



Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil

L. V. Alteio,^a  F. Schulz,^b R. Seshadri,^b N. Varghese,^b W. Rodriguez-Reillo,^a E. Ryan,^b D. Goudeau,^b  S. A. Eichorst,^c R. R. Malmstrom,^b R. M. Bowers,^b  L. A. Katz,^{a,d} J. L. Blanchard,^e  T. Woyke^b

^aGraduate Program in Organismic and Evolutionary Biology, University of Massachusetts Amherst, Amherst, Massachusetts, USA

^bDepartment of Energy, Joint Genome Institute, Berkeley, California, USA

^cCentre for Microbiology and Environmental Systems Science, Department of Microbiology and Ecosystem Science, Division of Microbial Ecology, University of Vienna, Vienna, Austria

^dDepartment of Biological Sciences, Smith College, Northampton, Massachusetts, USA

^eDepartment of Biology, University of Massachusetts Amherst, Amherst, Massachusetts, USA

ABSTRACT Soil ecosystems harbor diverse microorganisms and yet remain only partially characterized as neither single-cell sequencing nor whole-community sequencing offers a complete picture of these complex communities. Thus, the genetic and metabolic potential of this “uncultivated majority” remains underexplored. To address these challenges, we applied a pooled-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). The sorted and bulk MAGs increased the known phylogenetic diversity of soil taxa by 7.2% with respect to the Joint Genome Institute IMG/M database and showed clade-specific sequence recruitment patterns across diverse terrestrial soil metagenomes. Additionally, sorted-MAGs expanded the rare biosphere not captured through MAGs from bulk sequences, exemplified through phylogenetic and functional analyses of members of the phylum *Bacteroidetes*. Analysis of 67 *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.

IMPORTANCE Microbial ecologists have historically used cultivation-based approaches as well as amplicon sequencing and shotgun metagenomics to characterize microbial diversity in soil. However, challenges persist in the study of microbial diversity, including the recalcitrance of the majority of microorganisms to laboratory cultivation and limited sequence assembly from highly complex samples. The uncultivated majority thus remains a reservoir of untapped genetic diversity. To address some of the challenges associated with bulk metagenomics as well as low throughput of single-cell genomics, we applied flow cytometry-enabled mini-metagenomics to capture expanded microbial diversity from forest soil and compare it to soil bulk metagenomics. Our resulting data from this pooled-cell sorting approach combined with bulk metagenomics revealed increased phylogenetic diversity through novel soil taxa and rare biosphere members. In-depth analysis of genomes within the highly represented *Bacteroidetes* phylum provided insights into conserved and clade-specific patterns of carbon metabolism.


KEYWORDS flow cytometry, metagenomics, microbial ecology, soil microbiology

Citation Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, Goudeau D, Eichorst SA, Malmstrom RR, Bowers RM, Katz LA, Blanchard JL, Woyke T. 2020. Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* 5:e00768-19. <https://doi.org/10.1128/mSystems.00768-19>.

Editor Janet K. Jansson, Pacific Northwest National Laboratory

Copyright © 2020 Alteio et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to T. Woyke, twoyke@lbl.gov.

 Complementary metagenomic approaches provide a window into microbial diversity in Harvard forest soil.

Received 13 November 2019

Accepted 18 February 2020

Published 10 March 2020

Soil is considered to be among the most biologically diverse ecosystem types, and yet much of its microbial diversity remains poorly characterized (see, e.g., references 1 and 2). Each gram of soil is estimated to harbor 1,000 to 1,000,000 different bacterial species (see, e.g., references 3 to 7). Investigating soil microorganisms *in situ* is challenging due to the heterogeneous nature of the soil environment (see, e.g., references 8 to 10). As a result, terrestrial habitats remain immense reservoirs of untapped genetic and metabolic diversity (7, 11) encoded within microbial communities that drive important ecosystem-level processes, including nitrogen cycling and carbon dioxide flux (12–14). Soils are regarded as critical for global health, as they contain 3,000 Pg of carbon and have the potential to act as either a carbon source or a carbon sink, which is important to consider under conditions of climatic shift (15, 16). 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 (17, 18) with only a minute fraction of estimated bacterial diversity being successfully cultivated. Substantial efforts are being made to develop innovative cultivation techniques, including the ichip and droplet-based sorting coupled with laboratory cultivation (17, 19). These approaches have contributed to expansion of diversity within novel families. However, cultivation-independent investigations may further our understanding of microbial diversity by facilitating description of novel higher taxonomic ranks. Thus, challenges associated with direct study of soil microorganisms have yielded a large knowledge gap regarding terrestrial microbial diversity. Due to limitations associated with cultivation, relatively few isolate genomes are available as references for soil microbes (20). From the publicly available Integrated Microbial Genomics (IMG/M) database (21), 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. However, with soil estimated to contain 1,000 to 1,000,000 species per gram (9), these references represent only a small percentage of soil microbes.

In addition to culture-based approaches, amplicon studies have greatly contributed to our knowledge of microbial community structure (1, 22). However, amplicon sequencing primers that target the small-subunit (SSU) rRNA gene may not adequately amplify some organisms due to primer biases through mismatches (22). Additionally, estimates of organismal abundance may be conflated by variation in gene copy number (23). Phylogenetically divergent taxa may be overlooked using PCR-based approaches, thereby hampering our ability to describe an expanded diversity of organisms (22). High-throughput sequencing technologies combined with novel metagenome binning algorithms (24, 25) enable genome-resolved metagenomics studies and have greatly expanded the availability of reference genomes from uncultured taxa by circumventing challenges associated with cultivation- and amplicon-based studies (11, 26, 27). The more recent applications of directly sequencing DNA from soil microbial communities allow 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 (28–30), and population heterogeneity can hamper the efficiency of assembly, even of abundant microorganisms (31).

Population microheterogeneity of closely related strains within microbial communities makes the separation of individual strains challenging (32). Soils are typically dominated by a small set of highly abundant taxa (12), and the rare biosphere may therefore be overlooked in metagenomic studies despite playing an important role in soil biogeochemical processes (33). Lastly, bulk metagenomics can also include extracellular DNA from dead microorganisms, which may be abundant in the environment. The presence of this exogenous DNA has the potential to inflate estimates of diversity and genomic potential (34–36) and to 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 (36).

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 (37–39). 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 coupled to shotgun sequencing, termed mini-metagenomics, to forest soils collected from the Barre Woods soil warming experiment at the Harvard Forest Long-Term Ecological Research (LTER) site. This mini-metagenomic approach separates a researcher-defined number of cells from the larger community, which then undergo lysis and whole-genome multiple-displacement amplification (MDA), followed by sequencing.

Prior to the application of cell sorting to Harvard Forest soil in this study and in that by Schulz et al. (40), mini-metagenomics analysis of microorganisms had been used only in aqueous environments, including hot springs, hospital sink biofilms, and activated sludge (40–44). 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 susceptibility to the single-cell lysis step. The use of two overlapping metagenomic methods may allow us to capture a broader taxonomic diversity than the use of only one approach 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 (41). Hence, we evaluated this method as a tool to complement bulk metagenomics in uncovering the “microbial dark matter” in soil.

Here, we combined mini-metagenomics and bulk metagenomics as complementary approaches for capturing a more holistic perspective of microbial community diversity. We discovered additional diversity of uncultivated microorganisms in a forest soil microbial community and thus contribute to the known diversity of both major soil clades and understudied taxonomic groups, which can be used as reference sequences in future studies. Additionally, we provide an example of how the mini-metagenomics and bulk metagenomic approaches can be used in complement to investigate potential metabolism and ecological roles of microorganisms. 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 the known phylogenetic diversity (PD) of soil clades. Our data suggest that the sorted-MAGs represent some of the diversity of previously unsequenced organisms that are challenging to access using bulk approaches, offering insights into the functional potential of soil dark matter.

RESULTS AND DISCUSSION

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 totals of 1.2 Gbp and 1.3 Gbp, respectively (Fig. 1; see also Table S1 in the supplemental material).

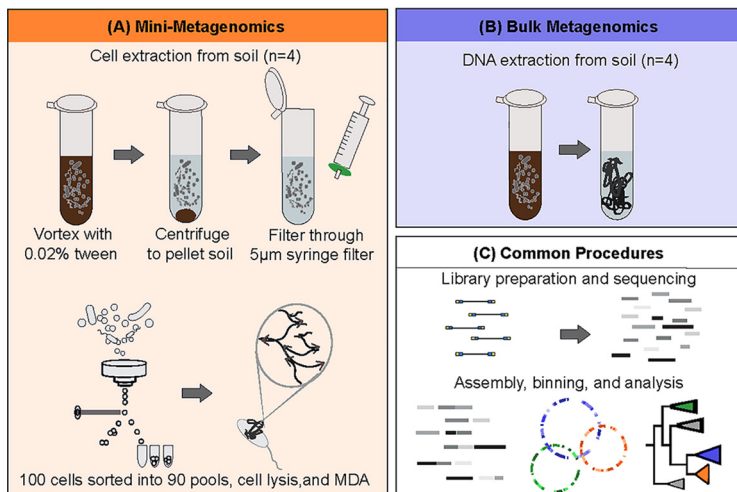


FIG 1 Overview of mini-metagenome and bulk metagenome approaches used in this study. (A) Mini-metagenomics performed on four soil samples, including one heated sample from the top organic soil, one heated sample from the lower mineral soil, one control organic sample, and one control mineral sample ($n = 4$). Cells were separated from soil particles using a mild detergent, followed by vortex mixing, centrifugation, and filtration through a 5- μm -pore-size 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 Materials and Methods.

Binning of assembled contigs produced 1,793 mini-metagenome assembled genomes (sorted-MAGs) and 275 bulk metagenome MAGs (Fig. 2; see also Fig. S1 in the supplemental material). Following CheckM quality assessment (45), 200 sorted-MAGs and 29 bulk MAGs surpassed completeness thresholds of $\geq 50\%$ complete, $\leq 10\%$

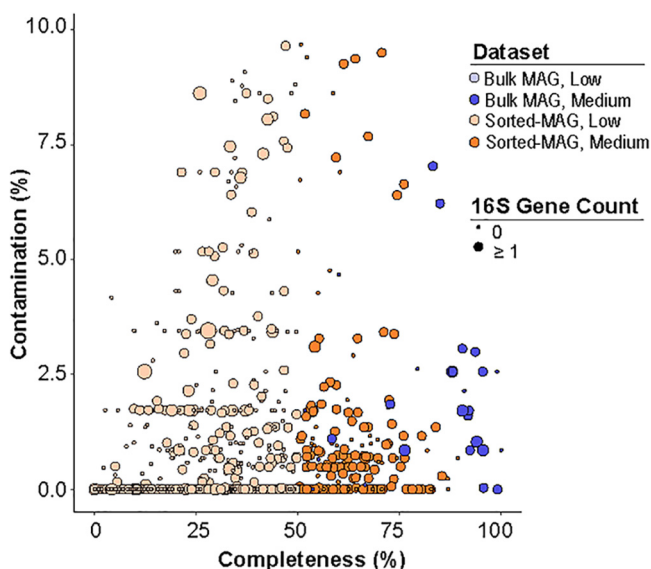


FIG 2 Assessment of sorted-MAG and MAG quality. Sorted-MAGs (orange, $n = 1,793$) and bulk MAGs from the four samples corresponding to those sorted with FACS (blue, $n = 275$) are represented. Medium-quality sorted-MAGs (dark orange, $n = 200$) and MAGs (dark blue, $n = 29$) are those with $\geq 50\%$ completeness, $\leq 10\%$ contamination, and $\leq 10\%$ strain heterogeneity based on analysis of CheckM marker genes (43). The size of each circle represents the number of 16S rRNA gene copies within each MAG.

contamination, and $\leq 10\%$ strain heterogeneity. We considered MAGs with less than 50% completeness to represent “low quality” based on MIMAG standards (46) and excluded them from additional analyses (Fig. 2; see also Fig. S1). Overall, quality filtering removed lower-quality sorted-MAGs on the basis of completeness, whereas bulk MAGs were removed due to higher degrees of contamination and strain-level heterogeneity. Assessment of MAG quality using CheckM showed average percent completeness of 81.5% in medium-quality bulk metagenome MAGs ($n = 29$), which was higher than the 61.9% seen with the medium-quality sorted-MAGs ($n = 200$; $P = 3.29 \times 10^{-7}$) (Fig. 2; see also Fig. S1). Assessed for marker gene contamination, bulk metagenome MAGs revealed an average estimated level of contamination of 1.92%, indicating an estimated level of contamination higher than the average of 0.98% contamination in the sorted-MAGs ($P = 0.01117$) (Fig. 2; see also Fig. S1). Analysis of strain-level heterogeneity across medium-quality MAGs and sorted-MAGs revealed a lower degree of multiple strain contamination in sorted-MAGs than in bulk MAGs as assessed by CheckM (45). The average level of strain heterogeneity for the bulk MAGs was 1.16%, compared to 0.04% in the sorted-MAGs ($P = 3.89 \times 10^{-6}$; Table S2). This decrease in strain heterogeneity seen using mini-metagenomics indicates that sorted-MAGs collapse fewer strains into a single MAG.

As one measure to compare mini-metagenomics and bulk metagenomics methods, we assessed GC content and found averages of 49.2% GC and 60.5% GC in sorted-MAGs and MAGs, respectively (Fig. S1; see also Table S2). Variation in GC content can be attributed to known biases in the single-cell workflow such as susceptibility of cells to sorting and lysis (37, 47), as well as amplification bias introduced during MDA (48). The cell isolation method used in mini-metagenomics reduces inflation of community diversity as a result of exogenous DNA. 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. Taking the data together, mini-metagenomics and bulk metagenomics generated a large number of quality MAGs that can be used as complementary data sets in genome-resolved studies to investigate broad microbial diversity.

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 (Fig. 3A) (21). 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 MAGs and isolate genomes from IMG/M (Fig. 3A; see also Fig. S2). 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-metagenome and bulk metagenomics revealed a broad diversity of soil bacteria, as well as a few archaeal taxa, and demonstrated the complementarity of these approaches for biological discovery. The sorted-MAGs expanded the known diversity of the taxa which were previously found to be abundant and ubiquitous across soil types (49), as well as of the taxa considered part of the rare biosphere that may still be widespread but remain at relatively low abundances in microbial communities (33). The more abundant taxa represented by the sorted-MAGs include *Bacteroidetes* ($n = 48$) and *Verrucomicrobia* ($n = 8$), while the

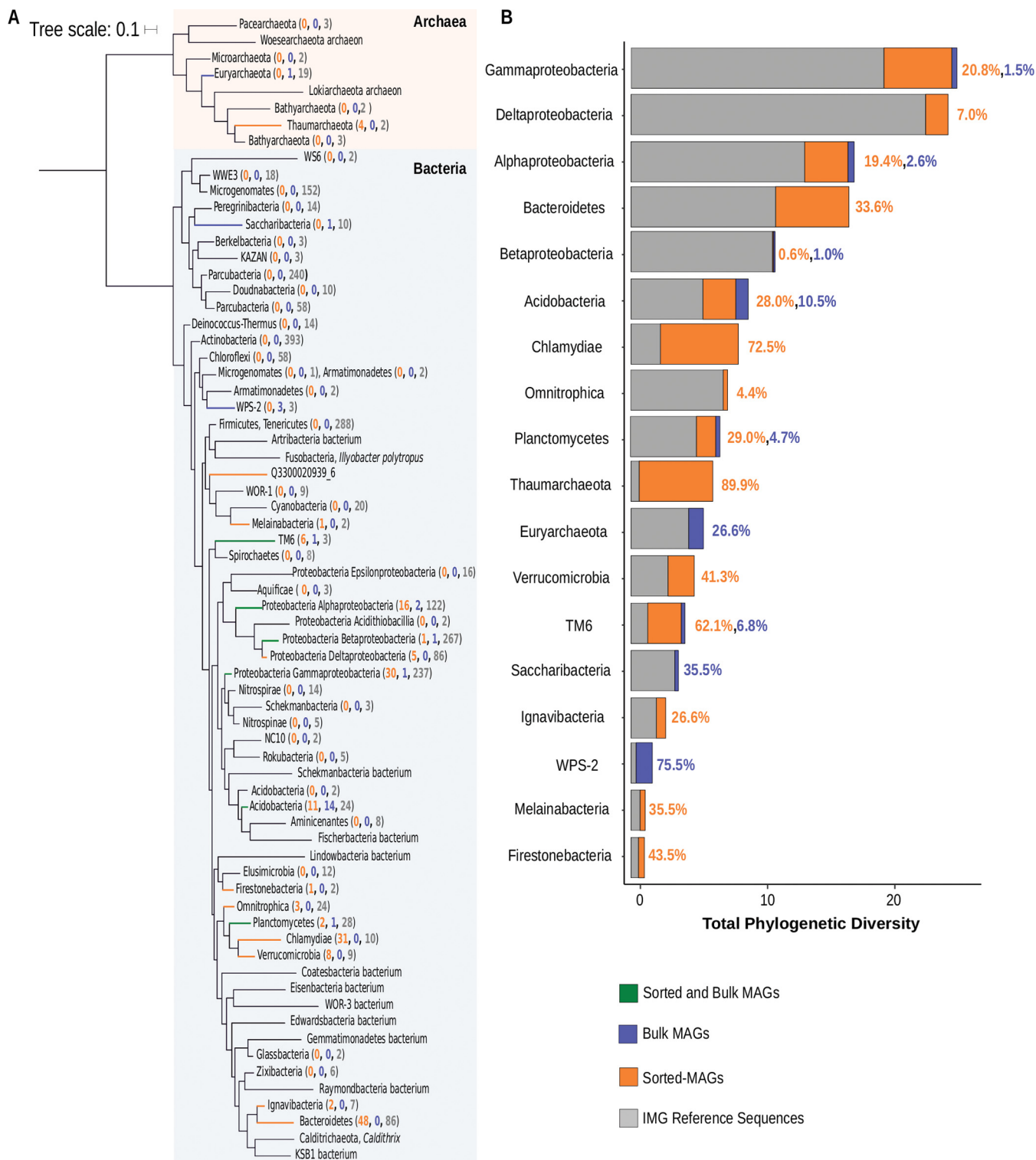


FIG 3 Phylogenetic diversity of soil taxa identified in this study. (A) Maximum likelihood tree of the phylogenetic distribution of medium-quality sorted-MAGs and bulk MAGs in the context of previously sequenced soil taxa. Colored branches represent clades that include sorted-MAGs and/or bulk MAGs. Orange branches include only sorted-MAGs, blue branches include only bulk MAGs, and green branches include both mini-metagenome and bulk MAGs. Numbers in orange represent numbers of contributed sorted-MAGs, blue numbers represent bulk MAGs, and gray numbers represent the number of reference sequences in each clade. (B) Phylogenetic diversity expansion through sorted-MAGs 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 total branch length from bulk MAGs. The percentage of increase in phylogenetic diversity from this study is shown next to each bar.

taxa with typically lower abundances in soils included *Thaumarchaeota* ($n = 4$), *Omni-trophica* ($n = 3$), *Ignavibacteria* ($n = 2$), *Melainabacteria* ($n = 1$), and *Firestonebacteria* ($n = 1$) (1, 49). Interestingly, numerous sorted-MAGs belonged to phyla typically comprised of pathogens and endosymbionts such as the *Chlamydiae* ($n = 31$) and *Gammaproteobacteria*, specifically within the order *Legionellales* ($n = 30$), as well as *TM6* ($n = 7$) (50–53) (Fig. 3A; see also Fig. S3). Genomes in the phylum *Chlamydiae* and in the order *Legionellales* within the phylum *Gammaproteobacteria* are considered entirely intracellular (54, 55). The phyla identified by sorted-MAGs represented abundant taxa found in previous soil community studies (1, 49, 56) in addition to the rare biosphere, demonstrating the utility of mini-metagenomics for expanding diversity beyond abundant soil taxa (Fig. 3; see also Fig. S3).

As for the bulk MAGs, some of these belonged to rare taxa not recovered through mini-metagenomics, including *WPS-2* ($n = 3$), *Euryarchaeota* ($n = 1$), and *Saccharibacteria* ($n = 1$). We assessed phylogenetic diversity (PD), the total amount of branch length contributed by sequences of interest within a phylogenetic tree, in the sorted-MAGs to determine the contribution of this single study to the known range of 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 (Fig. 3B). Mini-metagenomes expanded the range of available evidence not only of phylogenetic diversity within clades of known soil bacteria and archaea but also of 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 those of groups notoriously recalcitrant to laboratory cultivation, such as *TM6* (62.1%), *Verrucomicrobia* (41.3%), and *Acidobacteria* (28.0%) (42, 57). 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 (58). 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 (40), as these viruses are most often associated with eukaryotic host cells (59). The hypothesized liberation of these intracellular bacteria makes mini-metagenomic sequencing a useful tool for investigating the diversity and evolution of the intracellular life strategy (55, 60).

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* (6.8%), and *Alphaproteobacteria* (2.6%). Even in clades where more bulk-derived genomes were added than sorted-MAGs, such as in *Acidobacteria*, the sorted-MAGs were phylogenetically more diverse. These calculated increases in phylogenetic diversity with the addition of MAGs from this study are limited with regard to scope, as not all available reference sequences are publicly accessible in the IMG/M database. However, this database is updated monthly with newly uploaded sequences from GenBank (21).

Complementarity of mini-metagenomics and bulk metagenome sequencing.

Mini-metagenomics has not been widely applied in soils to date and will serve as a valuable tool for expanding our knowledge of soil biodiversity. In this study, we applied both bulk metagenomics and mini-metagenomics to compare analyses of complex community samples as well as to identify the advantages and disadvantages of each. This approach is capable of generating higher-quality MAGs than bulk metagenomics due to the reduction of strain-level microheterogeneity when selected pools of cells are sequenced (32). Although they are lower in estimated genome completeness than bulk MAGs, sorted-MAGs from soil also demonstrate a lower degree of strain heterogeneity, indicating that fewer genomic fragments from multiple organisms have been collapsed

into a single genome bin (45) (Table S2). The sorted-MAG reduced genome completeness is, at least in part, a likely result of uneven whole-genome amplification (WGA), as has been extensively reported in single-cell genomic studies (47). The larger number of sorted-MAGs presents opportunities for improved resolution for taxonomic classification and for genome-informed investigations of microbial metabolism and linking the potential metabolism to processes at the ecosystem level. Taxonomic classification of organisms using high-quality MAGs has become a critical approach for expanding knowledge of microbial diversity, given that we currently lack information for the majority of uncultivated organisms (61). Finally, although not applied in this study, FACS-based sample processing may be modified to achieve cell and/or population separation that is more highly targeted (62), thereby further expanding the utility of mini-metagenomics to detect microbial dark matter.

Although the mini-metagenomics approach produced a greater number of medium-quality genome bins than bulk metagenomics, this approach is not without challenges. In comparison to bulk metagenomics, the requirements associated with 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 and GC-based genome amplification skew, likely introduce bias during sample processing. The formation of extracellular polysaccharides is a strategy widely used by microorganisms to protect against changes in the environment, as well as for exchange of nutrients and materials (63). These matrices may support the maintenance of stable microbial consortia and cellular adhesion to soil particles (63). These larger aggregate structures are subject to exclusion in sample preparation steps, including filtration, prior to FACS. Methodological challenges such as these may be reflected in our data, where organisms which are typically abundant in forest soils, such as *Actinobacteria*, *Chloroflexi*, and *Firmicutes* (49), were present in low numbers using mini-metagenomics compared to traditional bulk metagenomics (Fig. 3; see also Fig. S3). 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 all were therefore excluded (Fig. 3; see also Fig. S3). An alternative DNA amplification method, termed WGA-X, has been developed which improves cell lysis and amplification of high-GC-content organisms over MDA (48). With this improved method of DNA amplification, more extensively representative mini-metagenomic sampling might be possible.

Bulk metagenomics presents fewer opportunities to introduce bias and may more accurately capture the total soil community than the mini-metagenomic approach. Using bulk metagenomics, DNA from the total soil sample is extracted and sequenced, which circumvents cell and particle size selection introduced via FACS. Thus, bulk metagenomics remains an invaluable tool for understanding the diversity of microbial communities, particularly that of the dominant soil microorganisms. Sorted-MAGs, however, provided additional genomic data covering broader phylogenetic diversity compared to the bulk MAGs, further enhancing biological discovery. The scientific question of interest should guide the selection of one approach over the other. We support the use of both approaches in complement to one another in order to capture the broadest scope of soil microbial diversity.

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 (Fig. 4; see also Fig. S2). We assessed these sorted-MAGs and bulk MAGs in the context of broader terrestrial community studies by comparing them

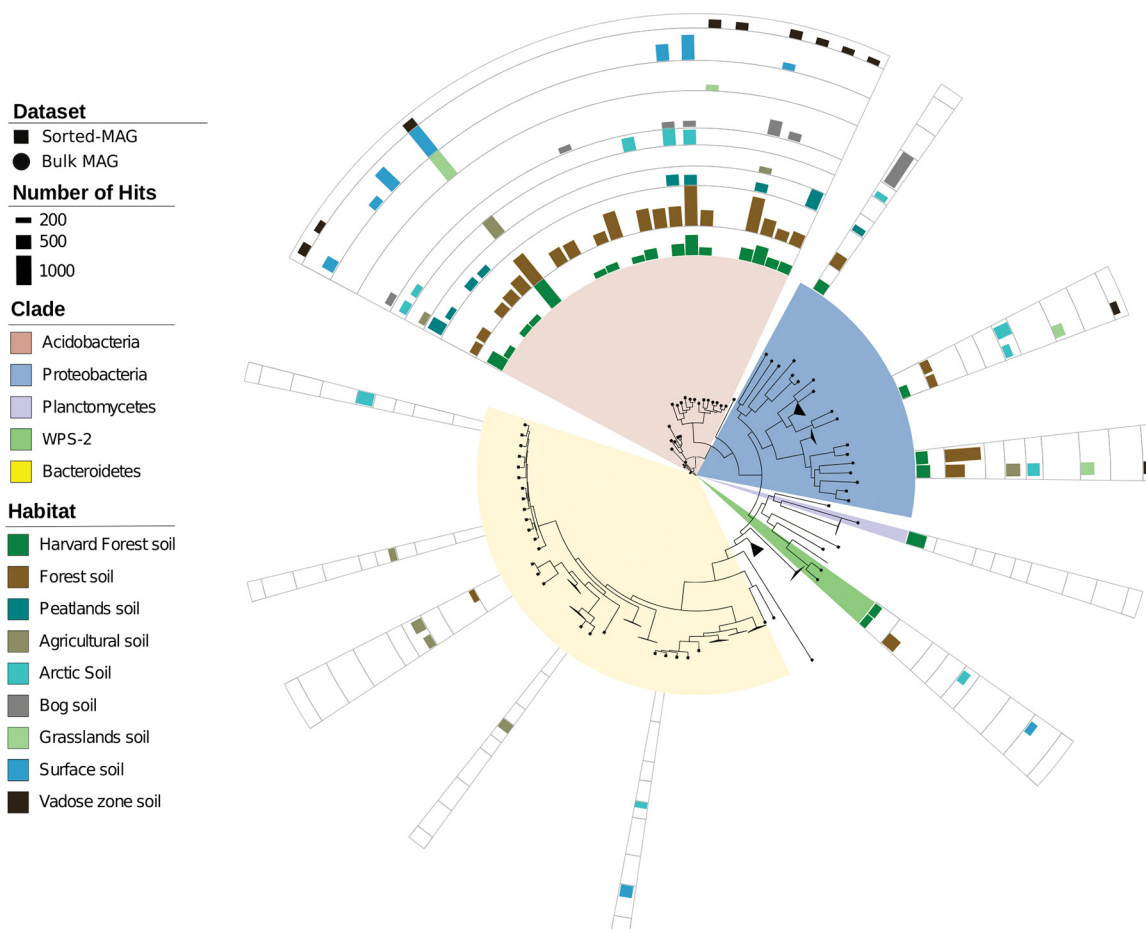


FIG 4 Comparison of MAGs from this study with published data from terrestrial metagenomes. Innermost is a maximum likelihood tree based on a concatenated alignment of 56 conserved marker proteins from medium-quality sorted-MAGs and bulk MAGs recovered in this study. Mini-metagenomes and bulk MAGs were 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 sorted-MAGs and bulk MAGs by metagenome samples arising from each terrestrial habitat type as specified in the legend. The height of the bar chart indicates the total number of sorted-MAGs and bulk MAG coding sequences that matched metagenome samples. The MAGs were considered matches if they had a minimum of 200 coding sequences with hits at $\geq 95\%$ amino acid identity over 70% alignment lengths to CDS of an individual metagenome. Further details are provided in Materials and Methods, and data corresponding to this figure are provided in Table S3. The figure was rendered using iTOL (96).

against 2,210 metagenomes from the 80 terrestrial studies using LAST (64) (Fig. 4; see also Table S3). We defined highly represented sorted-MAGs and MAGs as those with at least 200 protein coding sequences with hits to metagenome samples at $\geq 95\%$ amino acid identity (AAI) over a 70% alignment length (65, 66).

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) (Fig. 4; see also Table S3). 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 had not been found in previous terrestrial metagenome studies.

The phylum *Acidobacteria* was the most abundant phylum represented in the bulk MAGs (77%) and unbinned metagenome data (32%), compared to the sorted-MAGs (8.5%) (Fig. S3). In contrast, the phylum *Bacteroidetes* was highly represented by the sorted-MAGs (55.5%), compared to the bulk metagenome MAGs (0.1%) and unbinned

metagenome data (3.8%) (Fig. S3). The sorted-MAGs in the phylum *Bacteroidetes* increased the phylogenetic diversity of this group by 33.6% (Fig. 3) and appeared to be novel as they had a relatively low number of matches to protein coding sequences from publicly available soil metagenomes, with only 6 of 67 *Bacteroidetes* MAGs having similarity of at least 200 coding sequences with published soil metagenomes (Fig. 4). This presumed novelty could also contribute to computation challenges associated with sequence assembly, as only the most abundant taxa are overrepresented in public databases (29). And yet many of these sorted and bulk MAGs were not represented in previous Harvard Forest metagenomes (Fig. 3). Taking the data together, the low level of 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 phylogenetically diverse samples.

Biological insights into carbon metabolism in soil *Bacteroidetes*. *Bacteroidetes* spp. make up ~10% to the total microbial community in soils (1), and yet most of our knowledge about members of this phylum stems from sequenced isolates from vertebrate guts and aquatic habitats (67–69). Bacteria in the phylum *Bacteroidetes* are known to be important degraders of polysaccharides; however, little is known about the role of this abundant group in soils. Given the relatively small body of work on soil *Bacteroidetes* and the substantial contribution of 67 putatively novel sorted-MAGs from this study to phylogenetic diversity estimates (Fig. 3; see also Fig. 5), we further explored these sorted-MAGs from *Bacteroidetes* to gain insight into their physiological potential and assess functional similarities to previously known *Bacteroidetes*.

The genome sizes of the sorted-MAGs ranged from 1.6 to 5 Mb (Table S4), which is within the range of previously reported *Bacteroidetes* genome sizes of from 0.9 Mb (*Cardinium* endosymbiont) (70) to 9.1 Mb (*Chitinophaga pinensis*) (71). The finding of 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 (Fig. 5; see also Fig. S4) (45). These sorted-MAGs were distributed across three distinct families, including *Cytophagaceae*, *Chitinophagaceae*, and *Sphingobacteriaceae*, as well as a clade of unclassified sorted-MAGs (Fig. 5). *Bacteroidetes* are known to have a large set of genes that encode enzymes for carbohydrate degradation (69), including a broad array of glycoside hydrolases that are phylogenetically conserved (72). The distribution of CAZy gene families across these *Bacteroidetes* taxa exhibited clade-specific abundance patterns of glycoside hydrolases, glycosyl transferases, and carbohydrate binding modules (Fig. 5; see also Table S4) (73).

Sorted-MAGs within the *Cytophagaceae* family appeared to be specialized for polymeric carbon degradation, namely, degradation of cellulose, as they encode proteins in glycoside hydrolase family 5 which exhibit endocellulase activity (74, 75). 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 use cellulose, hemicellulose, and chitin. Seventeen of the 27 sorted-MAGs in the *Chitinophagaceae* family contained at least one chitinase in glycoside hydrolase family 18 or 19 (76) along with cellulases in glycoside hydrolase families 5, 8, and 9 and glycoside hydrolases in family 43 that may degrade hemicellulose and pectin (77) (Fig. 5; see also Fig. S5). In support of this conjecture, the sequenced genome of *Chitinophaga pinensis* (a member of the *Chitinophagaceae* family) contains genes to degrade leaf matter and fungal structures, suggesting its ability to degrade both cellulose and chitin (78). 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. 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

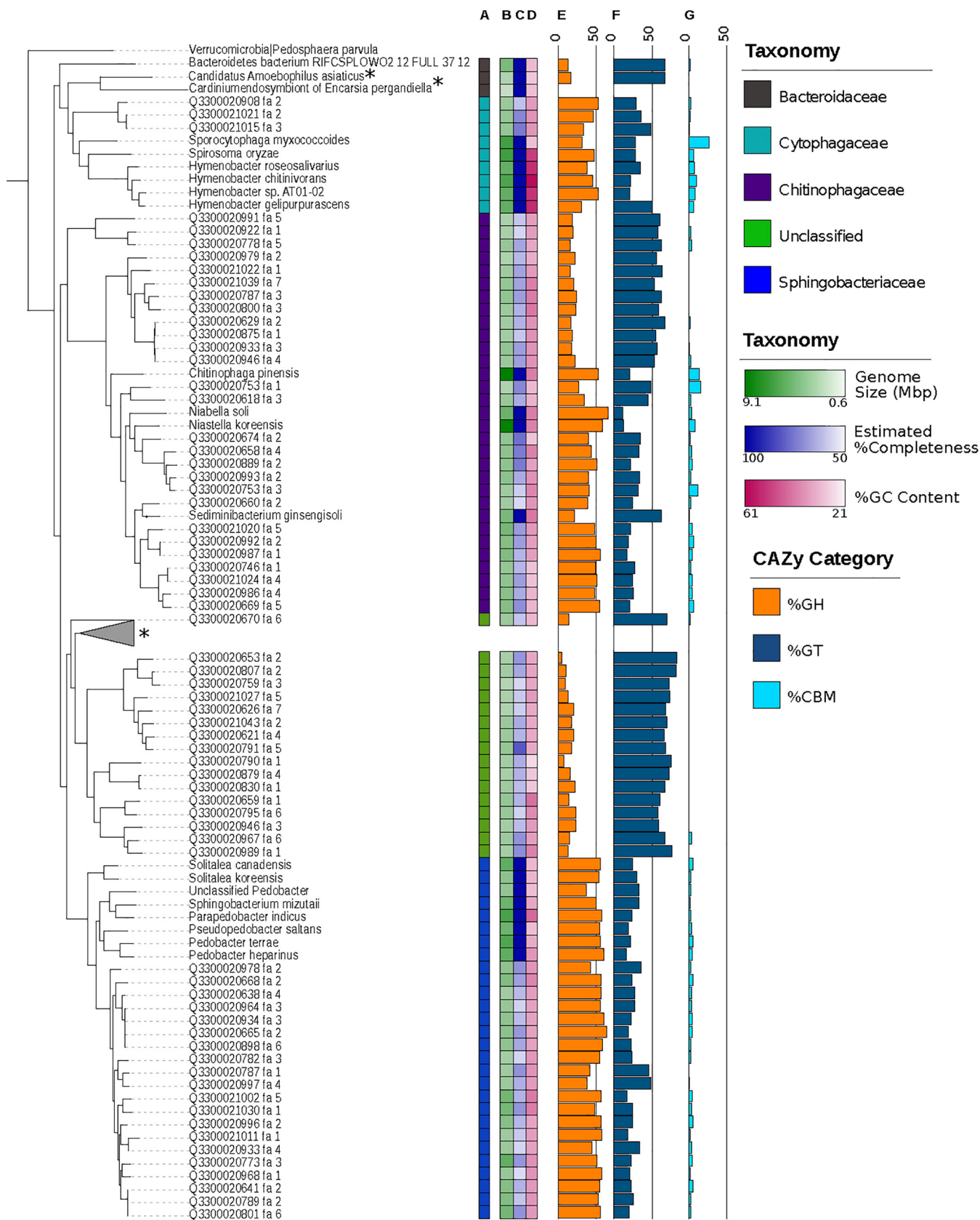


FIG 5 Insights into carbon metabolism within the phylum *Bacteroidetes*. A concatenated marker gene tree of 67 *Bacteroidetes* sorted-MAGs and 70 *Bacteroidetes* reference sequences from the IMG/M database shows clade-specific abundances of glycoside hydrolase and glycosyl transferases. The tree is rooted with *Pedospaera parvula* (phylum *Verrucomicrobia*). Column A shows the distribution of sorted-MAGs across three families of *Bacteroidetes*, including *Cytophagaceae*, (Continued on next page)

potential metabolic flexibility. This is consistent with previous investigations describing the family *Sphingobacteriaceae* as capable of degrading diverse polysaccharides (79).

Putatively novel *Bacteroidetes* sorted-MAGs stemming from experimental warming plots at the Harvard Forest Long-Term Ecological Research site spanned three different taxonomic families and harbored an extensive diversity of enzyme families, including those involved in hydrolysis of polymeric chitin, cellulose, and hemicellulose substrates. The genomic potential to utilize these labile carbon compounds is consistent with previous metagenomic investigations in soils of warmed plots (16, 80). Interestingly, the number of identified carbohydrate active enzyme genes increased with genome size for each of the six CAZy categories (Fig. 5; see also Fig. S4), illustrating that these organisms accumulated the capacity to degrade various carbohydrates, thereby expanding their niche for carbohydrate utilization in soil. And yet 17 sorted-MAGs belonged to an unclassified clade of *Bacteroidetes* spp. which were depleted in glycoside hydrolases and carbohydrate binding modules but retained a high number of glycosyl transferases (Fig. 5; see also Fig. S5), suggesting a limited role for these organisms in substrate decomposition. Rather, the relatively higher abundance of glycosyl transferase genes involved in the formation of glycosidic 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.

To further support the role of the *Bacteroidetes* in polymeric carbon degradation in soils, we investigated specific carbohydrate degradation using the KEGG database (81, 82) and predicted the completeness of metabolic pathways using KEGG-Decoder (83). The majority of sorted-MAGs in *Sphingobacteriaceae* and *Chitinophagaceae* have nearly complete pathways coding for alpha-amylase, beta-glucosidase, chitinase, and diacetylchitobiose deacetylase activity, further supporting the idea of a role of these organisms as generalists in polysaccharide degradation (Fig. S6). Additionally, seven of the sorted-MAGs within *Sphingobacteriaceae* contain nearly complete pathways for pullulanase. Consistent with analysis of carbohydrate degradation potential using the CAZy database (Fig. S5), 22 of the sorted-MAGs were found to contain only one complete pathway or no complete pathways for polymeric carbohydrate degradation (Fig. S6). This limited potential for carbohydrate utilization does not correlate with decreased genome completeness (Fig. S4). Rather, we hypothesize that these sorted-MAGs have an alternative survival strategy in the soil environment similar to those exhibited by other *Bacteroidetes*, including "*Candidatus* Amoebophilus asiaticus" (84), *Cardinium* sp. (85), "*Candidatus* Sulcia muelleri" (86), and *Blattabacterium* sp. (87), which are known symbionts (Fig. 5).

Similarly to known symbionts, the estimated GC contents of unclassified sorted-MAGs in this study were low relative to those of other *Bacteroidetes* sequences, with an average of 39.97% GC (88). These unclassified *Bacteroidetes* demonstrate limited ability for carbon utilization and reduced central carbon metabolism and chemotaxis (Fig. S6) while retaining genome sizes of 2.4 Mb on average, which are comparable to those of *Bacteroidetes* previously identified as host-associated species (Fig. 5; see also Fig. S5 and S6). 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 (87–89). The abundance of unclassified *Bacteroidetes* within this study may represent further evidence of the liberation of symbionts from host cells and vacuoles prior to FACS. Alternatively, the

FIG 5 Legend (Continued)

Chitinophagaceae, and *Sphingobacteriaceae*, and a clade of unclassified sorted-MAGs. Column B shows genome sizes, with the darkest color representing the largest genome of 9.1 megabases and the lightest representing a genome size of 0.6 megabases. Column C shows genome completeness based on CheckM marker genes, ranging from 50% to 80.5%, as a color gradient. Reference sequences represent isolates with complete genomes. Column D presents genome GC content as a color gradient that ranges from 21.13% to 61.24%. In columns E to G, percentages of genes annotated as glycoside hydrolases (column E), glycosyl transferases (column F), and carbohydrate binding modules (column G) are illustrated as bar charts with vertical lines denoting 0% and 50% of annotated genes. *Bacteroidetes* with known symbiotic relationships are indicated with an asterisk. The collapsed clade contains *Sulcia muelleri*, a known symbiont of sap-feeding insects, and *Blattabacterium* sp., a known symbiont of the cockroach *Blattella germanica*.

relatively low carbohydrate degradation potential of sorted-MAGs within the unclassified clade may be indicative of an opportunistic life strategy (74).

Conclusions. This application of mini-metagenomics 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 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 in low-abundance taxa. Taken together, the mini-metagenomics and bulk metagenomics approaches allow us to probe deeper into microbial diversity and function within heterogeneous environments beyond soil.

MATERIALS AND METHODS

Sample collection and incubation. Soils were collected on the 24th of 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, USA. This site consists of two 30-by-30-m plots: one which has remained at ambient soil temperature and one that has been artificially warmed since 2002 using heating cables buried at 10-cm depth (90). Soil respiration, nitrogen mineralization, and vegetation cover and growth as well as soil and litter chemistry have been measured over the course of the long-term experiment. The canopy overstory is dominated by paper birch and black birch (*Betula papyrifera* and *B. lenta*, respectively), red maple (*Acer rubrum*), black oak and red oak (*Quercus velutina* and *Q. rubra*, respectively), and American beech (*Fagus grandifolia*) (56).

Two intact soil cores were taken from subplots within the larger 30-by-30-m experimental plots, including a subplot within heated plot 2 and a subplot within control plot 12. The subplots included in this study were chosen at random. The collected soil cores 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 2-mm-pore-size mesh, resulting in a total of 4 individual soil samples.

Both treatments (heated and control) and soil horizons (organic and mineral) were represented by these four soil samples. Approximately 5 g of soil was immediately frozen in a dry ice/ethanol bath for DNA extraction and was then transported to the University of Massachusetts Amherst for storage at -80°C . Approximately 15 g of soil was transferred to a 50-ml Falcon tube for transportation on ice to the Joint Genome Institute (JGI) in Walnut Creek, CA, USA. Samples were further processed as described previously Schulz et al. (40). The study was limited to four soil samples in order to maintain the cost-effectiveness and overall efficiency of the techniques applied.

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 vortex mixing performed for 5 min. Samples were centrifuged for 5 min at $500 \times g$ to pellet large soil particles. Following centrifugation, the supernatant was filtered through a 5- μm -pore-size syringe filter to remove the remaining soil particulates. Samples were diluted 1:100 in phosphate-buffered saline (PBS) and stained with SYBR green (Thermo Fisher Scientific, Waltham, MA, USA). For each of the four soil samples, 90 pools of 100 SYBR-positive (SYBR⁺) cells were sorted into microwell plates using a BD Influx cell sorter (BD Biosciences, San Jose, CA, USA) to perform FACS. Sorted pools underwent cell lysis and whole-genome amplification using a Qiagen RepliG single-cell kit for multiple-displacement amplification (MDA) (Qiagen, Hilden, Germany). A total of 360 libraries were generated for sequencing with a Nextera XT v2 kit (Illumina, San Diego, CA, USA) with 9 rounds of PCR amplification.

Mini-metagenomes. Following library preparation, the 360 mini-metagenome libraries were sequenced on an Illumina NextSeq platform (Illumina, San Diego, CA, USA) at the DOE Joint Genome Institute (JGI; Walnut Creek, CA, USA). Pools of 90 libraries were processed in four sequencing runs with 2×150 -bp read lengths. Raw Illumina reads were quality filtered to remove contamination and low-quality reads using BBTools (v37.38) (91), resulting in 359 mini-metagenomes for downstream analysis, as one mini-metagenome did not pass quality filtering standards. Read normalization was performed using BBNorm (91), and error correction was conducted using Tadpole (91). Assembly of filtered, normalized Illumina reads was completed using SPAdes (v3.10.1) (92) with the following options: `-phred-offset 33 -t 16 -m 115 -sc -k 25,55,95`. All contig ends were trimmed of 200 bp, and contigs were discarded if the length was <2 kb or the level of read coverage was less than 2 using BMap (91) with the following options: `nodisk ambig, filterbycoverage.sh: mincov`.

Bulk metagenomes. Total DNA was extracted from ~ 0.25 g of soil using a DNeasy PowerSoil DNA extraction kit (Qiagen, Hilden, Germany). Extracted DNA was assessed using a Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and Qubit (Thermo Fisher Scientific, Waltham, MA, USA). Unamplified TruSeq libraries were prepared for 4 DNA samples prior to sequencing on an Illumina HiSeq-2000 platform (Illumina, San Diego, CA, USA) 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) (92) 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 BMap (v37.62) (91) with default parameters except `ambiguous=random`. The version of the processing pipeline used was `jgi_mga_meta_rqc.py`.

2.1.0. Of the 28 metagenome samples sequenced, only 4 were selected for inclusion in analysis for this study because they corresponded to those samples sorted using FACS.

Genome binning and quality assessment. Assembled contigs from the bulk and mini-metagenomes were binned into MAGs and sorted-MAGs based on tetranucleotide frequency using MetaBat2 (93). Sorted-MAGs were generated for mini-metagenomes without contig coverage patterns due to MDA bias. Genome bins were assessed for estimated completeness and estimated contamination marker genes included in the CheckM (45). Bulk metagenome MAGs and sorted-MAGs were filtered to $\geq 50\%$ completeness, $\leq 10\%$ contamination, and $\leq 10\%$ strain heterogeneity to retain medium-quality sorted-MAGs and bulk metagenome MAGs for downstream analysis (46). Following quality filtering, 200 medium-quality sorted-MAGs and 29 medium-quality bulk metagenome MAGs were used for phylogenomic analysis, metagenomic recruitment, and investigation of metabolic potential.

Phylogenetic tree construction and phylogenetic diversity. A concatenated marker gene phylogenetic tree was constructed for 200 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 (41, 92) was identified with *hmmsearch* (v3.1b2) (94) and specific hidden Markov models (HMMs) for each of the markers. For every marker protein, alignments were built with MAFFT (v7.294b) (95) and subsequently trimmed with BMGE using BLOSUM30 (96). MAGs and reference sequences were clustered at 95% average nucleotide identity with FastANI v1.0 (97), resulting in 170 sorted-MAGs, 25 bulk MAGs, and 2,341 reference sequences with distinct taxonomic classifications. Single-protein alignments were then concatenated, and a phylogenetic tree was inferred with FastTree2 using the options `-spr 4 -mlacc 2 -slownni -lg` (98) and was visualized using iTol (99).

The contribution of sorted-MAGs and bulk MAGs to phylogenetic diversity was determined by calculating the sum of the total branch lengths of the contributed genomes relative to the reference genomes (100). Total branch length was calculated for a phylogenetic tree containing only 2,341 bacterial and archaeal reference sequences from IMG/M (21). We then calculated the additional total branch lengths contributed by sorted-MAGs and bulk MAGs. The percentage of 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 (101). BLAST results were imported into MEGAN6 (102) for taxonomic assignment. The relative abundance of each phylum was computed and visualized in R using *ggplot2* (103).

Protein recruitment. Sorted-MAGs ($n = 200$) and bulk MAGs ($n = 29$) were dereplicated 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 (64) (Fig. 4; see also Table S3 in the supplemental material). Individual sorted-MAGs and MAGs were designated a match to metagenome samples if the following criteria were met: a minimum of 200 coding DNA sequences (CDS) with hits at $\geq 95\%$ 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 95% amino acid identity cutoff was chosen based on a study reported previously by Luo et al. (65), who asserted that organisms grouped at the “species” level typically show $>85\%$ AAI among themselves. Since our data set included divergent sublineages, 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 (Fig. 4; see also Table S4), and the results were plotted as a multibar chart in iTol (99).

Bacteroidetes phylogeny and metabolic predictions. A maximum likelihood tree for *Bacteroidetes* was constructed using IQTree (104) for the 67 sorted-MAGs and soil *Bacteroidetes* references from IMG/M. The tree was rooted with *Pedospaera parvula* in the phylum *Verrucomicrobia*. Family-level taxonomic classification and genome size and genome size based on CheckM marker gene assessment (45) were visualized using iTol (99). Functional annotation for sorted-MAGs was assigned using the Carbohydrate Active Enzyme (CAZy) database (73) implemented in dbCAN2 (105). The percentage of total annotated genes assigned to each gene family was calculated and is displayed in a multibar chart in iTol (99).

Additional metabolic annotations were assigned to the 67 *Bacteroidetes* sorted-MAGs using the GhostKoala server (82). Following annotation of protein coding genes, assigned knockouts (KOs) were used to estimate the completeness of selected pathways using KEGG-Decoder and a heat map was generated using “static” visualization mode to depict the completeness of each pathway (83).

Data availability. The bacterial and archaeal MAG data sets generated and analyzed in this study were deposited at NCBI GenBank under BioProject accession number PRJNA608716 and at https://bitbucket.org/lvalteio/forest_soil_mags_and_sortedmags/src, together with sequence alignments and phylogenetic trees generated in this study. Metagenomes and their corresponding metadata are available at IMG/M (<https://img.jgi.doe.gov/m>) under the taxon OIDs (identification numbers) indicated in Table S1.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.1 MB.

FIG S2, PDF file, 1.1 MB.

FIG S3, PDF file, 0.1 MB.

FIG S4, PDF file, 0.2 MB.

FIG S5, PDF file, 2.3 MB.

FIG S6, PDF file, 2.3 MB.

TABLE S1, CSV file, 0.03 MB.

TABLE S2, CSV file, 0.3 MB.

TABLE S3, CSV file, 0.7 MB.

TABLE S4, CSV file, 0.3 MB.

ACKNOWLEDGMENTS

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under contract no. DE-AC02-05CH11231. This research was further supported through a Facilities Integrating Collaborations for User Science (FICUS) grant (503125), a DOE Office of Science Graduate Student Research (SCGSR) Fellowship to L.V.A., and the Austrian Science Fund (FWF; project grant P26392-B20). The Harvard Forest experimental warming plots are supported by NSF grants DEB 1237491 (Long-Term Ecological Research) and DEB 1456528 (Long-Term Research in Environmental Biology). This material is based upon work supported by the National Institute of Food and Agriculture (NIFA), the U.S. Department of Agriculture (USDA), the Center for Agriculture, Food, and the Environment, and the Biology Department at the University of Massachusetts Amherst, under project number MAS00032.

The contents of this article are solely our responsibility and do not necessarily represent the official views of the USDA or NIFA.

We thank the Division of Computational Systems Biology for providing and maintaining the Life Science Compute Cluster (LiSC) at the University of Vienna. Additionally, we thank Alexander Truchon for assistance with sample collection and contributions to initial data analysis.

REFERENCES

- Fierer N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* 15:579–590. <https://doi.org/10.1038/nrmicro.2017.87>.
- Solden L, Lloyd K, Wrighton K. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr Opin Microbiol* 31:217–226. <https://doi.org/10.1016/j.mib.2016.04.020>.
- Amann R, Rosselló-Móra R. 2016. After all, only millions? *mBio* 7:e00999-16. <https://doi.org/10.1128/mBio.00999-16>.
- Gans J, Wolinsky M, Dunbar J. 2005. Microbiology: computational improvements reveal great bacterial diversity and high toxicity in soil. *Science* 309:1387–1390. <https://doi.org/10.1126/science.1112665>.
- Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* 113:5970–5975. <https://doi.org/10.1073/pnas.1521291113>.
- Lennon JT, Locey KJ. 2016. The underestimation of global microbial diversity. *mBio* 7:e01298-16. <https://doi.org/10.1128/mBio.01298-16>.
- Torsvik V, Øvreås L. 2002. Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* 5:240–245. [https://doi.org/10.1016/s1369-5274\(02\)00324-7](https://doi.org/10.1016/s1369-5274(02)00324-7).
- Lombard N, Prestat E, van Elsas JD, Simonet P. 2011. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol Ecol* 78:31–49. <https://doi.org/10.1111/j.1574-6941.2011.01140.x>.
- Urlich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* 3:e2527. <https://doi.org/10.1371/journal.pone.0002527>.
- Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. 2013. Micro-scale determinants of bacterial diversity in soil. *FEMS Microbiol Rev* 37:936–954. <https://doi.org/10.1111/1574-6976.12023>.
- Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, Brunel D, Frostegård Å, Heulin T, Jansson JK, Jurkevitch E, Kruus KL, Kowalchuk GA, Lagares A, Lappin-Scott HM, Lemanceau P, Le Paslier D, Mandic-Mulec I, Murrell JC, Myrold DD, Nalin R, Nannipieri P, Neufeld JD, O'Gara F, Parnell JJ, Pühler A, Pylro V, Ramos JL, Roesch LFW, Schlöter M, Schleper C, Szczyrba A, Sessitsch A, Sjöling S, Sørensen J, Sørensen SJ, Tebbe CC, Topp E, Tsiamis G, van Elsas JD, van Keulen G, Widmer F, Wagner M, Zhang T, Zhang X, Zhao L, Zhu Y-G, Vogel TM, Simonet P. 10 February 2016, posting date. Back to the future of soil metagenomics. *Front Microbiol* <https://doi.org/10.3389/fmicb.2016.00073>.
- Delgado-Baquerizo M, Maestre FT, Reich PB, Jeffries TC, Gaitan JJ, Encinar D, Berdugo M, Campbell CD, Singh BK. 2016. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nat Commun* 7:10541. <https://doi.org/10.1038/ncomms10541>.
- Graham EB, Knelman JE, Schindlbacher A, Siciliano S, Breulmann M, Yannarell A, Beman JM, Abell G, Philippot L, Prosser J, Foulquier A, Yuste JC, Glanville HC, Jones DL, Angel R, Salminen J, Newton RJ, Bürgmann H, Ingram LJ, Hamer U, Siljanen HMP, Peltoniemi K, Potthast K, Baneras L, Hartmann M, Banerjee S, Yu RQ, Nogaró G, Richter A, Koranda M, Castle SC, Goberna M, Song B, Chatterjee A, Nunes OC, Lopes AR, Cao Y, Kaisermann A, Hallin S, Strickland MS, Garcia-Pausas J, Barba J, Kang H, Isobe K, Papaspyrou S, Pastorelli R, Lagomarsino A, Lindström ES, Basiliko N, Nemergut DR. 24 February 2016, posting date. Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes? *Front Microbiol* <https://doi.org/10.3389/fmicb.2016.00214>.
- Lladó S, López-Mondéjar R, Baldrian P. 2017. Forest soil bacteria: diversity, involvement in ecosystem processes, and response to global change. *Microbiol Mol Biol Rev* 81:e00063-16. <https://doi.org/10.1128/MMBR.00063-16>.
- Hicks Pries CE, Castanha C, Porras RC, Torn MS. 2017. The whole-soil carbon flux in response to warming. *Science* 355:1420–1423. <https://doi.org/10.1126/science.aal1319>.
- Zhou J, Xue K, Xie J, Deng Y, Wu L, Cheng X, Fei S, Deng S, He Z, Van Nostrand JD, Luo Y. 2012. Microbial mediation of carbon-cycle feedbacks to climate warming. *Nat Clim Chang* 2:106–110. <https://doi.org/10.1038/nclimate1331>.
- Overmann J, Abt B, Sikorski J. 2017. Present and future of culturing

- bacteria. *Annu Rev Microbiol* 71:711–730. <https://doi.org/10.1146/annurev-micro-090816-093449>.
18. Pham VHT, Kim J. 2012. Cultivation of unculturable soil bacteria. *Trends Biotechnol* 30:475–484. <https://doi.org/10.1016/j.tibtech.2012.05.007>.
 19. Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K, Epstein SS. 2010. Use of icip for high-throughput in situ cultivation of “uncultivable microbial species. *Appl Environ Microbiol* 76:2445–2450. <https://doi.org/10.1128/AEM.01754-09>.
 20. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, Flater J, Tiedje JM, Hofmocker KS, Gelder B, Howe A. 2017. Strategies to improve reference databases for soil microbiomes. *ISME J* 11:829–834. <https://doi.org/10.1038/ismej.2016.168>.
 21. Chen I-M, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Hunt-emann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* 47:D666–D677. <https://doi.org/10.1093/nar/gky901>.
 22. Eloe-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC. 2016. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol* 1:15032. <https://doi.org/10.1038/nmicrobiol.2015.32>.
 23. Louca S, Doebeli M, Parfrey LW. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* 6:41. <https://doi.org/10.1186/s40168-018-0420-9>.
 24. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 31:533–538. <https://doi.org/10.1038/nbt.2579>.
 25. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337:1661–1665. <https://doi.org/10.1126/science.1224041>.
 26. Castelle CJ, Banfield JF. 8 March 2018, posting date. Major new microbial groups expand diversity and alter our understanding of the tree of life. <https://doi.org/10.1016/j.cell.2018.02.016>.
 27. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hensdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 11 April 2016, posting date. A new view of the tree of life. *Nat Microbiol* <https://doi.org/10.1038/nmicrobiol.2016.48>.
 28. Delmont TO, Eren AM, Maccario L, Prestat E, Esen ÖC, Pelletier E, Le Paslier D, Simonet P, Vogel TM. 2015. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Front Microbiol* 6:358. <https://doi.org/10.3389/fmicb.2015.00358>.
 29. Nayfach S, Pollard KS. 2016. Toward accurate and quantitative comparative metagenomics. *Cell* 166:1103–1116. <https://doi.org/10.1016/j.cell.2016.08.007>.
 30. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2:1533–1542. <https://doi.org/10.1038/s41564-017-0012-7>.
 31. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvočiūtė M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu Y-W, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin H-H, Liao Y-C, et al. 2017. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat Methods* 14:1063–1071. <https://doi.org/10.1038/nmeth.4458>.
 32. Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4:8. <https://doi.org/10.1186/s40168-016-0154-5>.
 33. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, Küsel K, Rillig MC, Rivett DW, Salles JF, van der Heijden MGA, Youssef NH, Zhang X, Wei Z, Hol W. 2017. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J* 11:853–862. <https://doi.org/10.1038/ismej.2016.174>.
 34. Nagler M, Insam H, Pietramellara G, Ascher-Jenull J. 2018. Extracellular DNA in natural environments: features, relevance and applications. *Appl Microbiol Biotechnol* 102:6343–6356. <https://doi.org/10.1007/s00253-018-9120-4>.
 35. Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol* 2:16242. <https://doi.org/10.1038/nmicrobiol.2016.242>.
 36. Lennon JT, Muscarella ME, Placella SA, Lehmkuhl BK. 2017. How, when, and where relic DNA biases estimates of microbial diversity. *bioRxiv* <https://www.biorxiv.org/content/10.1101/131284v4>.
 37. Blainey PC. 2013. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* 37:407–427. <https://doi.org/10.1111/1574-6976.12015>.
 38. Stepanauskas R. 2012. Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* 15:613–620. <https://doi.org/10.1016/j.mib.2012.09.001>.
 39. Woyke T, Doud DFR, Schulz F. 2017. The trajectory of microbial single-cell sequencing. *Nat Methods* 14:1045–1054. <https://doi.org/10.1038/nmeth.4469>.
 40. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, Blanchard J, Woyke T. 2018. Hidden diversity of soil giant viruses. *Nat Commun* 9:4881. <https://doi.org/10.1038/s41467-018-07335-2>.
 41. Yu FB, Blainey PC, Schulz F, Woyke T, Horowitz MA, Quake SR. 5 July 2017, posting date. Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife* <https://doi.org/10.7554/eLife.26580>.
 42. McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooshef S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS. 2013. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc Natl Acad Sci U S A* 110: E2390–E2399. <https://doi.org/10.1073/pnas.1219809110>.
 43. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with an expanded complement of translation system components. *Science* 356:82–85. <https://doi.org/10.1126/science.aal4657>.
 44. Berghuis BA, Yu B, Schulz F, Blainey PC, Woyke T, Quake SR. 2019. Hydrogenotrophic methanogenesis in archaeal phylum Verstraetearchaeota reveals the shared ancestry of all methanogens. *Proc Natl Acad Sci U S A* 116:5037–5044. <https://doi.org/10.1073/pnas.1815631116>.
 45. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>.
 46. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooshef S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Etema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattai T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C; Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 35:725–731. <https://doi.org/10.1038/nbt.3893>.
 47. Clingenpeel S, Clum A, Schwientek P, Rinke C, Woyke T. 8 January 2015, posting date. Reconstructing each cell’s genome within complex microbial communities – dream or reality? *Front Microbiol* <https://doi.org/10.3389/fmicb.2014.00771>.
 48. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povilaitis T, Thompson BP, Mascena CJ, Bellows WK, Lubys A. 20 July 2017, posting date. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun* <https://doi.org/10.1038/s41467-017-00128-z>.
 49. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, Maestre FT, Singh BK, Fierer N. 2018. A global atlas of the dominant bacteria found in soil. *Science* 359:320–325. <https://doi.org/10.1126/science.aap9516>.
 50. Horn M. 2008. Chlamydiae as symbionts in eukaryotes. *Annu Rev Microbiol* 62:113–131. <https://doi.org/10.1146/annurev.micro.62.081307.162818>.

51. Schulz F, Tysl T, Pizzetti I, Dyková I, Fazi S, Kostka M, Horn M. 2015. Marine amoebae with cytoplasmic and perinuclear symbionts deeply branching in the Gammaproteobacteria. *Sci Rep* 5:13381. <https://doi.org/10.1038/srep13381>.
52. Pagnier I, Yutin N, Croce O, Makarova KS, Wolf YI, Benamar S, Raoult D, Koonin EV, La Scola B. 2015. *Babela massiliensis*, a representative of a widespread bacterial phylum with unusual adaptations to parasitism in amoebae. *Biol Direct* 10:13. <https://doi.org/10.1186/s13062-015-0043-z>.
53. Deeg CM, Zimmer MM, George E, Husnik F, Keeling PJ, Suttle CA. 2018. *Chromulinavorax destructans*, a pathogenic TM6 bacterium with an unusual replication strategy targeting protist mitochondrion. *bioRxiv* <https://doi.org/10.1101/379388>.
54. Elwell C, Mirrashidi K, Engel J. 2016. Chlamydia cell biology and pathogenesis. *Nat Rev Microbiol* 14:385–400. <https://doi.org/10.1038/nrmicro.2016.30>.
55. Graells T, Ishak H, Larsson M, Guy L. 1 December 2018, posting date. The all-intracellular order Legionellales is unexpectedly diverse, globally distributed and lowly abundant. *FEMS Microbiol Ecol* 94 <https://doi.org/10.1093/femsec/fiy185>.
56. DeAngelis KM, Pold G, Topçuoğlu BD, van Diepen LTA, Varney RM, Blanchard JL, Melillo J, Frey SD. 13 February 2015, posting date. Long-term forest soil warming alters microbial communities in temperate forest soils. *Front Microbiol* <https://doi.org/10.3389/fmicb.2015.00104>.
57. Eichorst SA, Trojan D, Roux S, Herbold C, Rattei T, Woebken D. 2018. Genomic insights into the *Acidobacteria* reveal strategies for their success in terrestrial environments. *Environ Microbiol* 20:1041–1063. <https://doi.org/10.1111/1462-2920.14043>.
58. Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M. 2014. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J* 8:115–125. <https://doi.org/10.1038/ismej.2013.142>.
59. Aherfi S, Colson P, La Scola B, Raoult D. 2016. Giant viruses of amoebas: an update. *Front Microbiol* 7:349. <https://doi.org/10.3389/fmicb.2016.00349>.
60. Duron O, Doublet P, Vavre F, Bouchon D. 12 October 2018, posting date. The importance of revisiting legionellales diversity. <https://doi.org/10.1016/j.pt.2018.09.008>.
61. Konstantinidis KT, Rosselló-Móra R, Amann R. 2017. Uncultivated microbes in need of their own taxonomy. *ISME J* 11:2399–2406. <https://doi.org/10.1038/ismej.2017.113>.
62. Doud DFR, Woyke T. 2017. Novel approaches in function-driven single-cell genomics. *FEMS Microbiol Rev* 41:538–548. <https://doi.org/10.1093/femsre/fux009>.
63. Costa OYA, Raaijmakers JM, Kuramae EE. 2018. Microbial extracellular polymeric substances: ecological function and impact on soil aggregation. *Front Microbiol* 9:1636. <https://doi.org/10.3389/fmicb.2018.01636>.
64. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* 21:487–493. <https://doi.org/10.1101/gr.113985.110>.
65. Luo C, Rodriguez-R LM, Konstantinidis KT. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res* 42:e73. <https://doi.org/10.1093/nar/gku169>.
66. Seshadri R, Leahy SC, Attwood GT, Teh KH, Lambie SC, Cookson AL, Eloe-Fadrosh EA, Pavlopoulos GA, Hadjithomas M, Varghese NJ, Paez-Espino D, Hungate1000 Project Collaborators, Perry R, Henderson G, Creevey CJ, Terrapon N, Lapebie P, Drula E, Lombard V, Rubin E, Kyrpides NC, Henrissat B, Woyke T, Ivanova NN, Kelly WJ. 2018. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat Biotechnol* 36:359–367. <https://doi.org/10.1038/nbt.4110>.
67. Fernández-Gómez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González JM, Pedrós-Alió C. 2013. Ecology of marine Bacteroidetes: a comparative genomics approach. *ISME J* 7:1026–1037. <https://doi.org/10.1038/ismej.2012.169>.
68. Kabisch A, Otto A, König S, Becher D, Albrecht D, Schüler M, Teeling H, Amann RL, Schweder T. 2014. Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes ‘Gramella forsetii’ KT0803. *ISME J* 8:1492–1502. <https://doi.org/10.1038/ismej.2014.4>.
69. Thomas F, Hehemann JH, Rebuffet E, Czjzek M, Michel G. 30 May 2011, posting date. Environmental and gut Bacteroidetes: the food connection. *Front Microbiol* <https://doi.org/10.3389/fmicb.2011.00093>.
70. Penz T, Schmitz-Esser S, Kelly SE, Cass BN, Müller A, Woyke T, Malfatti SA, Hunter MS, Horn M. 2012. Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium* hertigii. *PLoS Genet* 8:e1003012. <https://doi.org/10.1371/journal.pgen.1003012>.
71. Glavina Del Rio T, Abt B, Spring S, Lapidus A, Nolan M, Tice H, Copeland A, Cheng J-F, Chen F, Bruce D, Goodwin L, Pitluck S, Ivanova N, Mavromatis K, Mikhailova N, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang Y-J, Jeffries CD, Chain P, Saunders E, Detter JC, Brettin T, Rohde M, Göker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-P, Lucas S. 2010. Complete genome sequence of *Chitinophaga pinensis* type strain (UQM 2034). *Stand Genomic Sci* 2:87–95. <https://doi.org/10.4056/signs.661199>.
72. Berlemont C, Martiny RC. 2015. Genomic potential for polysaccharide deconstruction in bacteria. *Appl Environ Microbiol* 81:1513–1519. <https://doi.org/10.1128/AEM.03718-14>.
73. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42:D490–D495. <https://doi.org/10.1093/nar/gkt1178>.
74. Berlemont R, Martiny AC. 2013. Phylogenetic distribution of potential cellulases in bacteria. *Appl Environ Microbiol* 79:1545–1554. <https://doi.org/10.1128/AEM.03305-12>.
75. Taillefer M, Arntzen MØ, Henrissat B, Pope PB, Larsbrink J. 2018. Proteomic dissection of the cellulolytic machineries used by soil-dwelling *Bacteroidetes*. *mSystems* 3:e00240-18. <https://doi.org/10.1128/mSystems.00240-18>.
76. Hoell IA, Vaaje-Kolstad G, Eijsink V. 2010. Structure and function of enzymes acting on chitin and chitosan. *Biotechnol Genet Eng Rev* 27:331–366. <https://doi.org/10.1080/02648725.2010.10648156>.
77. Mewis K, Lenfant N, Lombard V, Henrissat B. 2016. Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization. *Appl Environ Microbiol* 82:1686–1692. <https://doi.org/10.1128/AEM.03453-15>.
78. McKee LS, Martínez-Abad A, Ruthes AC, Vilaplana F, Brumer H. 2019. Focused metabolism of β -glucans by the soil Bacteroidetes species *Chitinophaga pinensis*. *Appl Environ Microbiol* 85:e02231-18. <https://doi.org/10.1128/AEM.02231-18>.
79. Shen L, Liu Y, Xu B, Wang N, Zhao H, Liu X, Liu F. 2017. Comparative genomic analysis reveals the environmental impacts on two Arctic bacter strains including sixteen Sphingobacteriaceae species. *Sci Rep* 7. <https://doi.org/10.1038/s41598-017-02191-4>.
80. Luo Y, Wan S, Hui D, Wallace LL. 2001. Acclimatization of soil respiration to warming in a tall grass prairie. *Nature* 413:622–625. <https://doi.org/10.1038/35098065>.
81. Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28:27–30. <https://doi.org/10.1093/nar/28.1.27>.
82. Kanehisa M, Sato Y, Morishima K, Sternberg M. 14 November 2015, posting date. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* <https://doi.org/10.1016/j.jmb.2015.11.006>.
83. Graham ED, Heidelberg JF, Tully BJ. 2018. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J* 12:1861–1866. <https://doi.org/10.1038/s41396-018-0091-3>.
84. Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M. 2010. The genome of the amoeba symbiont ‘*Candidatus Amoebophilus asiaticus*’ reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol* 192:1045–1057. <https://doi.org/10.1128/JB.01379-09>.
85. Zchori-Fein E, Perlman SJ, Kelly SE, Katzir N, Hunter MS. 2004. Characterization of a ‘*Bacteroidetes*’ symbiont in *Encarsia* wasps (Hymenoptera: Aphelinidae): proposal of ‘*Candidatus Cardinium hertigii*’. *Int J Syst Evol Microbiol* 54:961–968. <https://doi.org/10.1099/ijs.0.02957-0>.
86. Chang H-H, Cho S-T, Canale MC, Mugford ST, Lopes JRS, Hogenhout SA, Kuo C-H. 29 January 2015, posting date. Complete genome sequence of ‘*Candidatus Sulcia muelleri*’ ML, an obligate nutritional symbiont of maize leafhopper (*Dalbulus maidis*). *Genome Announc* <https://doi.org/10.1128/genomeA.01483-14>.
87. Ló Pez-Sánchez MJ, Neef A, Peretó J, Patiñ O-Navarrete R, Pignatelli M. 2009. Evolutionary convergence and nitrogen metabolism in *Blattabacteria* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. *PLoS Genet* 5:1000721. <https://doi.org/10.1371/journal.pgen.1000721>.
88. Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323:379–382. <https://doi.org/10.1126/science.1167140>.

89. McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10:13–26. <https://doi.org/10.1038/nrmicro2670>.
90. Melillo JM, Steudler PA, Aber JD, Newkirk K, Lux H, Bowles FP, Catricala C, Magill A, Ahrens T, Morrisseau S. 2002. Soil warming and carbon-cycle feedbacks to the climate system. *Science* 298:2173–2176. <https://doi.org/10.1126/science.1074153>.
91. Bushnell B. 2002. BBTtools. <https://sourceforge.net/projects/bbmap/>.
92. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>.
93. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. <https://doi.org/10.7717/peerj.1165>.
94. Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>.
95. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066. <https://doi.org/10.1093/nar/gkf436>.
96. Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10:210. <https://doi.org/10.1186/1471-2148-10-210>.
97. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9:5114. <https://doi.org/10.1038/s41467-018-07641-9>.
98. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. <https://doi.org/10.1371/journal.pone.0009490>.
99. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44:W242–W245. <https://doi.org/10.1093/nar/gkw290>.
100. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng J-F, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk H-P, Eisen JA. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056–1060. <https://doi.org/10.1038/nature08656>.
101. Buchfink B, Xie C, Huson DH. 17 November 2014, posting date. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* <https://doi.org/10.1038/nmeth.3176>.
102. Huson DH, Albrecht B, Bağcı C, Bessarab I, Górska A, Jolic D, Williams R. 2018. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol Direct* 13:6. <https://doi.org/10.1186/s13062-018-0208-7>.
103. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
104. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>.
105. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 46:95–101. <https://doi.org/10.1093/nar/gky418>.