorted by chance. A plausible scenario could be that host vacuoles already filled with giant viruses may have been recovered during sorting, thereby delivering several clonal copies of a giant virus genome into a single mini-metagenome pool. This would
enable genome assembly of higher quality and completeness, as previously shown for polyploid bacterial symbionts (Legendre et al., 2014; Wu et al., 2009).

Of the few available studies that have used this mini-metagenomes method, one describes the discovery of a novel intracellular bacterium 30 and another a new group of giant viruses (Legendre et al., 2018), suggesting mini-metagenomics is a promising method for elucidating the hidden diversity of intracellular entities such as giant viruses. As shown by the MCP diversity in the unbinned metagenome fraction many novel giant viruses are readily awaiting discovery. Importantly, the mini-metagenomics approach has not been exhaustively performed in soil or any other ecosystem and thus represents a promising toolkit for exploring the untapped diversity in the giant virus universe.

4.5 Methods

4.5.1 Sampling and sample preparation

Fourteen forest soil cores from the Barre Woods warming experiment located at the Harvard Forest Long-Term Ecological Research site (Petersham, MA) were collected and subsampled into organic horizon and mineral zone, resulting in twenty-eight total samples. Mineral zone samples were flash-frozen while organic horizons were incubated with deuterium oxide for 2 weeks prior to freezing to label the active microbial communities. This incubation was carried out as part of a different experiment that will be addressed in a later manuscript. Total DNA and RNA were extracted from twenty-eight soil samples for bulk metagenomics and metatranscriptomics using the MoBio PowerSoil DNA and RNA kits, respectively. Bacterial and Plant rRNA depletion was performed on the RNA samples prior to sequencing. Of these 28 soil samples, a subset of four encompassing two organic and two mineral layers were selected for
mini-metagenomics. Cells, and presumably viral particles and/or eukaryote vacuoles containing them, were separated from soil particles using a mild detergent, followed by vortexing, centrifugation, and filtration through a 5 μm syringe filter. The filtrates were stained with SYBR Green nucleic acid stain. For each of the four samples, ninety pools containing 100 SYBR+ particles were sorted into microwell plates using Fluorescence Activated Cell Sorting (FACS). Sorted pools underwent lysis and whole genome amplification through Multiple Displacement Amplification (MDA) following methods outlined previously 46. A total of 360 sequencing libraries were generated with the Nextera XT v2 kit (Illumina) with 9 rounds of PCR amplification.

### 4.5.2 Mini-Metagenomes

The 360 libraries derived from sorted particles were sequenced at the DOE Joint Genome Institute (JGI, Walnut Creek, CA) using the Illumina NextSeq platform. Pools of 90 libraries were processed in four sequencing runs that generated 2X150bp read lengths. Raw Illumina reads were quality filtered to remove contamination and low quality reads using BBTools (http://bbtools.jgi.doe.gov, version 37.38). Read normalization was performed using BBNorm (http://bbtools.jgi.doe.gov) and error correction with Tadpole (http://bbtools.jgi.doe.gov). Assembly of filtered, normalized Illumina reads was performed using SPAdes (v3.10.1) (Legendre et al., 2018; Pagnier et al., 2013) with the following options: --phred-offset 33 -t 16 -m 115 --sc -k 25,55,95. All contig ends were then trimmed of 200bp and contigs were discarded if the length was <2kb or read coverage less than 2 using BBMap (http://bbtools.jgi.doe.gov) with the following options: nodisk ambig, filterbycoverage.sh: mincov.
4.5.3 Bulk Metagenomes

Unamplified TruSeq libraries were prepared for the 28 DNA samples for metagenomic sequencing on the Illumina HiSeq-2000 platform at the DOE JGI. Raw Illumina reads were trimmed, quality filtered, and corrected using bfc (version r181) (S. W. Wilhelm, Coy, Gann, Moniruzzaman, & Stough, 2016) with the following options: -1 -s 10g -k 21 -t 10. Following quality filtering, reads were assembled using SPAdes (v3.11.1) (Schulz et al., 2017) with the following options: -m 2000 --only-assembler -k 33,55,77,99,127 --meta -t 32. The entire filtered read set was mapped to the final assembly and coverage information generated using bbmap (http://bbtools.jgi.doe.gov, version 37.62) with default parameters except ambiguous=random. The version of the processing pipeline was jgi_mga_meta_rqle.py, 2.1.0.

4.5.4 Metatranscriptomes

Libraries were prepared and sequenced on the Illumina NextSeq platform at the DOE JGI. Following sequencing, metatranscriptome reads were quality cleaned and a combined assembly was generated using the MEGAHIT assembler (v1.1.2) (Bankevich et al., 2012) using the following options: -m 0.2 --k-list 23,43,63,83,103,123 --continue -o out.megahit --12. These cleaned reads were aligned to metagenome reference sequences using BBMap (http://bbtools.jgi.doe.gov, version 37.38) with the following options: nodisk=true interleaved=true ambiguous=random.

4.5.5 Metagenome Binning

Contigs were organized into genome bins based on tetranucleotide sequence composition with MetaBat2 (Heng Li, 2015). Genome bins were generated for mini-metagenomes without contig coverage patterns due to MDA bias (Bankevich et al., 2012; Nurk, Meleshko, Korobeynikov,
& Pevzner, n.d.). Coverage was determined for the bulk metagenomes by mapping reads to the completed assemblies using the Burrows-Wheeler aligner (D. Li et al., 2015).

4.5.6 Screening for giant viruses

Metagenomic bins were screened for presence of the 20 ancestral NCVOGs (Kang et al., 2015) with hmmsearch (version 3.1b2, hmmer.org). Bins with more than 5 different hits and/or that contained the NCLDV MCP gene (NCVOG0022) were selected and further evaluated (see below).

4.5.7 Annotation and quality control of viral genome bins

Gene calling was performed with GeneMarkS using the virus model (Woyke et al., 2009). For functional annotation proteins were blasted against previously established NCVOGs (H. Li & Durbin, 2009) and the NCBI non redundant database (nr) using Diamond blastp (Yutin et al., 2009) with an e-value cutoff of 1.0e-5. In addition, protein domains were identified by hmmsearch (version 3.1b2, hmmer.org) against Pfam-A (version 29.0) (Borodovsky & Lomsadze, 2011), and tRNAs and introns were identified using tRNAscan-SE (Yutin et al., 2009) and cmsearch from the Infernal package (Buchfink et al., 2014) against the Rfam database (version 13.0) (Finn et al., 2016). Nearly identical sequences within genome bins (>100 bp, identity > 94%) were detected using the MUMmer repeat-match algorithm (Lowe & Chan, 2016) and visualized with Circos (Nawrocki & Eddy, 2013) together with the respective genome bins. For all MAGs, paralogs and best diamond blastp vs NCBI nr hits were visualized with Circos (Kalvari et al., 2018). Furthermore, distribution of read depth across contigs was evaluated and regions with low average coverage were identified.
4.5.8 Experimental benchmarking of the mini-metagenomics approach

Benchmarking of the mini-metagenomics approach to assess potential chimera formation during MDA was performed by randomly sorting 10 cells from a bacterial mock community consisting of five different bacterial isolates; Escherichia coli K12, Echinicola vietnamensis DSM 17526, Shewanella oneidensis MR-1, Pseudomonas putida F1 and Meiothermus ruber. In total 59 of these 10-cell sorts were subject to MDA and sequencing. Resulting reads were filtered, assembled and analyzed with the same bioinformatics pipeline used for the mini-metagenomes generated in this study. Assembly statistics of recovered MAGs were generated with MetaQUAST (Delcher, Salzberg, & Phillippy, 2003).

4.5.9 Computational benchmarking of giant virus metagenomic binning

In addition, benchmarking of the binning workflow was performed to assess its applicability to giant virus data. First, binning of a simulated mock community consisting of 12 giant viruses was tested, each a representative of a subfamily or family in the NCLDV. In addition, the herein newly discovered giant viruses were used as template for a second simulated mock community. In brief, MDA was simulated on the genomes of the mock communities with MDAsim (Krzywinski et al., 2009) (https://github.com/hzi-bifo/mdasim/releases/v2.1.1). In the following, Illumina reads were generated with ART (Krzywinski et al., 2009) and the same bioinformatics pipeline used for the mini-metagenomes in this study employed for read error-correction, normalization, assembly and binning.

4.5.10 Phylogenomics

To remove redundancy, the set of 186 published NCLDV genomes and 16 novel soil giant viruses were clustered at an average nucleotide identity (ANI) of 95% with at least 100 kb aligned
fraction using fastANI (Mikheenko, Saveliev, & Gurevich, 2016) resulting in 132 clusters and singletons. None of the newly discovered viruses clustered with any other virus. The three most incomplete novel giant virus genomes were removed from the data set. To infer the positions of novel soil giant viruses in the NCLDV, five core NCLDV proteins (Tagliavi & Draghici, 2012) were selected: DNA polymerase elongation subunit family B (NCVOG0038), D5-like helicase-primase (NCVOG0023), packaging ATPase (NCVOG0249) and DNA or RNA helicases of superfamily II (NCVOG0076) and Poxvirus Late Transcription Factor VLTF3-like (NCVOG0262), and identified with hmmsearch (version 3.1b2, hmmer.org). Three of the MAGs derived from mini-metagenomes were excluded from the analysis as they had less than three conserved NCLDV proteins. Protein sequences were aligned using mafft (W. Huang, Li, Myers, & Marth, 2012). Gapped columns in alignments (less than 10% sequence information) and columns with low information content were removed from the alignment with trimal (Jain et al., 2018). Phylogenetic trees for each protein and for a concatenated alignment of all five proteins were constructed using IQ-tree with LG+F+R6 as suggested by model test as best-fit substitution model (Yutin et al., 2009). PD was calculated as described previously (Katoh et al., 2002); briefly, the total branch lengths of phylogenetic trees with and without the novel soil NCLDV genomes were calculated and compared to determine the gain in PD.

4.5.11 Major capsid protein analysis

Bulk metagenome assemblies and 186 published NCLDV genomes and 16 soil MAGs were screened for presence of the NCLDV MCP gene (NCVOG0022) (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009) with hmmsearch (version 3.1b2, hmmer.org) and applying a cutoff of 1e-6. This cutoff has been evaluated against ~60,000 available microbial, eukaryotic and non-NCLDV genomes in the Integrated Microbial Genomes database (Nguyen et al., 2015) yielding in only few
false positives. Resulting protein hits were extracted from the metagenome and to reduce redundancy clustered with cd-hit at a sequence similarity of 95% (Wu et al., 2009). Cluster representatives were then subject to diamond blastp (Schulz et al., 2017; Yutin et al., 2009) against nr database (June 2018) and proteins which had hits but no NCLDV MCP in the top 10 were excluded from further analysis as potentially false positives. For tree construction, MCPs were extracted and aligned with mafft-ginsi (--unalignlevel 0.8, --allowshift) (Chen et al., 2016). Gapped columns in the alignment (less than 10% sequence information) were removed with trimal (W. Li & Godzik, 2006) and proteins with less than 50 aligned amino acids were removed. A phylogenetic tree was constructed with IQ-tree and the LG+F+R8 as suggested by model test as the best-fit substitution model (Buchfink et al., 2014).

4.5.12 Gene sharing network

Protein families were inferred with OrthoFinder 1.03 71 on a representative dataset of 93 NCLDV genomes for comparative analysis (after de-replication using 95% ANI clustering (Katoh et al., 2002), details described above, and removal of 36 Poxviruses). For each pair of NCLDV genomes (ANI 95% cluster representatives) the average percentage of proteins in shared orthogroups in relation to the total number of proteins in the respective genome was calculated and used as edge weight in the network. The network was created in Gephi (Capella-Gutierrez et al., 2009) using a force layout and filtered at an edge weight of 18%.
Figure 4.1. Discovery pipeline for soil giant viruses. a Overall workflow. Fourteen forest soil cores from Barre Woods long-term experimental warming site were sub-sampled into organic horizon and mineral zone resulting in twenty-eight total samples. Total DNA and RNA were extracted from twenty-eight soil samples for bulk metagenomics and metatranscriptomics. Of these samples, a subset of four encompassing two organic and two mineral layers were selected for flow-sorted mini-metagenomics. Cells, and presumably viral particles, were separated from soil, stained with SYBR Green nucleic acid stain and sorted using Fluorescence Activated Cell Sorting (FACS). Ninety sorted pools of 100 SYBR+ particles underwent lysis, whole genome amplification, library preparation and sequencing on the Illumina NextSeq platform. Phylogenomic analysis of metagenome assembled genomes (MAGs) facilitated the identification of novel giant viruses. There was no correlation of presence or absence of giant viruses and sample treatment. b Data analysis summary. Fifteen giant virus MAGs (orange circles) were recovered from sorted samples, while only one giant virus MAG (turquoise circle) was recovered from the bulk metagenomes. The other 1778 MAGs from the mini-metagenomes (gray circles) and 1772 MAGs from the bulk metagenomes (gray circles) were of microbial origin and not analyzed further in this study. Mapping of bulk metagenome reads to MAGs revealed ~9X coverage of the bulk-metagenome derived MAG and <1X coverage of MAGs derived from mini-metagenomes, confirming the inability to recover these novel giant virus genomes using bulk metagenomics despite deep sequencing efforts. Assembly and mapping of metatranscriptome data indicated expression of only few of the novel giant virus genes of MAGs derived from mini-metagenomes.
Figure 4.2. Expansion of NCLDV diversity by soil giant viruses. 

a. Phylogenetic tree (IQ-tree LG+F+R6) of NCLDV inferred from a concatenated protein alignment of five core nucleocytoplasmic virus orthologous genes (NCVOGs). The tree was built from a representative set of NCDLV genomes after de-replication by ANI clustering (95% id). Novel soil NCLDV lineages and existing major NCLDV lineages grouping together with soil NCLDV are highlighted in black. The scale bar represents substitutions per site. Branches are collapsed if support was low (<50), filled circles indicate moderate support (50-80, white) or high support (80-97, black), branches without circles are fully supported (>97).

b. Detailed phylogenetic tree of the Mimiviridae. Diameter of filled circles correlates with assembly size and shades of gray with GC% ranging from 20% (light gray) to 60% (dark gray). Bar plots summarize total number of encoded aminoacyl-tRNA synthetases (aaRS) and tRNAs. In addition, completeness was estimated based on number of identified marker genes out of 20 ancestral NCVOGs.

c. Increase of phylogenetic diversity (PD) after adding the soil NCLDV MAGs (black) to representative sets of NCLDV reference genomes (gray).
Figure 4.3. Genome novelty of soil giant viruses. **A** Nucleocyttoplasmic large DNA virus (NCLDV) gene sharing network, with nodes representing genomes, node diameter correlating with genome size, edge diameter and color intensity with normalized percentage of genes in shared gene families between node pairs above a threshold of 18%. **B** Circular representation of the Ca. Sylvanvirus genome. From outside to inside: Blue filled circles depict location of encoded tRNAs. The second ring displays positions of genes (gray) either on the minus or the plus strand. The next track illustrates GC content in gray ranging from 20% (white) to 60% (dark gray). The fourth track shows color-coded origin of proteins with best blastp hits to cellular homologs. Best hits against viral proteins are indicated in white and further broken down based on their taxonomic origin color-coded on the most inner track. Lines in the middle of the plot connect paralogs (gray) and nearly identical repeats (orange). The pie chart in the center of the plot summaries the percentage of genes with and without cellular homologs, which are further broken down based on best blastp hits in the adjacent bar plot. **C** Percentage of genes in NCLDV genes with bacterial or eukaryotic homologs with no blastp hits (in the NCBI NR database, highlighting the unique position of Ca. Sylvanvirus).
Figure 4.4. Hidden diversity of giant viruses in bulk metagenomes. a Total number of major capsid proteins (MCPs) found in reference NCLDV genomes, MAGs or recovered from bulk metagenomes on contigs > 1 kb and contigs < 1 kb (dark gray), colored by taxonomy. b Size and cover of bulk metagenome contigs containing MCP genes, either from the unbinned fraction (filled blue circles) or the MAGs (filled pink circles). c Phylogenetic tree of the MCPs of nucleocytoplasmic large DNA virus (NCLDV). Branches are color-coded based on taxonomic origin of MCPs inferred by relationship in the tree to MCPs of known reference NCLDV. MCPs of novel giant viruses from this study which are not members of the Mimiviridae are indicated in red. Branches labelled with a circle represent novel MCP from MAGs generated in this study while stars indicate MCPs recovered from the unassembled fraction (contigs > 1 kb) of bulk-metagenomes. Circles and stars are filled in color if taxonomy could be assigned based on the tree and in black if it was not possible to assign taxonomy.
Table S1. RDP Classifier results for full-length 16S sequences extracted from acidobacteria MAGs.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Contig</th>
<th>Sequence Orientation</th>
<th>Orientation</th>
<th>%ID</th>
<th>Domain</th>
<th>%ID</th>
<th>Phylum</th>
<th>%ID</th>
<th>Class</th>
<th>%ID</th>
<th>Subclass</th>
<th>%ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF_2_S4</td>
<td>Ga0206966_1374/5913-4397</td>
<td>+</td>
<td>Root</td>
<td>100.00</td>
<td>Bacteria</td>
<td>100.00</td>
<td>Acidobacteria</td>
<td>100.00</td>
<td>Acidobacteria_Gp2</td>
<td>100.00</td>
<td>Gp2</td>
<td>100.00</td>
</tr>
<tr>
<td>HF_2_S1</td>
<td>Ga0206813_1010/1-1427</td>
<td>+</td>
<td>Root</td>
<td>100.00</td>
<td>Bacteria</td>
<td>100.00</td>
<td>Acidobacteria</td>
<td>100.00</td>
<td>Acidobacteria_Gp2</td>
<td>100.00</td>
<td>Gp2</td>
<td>100.00</td>
</tr>
<tr>
<td>HF_2_S2</td>
<td>Ga0207015_1236/876-2409</td>
<td>+</td>
<td>Root</td>
<td>100.00</td>
<td>Bacteria</td>
<td>100.00</td>
<td>Acidobacteria</td>
<td>100.00</td>
<td>Acidobacteria_Gp2</td>
<td>100.00</td>
<td>Gp2</td>
<td>100.00</td>
</tr>
<tr>
<td>HF_2_M3</td>
<td>Ga0210394_10000093/857-1</td>
<td>+</td>
<td>Root</td>
<td>100.00</td>
<td>Bacteria</td>
<td>100.00</td>
<td>Acidobacteria</td>
<td>100.00</td>
<td>Acidobacteria_Gp2</td>
<td>100.00</td>
<td>Gp2</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Figure S1. Full concatenated marker gene tree of acidobacterial MAGs and references. Genome sequences and MAGs were assessed for the presence of 120 marker genes in the Genome Taxonomy Database (GTDB; Parks et al., 2018). Acidobacterial subdivisions are displayed next to each clade, with subdivision 2 highlighted in green. The MAGs from which a full-length 16S rRNA gene could be extracted are marked with an asterisk (*). Genomes from cultured acidobacteria are in bold. Bootstrap support values ≥95% (●) and ≥70% (●).
Figure S2. Genomic similarity across the 5 subdivision 2 acidobacteria characterized in this study. Average nucleotide identity (ANI, in blue) and average amino acid identity (AAI, in green) of subdivision 2 MAGs from this study. ANI and AAI were calculated using the ANI/AAI matrix calculator tool by (L. Rodriguez-R & Konstantinidis, 2016) and visualized using the ggplot2 package in R.
**Figure S3.** Genomic similarity across subdivision 1 acidobacteria. Average nucleotide identity (ANI, in blue) and average amino acid identity (AAI, in green) across subdivision 1 acidobacteria MAGs from this study, and reference genomes from isolates. ANI and AAI were calculated using the ANI/AAI matrix calculator tool by Kostantinidis et al. and visualized using the ggplot2 package in R.
Figure S4. PCoA clustering COG/NOG categories across acidobacteria MAGs and reference genomes. All points are labeled with the genome name. Acidobacterial subdivisions (Sds) are represented by color, with subdivision 1 genomes and MAGs in yellow, subdivision 2 MAGs in green, subdivision 2 genomes in purple and all other subdivisions in blue.
Figure S5. Heatmap of EggNOG genes across Acidobacteria subdivisions. Acidobacterial subdivisions (SDs) are labeled on the x-axis of the heatmap. Black represents the presence of the gene in the genome, and white represent the absence of the gene.
Figure S6. PcoA clustering CAZy categories across acidobacteria MAGs and reference genomes. All points are labeled with the genome name. Acidobacterial subdivisions (Sds) are represented by colors with subdivision 1 MAGs and reference genomes in yellow, subdivision 2 MAGs in green, subdivision 3 genomes in purple, and all other subdivisions in blue.
Figure S7. Heatmap of CAZy gene families across acidobacterial subdivisions. Acidobacterial subdivisions (SDs) are labeled on the x-axis. Names on the y-axis display the annotated gene families from the CAZy database. The clustering in the second y-axis is based on hierarchical clustering of the genomes. Black color in the heatmap represents the presence of the gene, and white represents the absence of the gene.
Figure S8. Relative abundance of CAZy gene families across acidobacteria MAGs and reference genome sequences. Reference genomes and MAGs are displayed on the x-axis and the relative abundance of each gene family in each genome. The legend displays each of the CAZy gene families, including Auxiliary Activities (AA in dark green), Carbohydrate Binding Modules (CBM in light green), Carbohydrate Esterases (CE in orange), Glycoside Hydrolases (GH in light blue), Glycosyl Transferases (GT in purple) and Polysaccharide Lyases (PL in yellow).
Figure S9. Phylogenetic tree of group 1h/5 [NiFe] hydrogenases. Maximum likelihood tree based on the deduced amino acid sequence (~600 amino acid positions) of the group 1h/5 [NiFe]-hydrogenase large subunit. The tree was bootstrapped 1,000 times and the consensus support is displayed ≥99% (●) and ≥90% (●). The outgroup (not shown) was the group 1 [NiFe]-hydrogenase from Desulfovibrio vulgaris (WP_010939204.1). The scale bar depicts 0.01 changes per amino acid. The yellow box indicates the clade of Acidobacteria. Sequences in red represent those genes from acidobacterial MAGs generated in this study.
Figure S10. Glycoside hydrolases expressed across subdivision 2 acidobacteria MAGs. The x-axis includes the five, high-quality acidobacterial MAGs classified as subdivision 2. The y-axis contains each of the CAZy gene families expressed in the genomes of these subdivision 2 acidobacteria, and the second y-axis shows the pattern of hierarchical clustering based on the genes within the genomes. Purple color represents that the gene family is expressed in the subdivision 2 acidobacterial genome, and white represents the lack of expression of that gene family.
Figure S11. Relative abundance of CAZy gene families expressed across subdivision 2 MAGs. The x-axis displays the five high-quality subdivision 2 acidobacterial MAGs generated in this study. The y-axis displays the relative abundance of each CAZy gene family in each genome. CAZy gene families are represented in the legend as Auxiliary Activities (AA in dark green), Carbohydrate Binding Modules (CBM in light green), Carbohydrate Esterases (CE in orange), Glycoside Hydrolases (GH in light blue), Glycosyl Transferases (GT in purple) and Polysaccharide Lyases (PL in yellow).
Figure S12. Distribution of acidobacterial classes in the mineral and organic soil horizons. Taxonomic classification was completed using the RDP Classifier on 16S rRNA genes extracted from metagenome and mini-metagenome reads using PhyloFlash. The x-axis describes the soil horizon (Mineral or Organic) from which the soil samples were taken. The y-axis displays the relative abundance of each acidobacterial class.
BIBLIOGRAPHY


Barns, S. M., Cain, E. C., Sommerville, L., & Kuske, C. R. (2007). Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within


Gupta, R. S., Bhandari, V., & Naushad, H. S. (2012). Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide


139


Pagnier, I., Yutin, N., Croce, O., Makarova, K. S., Wolf, Y. I., Benamar, S., ... La Scola, B. (2015). Babela massiliensis, a representative of a widespread bacterial phylum with unusual
adaptations to parasitism in amoebae. *Biology Direct, 10*(1), 13.
https://doi.org/10.1186/s13062-015-0043-z

https://doi.org/10.1016/j.tim.2017.02.015


https://doi.org/10.1038/nbt.4229

https://doi.org/10.1101/gr.186072.114

https://doi.org/10.1038/s41564-017-0012-7


https://doi.org/10.1016/S0005-2728(01)00169-4


https://doi.org/10.1038/ismej.2011.20


