PLOS ONE

RESEARCH ARTICLE

Identification of the Core Set of Carbon-Associated Genes in a Bioenergy Grassland Soil

Adina Howe[1], Fan Yang[1], Ryan J. Williams[1], Folker Meyer[2], Kirsten S. Hofmockel[3,4]*

1 Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA, 50011, United States of America, **2** Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, 60439, United States of America, **3** Department of Ecology and Evolutionary Biology, Iowa State University, Ames, IA, 50011, United States of America, **4** Pacific Northwest National Laboratory, Richland, WA, 99352, United States of America

* khof@iastate.edu

OPEN ACCESS

Citation: Howe A, Yang F, Williams RJ, Meyer F, Hofmockel KS (2016) Identification of the Core Set of Carbon-Associated Genes in a Bioenergy Grassland Soil. PLoS ONE 11(11): e0166578. doi:10.1371/journal.pone.0166578

Editor: John J. Kelly, Loyola University Chicago, UNITED STATES

Received: July 25, 2016

Accepted: October 31, 2016

Published: November 17, 2016

Data Availability Statement: All metagenome sequencing files are available from the MG-RAST database (accession numbers are provided in S2 Table).

Funding: This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number SC0010775. Fieldwork was supported by the USDA National Institute of Food and Agriculture Carbon Cycle Science Program grant number 2011-01033.

Abstract

Despite the central role of soil microbial communities in global carbon (C) cycling, little is known about soil microbial community structure and even less about their metabolic pathways. Efforts to characterize soil communities often focus on identifying *differences* in gene content across environmental gradients, but an alternative question is what genes are *similar* in soils. These genes may indicate critical species or potential functions that are required in all soils. Here we identified the "core" set of C cycling sequences widely present in multiple soil metagenomes from a fertilized prairie (FP). Of 226,887 sequences associated with known enzymes involved in the synthesis, metabolism, and transport of carbohydrates, 843 were identified to be consistently prevalent across four replicate soil metagenomes. This core metagenome was functionally and taxonomically diverse, representing five enzyme classes and 99 enzyme families within the CAZy database. Though it only comprised 0.4% of all CAZy-associated genes identified in FP metagenomes, the core was found to be comprised of functions similar to those within cumulative soils. The FP CAZy-associated core sequences were present in multiple publicly available soil metagenomes and most similar to soils sharing geographic proximity. In soil ecosystems, where high diversity remains a key challenge for metagenomic investigations, these core genes represent a subset of critical functions necessary for carbohydrate metabolism, which can be targeted to evaluate important C fluxes in these and other similar soils.

Introduction

Soil microbial communities are of critical importance; they influence nutrient availability, decomposition rates, greenhouse gas emissions, soil fertility, and agricultural production [1–3]. Despite decades of research, we still know very little about soil microbial community structure and functioning, especially in agricultural soils. Sequencing-based approaches, particularly metagenomics, have greatly enhanced the resolution at which we can investigate genes

Competing Interests: The authors have declared that no competing interests exist.

contained within soil microbial communities [4–6]. Nonetheless we are becoming increasingly aware that the incredible diversity present in soils requires very deep sampling to capture (estimated Terabasepairs of sequencing) [7,8]. Efforts to characterize soil communities often focus on identifying *differences* in gene content across environmental gradients (e.g., genes present under varying land use history, nutrient loads, soil moisture content) [6,9–11]. An alternative question is what genes are *similar* in soils? In other words, is there a core set of sequences, genes, or functions that are present in *all* soils? Are the core genes found in one soil type, field, or even plot representative in soils? A similar effort in the gut microbiome environment identified gut-associated genes shared between multiple humans (124 European individuals) and provided a transformative reference gene set describing the minimal gut metagenome among these individuals and its encoded functions [12]. The microbial diversity in soils is magnitudes higher than the gut microbiome [13], suggesting that a soil core, if present, would be much smaller. We explore the presence of core gene sequences in a single experimental field and evaluate the insight it provides for soil function.

A single field site was selected for characterization of a soil core. Within a field, high levels of local spatial variation of microbial community structure have been observed [14–16]. Consequently, our study focused on the identification of a plot scale core soil microbial community in fertilized prairie (FP) whole soil (WS) metagenomes from a single experiment. The resulting soil core was also compared to several other soil metagenomes, including soil aggregate metagenomes from the same plot (e.g., originating from sieved partitions of the same FP WS), metagenomes from soils located nearby, and publicly available soil metagenomes. We expect that genes that are ubiquitously present in multiple soils may represent functions that are critical to soil processes. To evaluate the functions represented in our soil core, we identified genes associated with carbon (C) cycling and evaluated their contributions to microbial biomass synthesis and decomposition in these soils.

Materials and Methods

Samples were collected with permission from the Committee for Agricultural Development, a nonprofit organization in Iowa that owns the property.

Study site

Soil was collected from the Iowa State Comparison of Biofuel Systems (COBS) experimental site located on the South Reynoldson Farm in Boone County, IA (41˚55'14.42"N, 93˚44'58.96"W); see [17] for a detailed site description. Soils consisted of loams in the Nicollet (Fine-loamy, mixed, superactive, mesic Aquic Hapludoll) and Webster (Fine-loamy, mixed, superactive, mesic Typic Endoaquoll) series with less than 3% slope. Sand content ranged from 27% to 53% across the site and clay content ranged from 17% to 32%. In the 5 years prior to sampling, average growing season precipitation at the site was 91.8 cm and mean annual temperature was 9˚C. Four replicate blocks contain four plots (27 x 61 m^2) of each planting treatment in a randomized complete block design. The present study includes samples from plots planted with fertilized native tallgrass prairie (31 species). Soil cores (5.5 cm x 10 cm) were collected in July 2012 as described in [17]. Subsamples of soil were separated into soil aggregate fractions by an optimal sieving method prior to DNA sequencing. Biogeochemical analyses of these samples has been previously reported [18,19].

DNA extraction and library preparation

For each soil sample, DNA was extracted from 0.25 g of soil by using MoBio PowerLyzer PowerSoil DNA Isolation Kit (MoBio, Carlsbad, CA). DNA was quantified using Nanodrop

and approximately 1 μg of DNA per sample was used for metagenomic sequencing. Metagenome libraries were prepared with IntegenX PrepX DNA Library Kit with 180 bp overlapping inserts and subsequently size-selected prior to sequencing on an Illumina HiSeq2000. Library preparation and sequencing were performed at Argonne National Laboratory (Argonne, IL).

Assembly and coverage of soil metagenome

An assembly of all soil metagenomes available from this site was used to generate a reference set of contigs for this study as previously described in [7]. All sequencing reads originating from FP samples (both WS and varying sizes of soil aggregates, n = 20) were combined and assembled (sequencing reads available for all data in MG-RAST, see S1 and S2 Tables) to create a cumulative reference metagenome. This reference was used to identify shared core contig sequences among all FP WS metagenomes. Prior to assembly, Illumina adapters were trimmed with Trimmomatic (v0.27, [20]) using Illumina TruSeq2-PE with threshold of seed mismatches, palindrome clip threshold, and simple clip threshold as 2, 30, and 10, respectively. Remaining paired end sequences were merged with PANDAseq [21]. The resulting sequences were normalized using the *khmer* package [22–24] and methods previously described in [7] with the following parameters: -k 20 -C 10 -N 4 -x 100e9. High abundance k-mers with coverage $>$ 50 were trimmed from sequences, and the remaining sequences were partitioned as previously described in [7,25] with the following parameters: -k 32 -N 4 -x 80e9. Assembly was performed with the Velvet assembler (v 1.2.10, [26]) using odd k-mer lengths from 33 to 65. Resulting assembled contiguous sequences (contigs) were merged as described previously [7] using CD-HIT (v4.6, [27,28] and minimus2 (Amos v3.1.0, [29])

Abundances of contigs associated with each sample was estimated through the alignment of sequencing reads with assembled contigs using Bowtie2 (v2.0.5, default parameters, [30]). Coverage of each contig was estimated as the maximum base pair coverage of the assembled sequence, requiring a minimum of coverage length of 100 bp. This abundance estimation was intentionally chosen to be liberal, intending to capture representative core contigs even at low abundance in individual samples. To be identified as present in *all* samples, contigs were required to be present at 5 or greater base pair coverage in all four local soil metagenomes. For core contigs, the MG-RAST automated annotation system was used to identify associated function of assembled contigs (MG-RAST ID 4519723.3) [31]. In order to standardize samples with various sampling depths, the total number of single-copy *recA* genes in each sample was estimated, using annotations from MG-RAST, requiring a sequence alignment (E-value $<$ 1e-5) to *recA* (Subsystem ID SS04542). The coverage of each gene was estimated from the base pair estimated coverage of its originating assembled contig divided by the total estimated *recA* genes identified in each sample. To identify carbon-associated genes, all assembled contigs were compared to known proteins in the Carbohydrate Active Enzyme (CAZy) database [32] using NCBI BLAST (v2.2.25, blastx), requiring an E-value $\leq$ 1e-5. The best scoring alignment to the CAZy database for each sequence was used for characterization (both function and taxonomic origin). If multiple annotations shared identical best score alignments, one was chosen at random. The resulting set of CAZy-associated assembled contigs comprised the dataset referred to as the cumulative FP-CAZy metagenome. Annotations associated with each contig as well as analysis performed in this study can be found at https://github.com/germs-lab/carbon-core-soil-paper.

Characterization of the fertilized prairie whole soil carbon core community

To identify the genes present in all of the available metagenomes of the FP WS, we defined the *core* carbohydrate-associated contigs as sequences that were present a minimum abundance of

5 or greater base pair coverage in each of the four WS field replicate metagenomes; these core contigs are hereafter referred to as the FP-CAZy core. As we focused on carbohydrate-associated genes, only contigs that shared sequence similarity to a known protein within the CAZy database were considered.

To examine the content of the FP-CAZy core, a phylogenetic tree (based on bacterial and archaeal 16S rRNA genes) was constructed for bacteria associated with FP-CAZy core contigs (S1 Fig). As core proteins cannot be aligned easily to construct a phylogenetic tree, representative 16S rRNA genes associated with phyla associated with CAZy genes were obtained. Core sequences were searched against GenBank (release 198.0) from which organisms with full genomes were identified. Bacterial and archaeal 16S rRNA genes were extracted from these full genomes to build the phylogenetic tree. For microorganisms containing multiple 16S rRNA genes, the first gene sequence identified in the GenBank record was selected as representative. The 16S rRNA gene sequences were aligned by using RDP (Ribosomal Database Project, release 11, [33]) Aligner. MEGA 5.2 (Molecular Evolutionary Genetics Analysis, [34]) was used to construct the phylogenetic tree. Specifically, the phylogeny was inferred using maximum likelihood heuristic method (nearest neighbor interchange) with a general time reversible nucleotide substitution model (discrete gamma distribution with 5 categories and allowing the presence of invariant sites). The phylogeny was tested using the boostrap method (999 times). For every microorganism in the phylogenetic tree, the abundance of associated FP-CAZy core contigs associated to that phyla (standardized against the recA gene abundance) was also calculated. For each metagenome CAZy annotations were grouped into six CAZy classes, glycosyltransferases (GT), glycoside hydrolases (GH), carbohydrate esterases (CE), carbohydrate-binding modules (CB), polysaccharide lyases (PL) and unknown. The abundance of each CAZy class was summed based on phylogenetic identification within the same sample, and averaged across 4 replicates (at 95% confidence intervals).

The resulting FP-CAZy core contigs were compared to other soil metagenomes. Within the COBS experimental site, these other metagenomes included soil aggregate fractions isolated from FP WS (n = 14), as well as microaggregates isolated from unfertilized prairie (n = 4) and continuous corn (n = 2) cropping systems (S2 Table). FP-CAZy core sequences were considered present in these samples if identified in at least one field replicate with 5 or greater base pair coverage. To assess the relevance core sequences more broadly, we analyzed several public soil metagenomes (Fierer et al 2012, Howe et al 2014), where presence of FP-CAZy core sequences were considered to be present if sequence similarity of the best scoring BLAST alignments of core sequence to metagenome reads had a minimum E-value score of 1e-5, length of 70, and identity of 70%.

Results

Characterization of the FP-CAZy core

The total size of each FP WS replicate metagenome ranged from 4.3 to 20.5 Gbp (average 10.3 ± 7.1 Gbp). Within these metagenomes, a total of 226,887 contigs were identified with shared sequence similarity to known CAZy proteins. Among these contigs, a total of 11,193 contigs were identified as present in all four replicates (at varying abundances). Requiring a minimal abundance of greater than 5-fold bp coverage, a total of 911 sequences were present in all four metagenomes and comprised the FP-CAZy core metagenome (representing 499,329 bp of 41.4 Gbp). Given the conservative requirements for core sequences, these represent a lower bound estimate of shared sequences within the field replicated FP metagenomes. To evaluate the effects of sequencing depth, we estimated the total number of core k-mers in subsets of the four WS metagenomes, 100,000 to 100 million reads (278 to 539 million unique k-

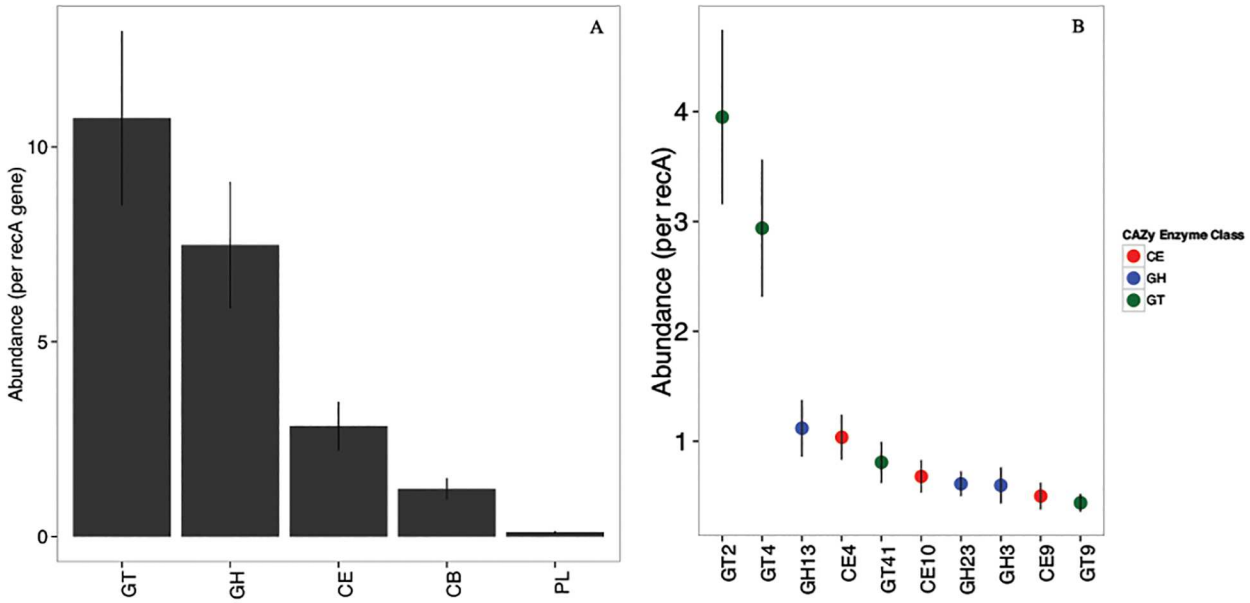

Fig 1. Functional profile (mean ± SE) of CAZy enzyme classes (A) and 10 most abundant enzyme families (B) represented in FP-CAZy core sequences (GT, glycosyltransferase; GH, glycoside hydrolase; CE, carbohydrate esterase; CB, carbohydrate-binding module; PL, polysaccharide lyase).

doi:10.1371/journal.pone.0166578.g001

mers average, n = 4 bootstraps). We found that the total core k-mers comprised 4.8 to 8.9% of total unique k-mers, suggesting that with more sequencing, the core size continue to grow linearly. The sequencing depth represented within this study (minimum 4.3 Gbp) is greater than the average sequencing depth of 1.67 Gbp (median 0.47 Gbp) of 1,093 public soil-associated metagenomes in MG-RAST (October 27, 2015, material = peat soil, sediment, soil, agricultural soil, alpine soil, arable soil, bulk soil, clay soil, farm soil, grassland soil, lawn soil, leafy wood soil, paddy field soil, rhizosphere, sandy sediment, volcanic soil, and xylene contaminated soil). Even with this above average sequencing effort, we were able to identify only 911 core sequences among the four FP metagenomes. When we further limited our analysis to CAZy-associated proteins related to bacteria, archaea, viruses, and fungi, our core consisted of a total of 843 contigs shared between all four FP replicates.

We next explored the content of our core, evaluating the putative functions and taxonomy of proteins sharing similarity to core contigs. Functions and taxonomy associated with core sequences were quantified in their prevalence (number of unique occurrences within the core) and relative abundance (cumulative abundance). Overall, the FP-CAZy core of 843 contigs represented five major enzyme classes and 99 enzyme families. The average relative abundance of proteins indicated GT > GH > CE > CB > PL (Fig 1A, S1 Fig). Among these, the most abundant enzyme families included GT2, GT4, and GH13, which were associated with a total prevalence of 145, 109, and 38 core sequences and present at estimated average abundances of 4.0, 2.9, and 1.1 copies per *recA* sequence, respectively (Fig 1B). The most represented taxa of the FP-CAZy core were similar to Proteobacteria, Actinobacteria, and Firmicutes, related to a total prevalence count of 266, 120, and 70 unique assembled sequences, respectively. The relative abundances of phyla associated with core sequences revealed diverse membership (Fig 2), with sequences associated with Proteobacteria broadly represented in association with enzymes GT, GH, and CE (S1 Fig). Fungal sequences associated with GH and CE enzymes were also abundant. A few core enzyme classes were represented by only a single phylum, the

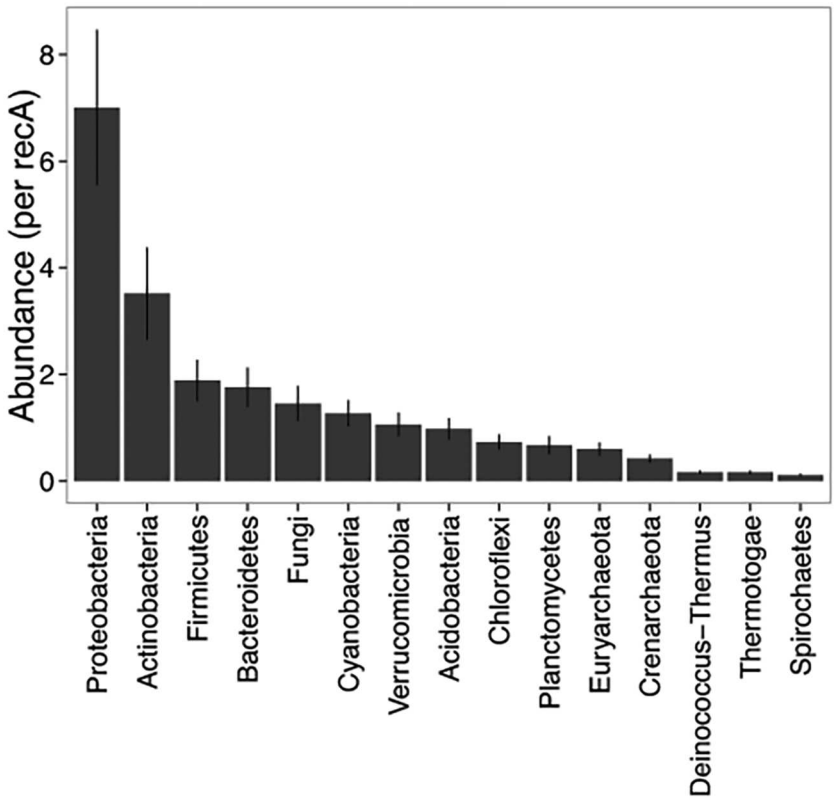

Fig 2. Abundance of phyla represented in the FP CAZy core (mean ± SE; n = 4).

doi:10.1371/journal.pone.0166578.g002

most abundant classes including GH94 (37 copies/1000 *recA*, Proteobacteria) and GH76 (29 copies/100 *recA*, Fungi).

Because the identification of core sequences required similarity to protein encoding sequences within the CAZy database, we evaluated potential biasing of the core from the CAZy database itself. We randomly selected 50,000 proteins each from the CAZy database and the cumulative metagenome to generate a random set of CAZy classes and associated taxonomy. The simulated random distribution was performed 1000 times to compare to the observed CAZy classes and phylogeny of the core sequences. We found that protein distributions were significantly different for both CAZy class and phyla (ANOVA, $p < 0.05$), suggesting that observations here not the result of database bias.

Representation of FP-CAZy core in cumulative FP and other soil metagenomes

The total number of FP-CAZy core sequences represent only 0.4% of all CAZy proteins identified in the cumulative FP metagenomes (843 out of 226,887 unique sequences). With greater sequencing depth, we would expect this number to increase significantly. By abundance, the core comprises ~1% of the most abundant ($\geq$ 5-fold bp coverage) CAZy-associated sequences. We compared core functions and taxonomy to those all observed functions and taxa in the FP metagenomes. The distributions of the abundances of identified CAZy enzyme classes in the core and cumulative FP metagenome were similar though the relative abundances of CB and GT differed by up to 4% (S3 Fig). We observed that at the enzyme family level, abundances in the core and cumulative datasets contrasted, with the exception of GT2 and CE10 (p-

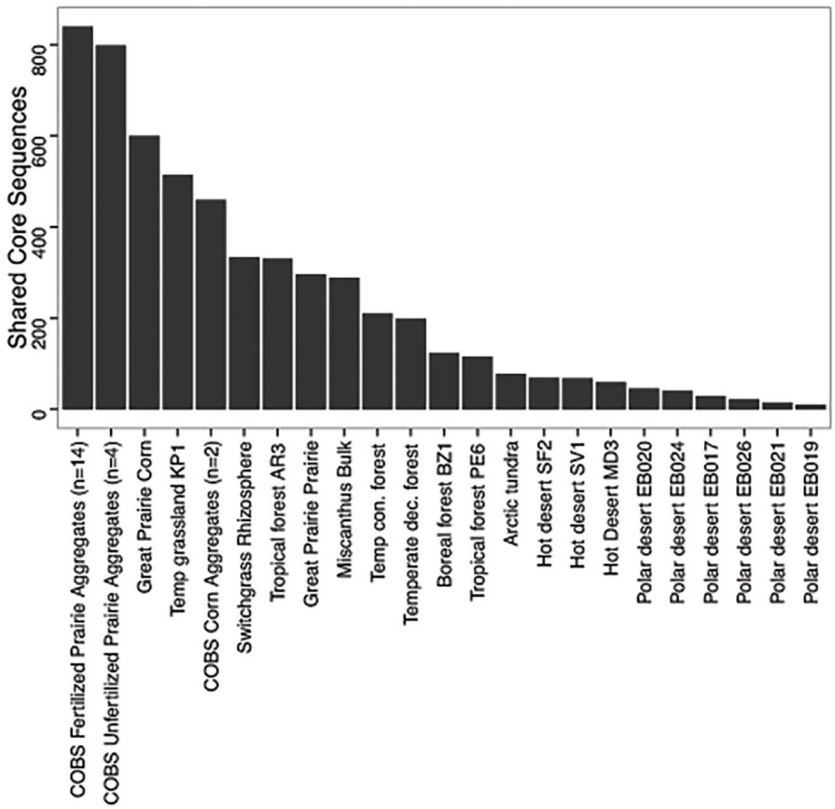

Fig 3. Number of shared fertilized prairie metagenome core sequences in global soil metagenomes sharing sequence similarity (E-value 1e-5).

doi:10.1371/journal.pone.0166578.g003

value > 0.2). Similarly, the distribution of phyla-associated with dominant enzyme families (e.g., GT2, GH13, CE10, and GT4) was significantly different between the core and cumulative metagenomes (S4 Fig). For example, proteins associated with GT2 originating from Actinobacteria (14 sequences) and Firmicutes (6 sequences) were significantly enriched in the core.

Comparison of the FP-CAZy core to metagenomes of FP aggregates as well as adjacent corn and unfertilized prairie aggregates (Bach and Hofmockel 2014) revealed substantial sequence similarity. The large majority of FP-CAZy core sequences, 840 out of 843, were also identified within FP soil aggregate fractions (sieved fractions of whole soil samples). In adjacent soil samples (microaggregates of unfertilized prairie (n = 4) and corn (n = 2) fields, a total of 792 and 600 FP-CAZy core sequences were observed in unfertilized prairie and corn metagenomes, respectively. Comparing the FP-CAZy core to other globally distributed soil metagenomes (Fig 3, S2 Table) revealed the most shared sequences (e.g., sequence similarity) with other grassland (405 ± 109) and agricultural soils (535 ± 95) and fewer shared core sequences with forest (196 ± 39), tundra (78), and desert (40 ± 8) soils. However, at the functional level (as opposed to sequence similarity), the relative distribution of CAZy enzyme classes was broadly similar (Fig 4).

Discussion

The soil represents arguably the most challenging environment to access with modern molecular microbial ecology. Its high diversity and spatial heterogeneity, even at the meter scale,

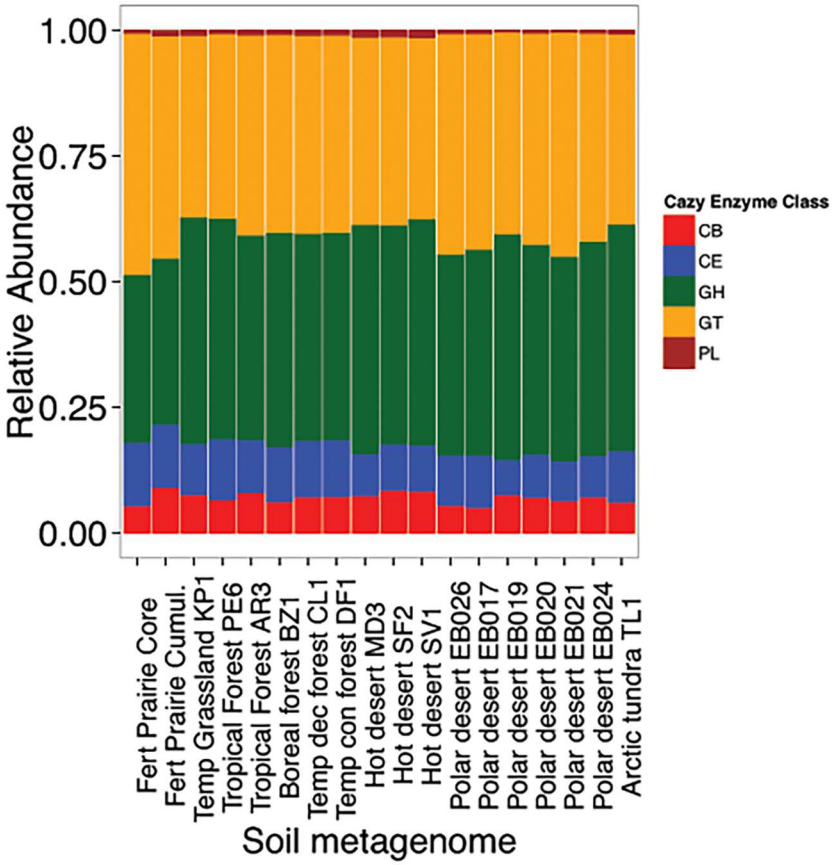

Fig 4. Functional distribution of the presence of CAZy enzyme classes in global soil metagenomes.

doi:10.1371/journal.pone.0166578.g004

make it difficult to sample and characterize. Despite these difficulties, the importance of microbial communities in terrestrial biogeochemistry and ecosystem C- cycling are well agreed upon [35,36]. Understanding the functional capacity of soil microorganisms remains an important goal for understanding ecosystem health and stability [37]. We explored the insights a minimal local soil core metagenome could provide for identifying key enzymes, microorganisms, and ecological functions related to soil C cycling. Among all C-cycling enzymes identified in our soil metagenomes, the CAZy-associated soil core represented only 0.4% of genes. Despite representing only a fraction of all genes observed in FP metagenomes, the identified FP-CAZy core shared similar functions to whole metagenomes, supporting our hypothesis that core sequences represent a set of minimum C-cycling functions necessary in FP soils. The most dominant functions identified within the core and cumulative metagenomes were GT2 and GT4, which are involved in the formation of cell wall polysaccharides of diverse organisms including bacteria, archaea, fungi, and plants, as well as numerous biological processes such as pathogen protection, intercellular signaling, and biofilm production [38]. In the core, these enzymes have been observed as originating from Proteobacteria. Sequences similar to Bacteroidetes and Actinobacteria dominated the GT2 family in core functions, while Verrucomicrobia and Euryarchaeota genes were prevalent in GT4 family. We observe that broad membership may provide core enzymes in soils, supporting previous observations that high biodiversity may help to stabilize carbon cycling [39,40].

We found that genes associated with amylolytic enzymes (GH13) were highly prevalent in the core, highlighting the central importance of breakdown and utilization of starch and related oligo- and polysaccharides. In general, genes known to be associated with GH13 contribute to trehalose synthesis, a compound that is used by both plants and fungi to store carbon and energy, as well as protecting bacterial cells from physical and chemical stresses [41]. Abundant core sequences associated with GH13 are similar to genes of Proteobacteria, Actinobacteria, and Planctomycetes, suggesting that these organisms may play a central role in starch utilization. Although Proteobacteria and Actinobacteria are commonly associated with C cycling in soil, Planctomycetes have only recently been identified in agricultural soils [42], tundra soils [43], and Arctic peats [44], demonstrating its global presence, and potential functional importance in soils. Additionally, Planctomycetes have been associated with decomposition of cellulose within agricultural soils [45], making this a noteworthy phylum for further investigations focused on soil C cycling.

Genes that were enriched in the core relative to the cumulative metagenome were hypothesized to play critical roles in soil C-cycling. These genes included sequences sharing similarity to GH13 associated Actinobacteria and Planctomycetes, GT4 associated Actinobacteria, Firmicutes, Euryarchaeota, and Verrucomicrobia, GT2 associated Actinobacteria and Firmicutes, and CE10 associated Fungi and Proteobacteria sequences. These enriched core functions represented a relatively small diversity of all observed carbon related functions. Enriched core genes comprised 3% of cumulative GT2-associated Actinobacteria and Firmicutes proteins; 3% of GT4-associated Actinobacteria, Firmicutes, Euryarchaeota, and Verrucomicrobia proteins; 14% of CE10-associated Proteobacteria and Fungi proteins; and 5% of GH13-associated Actinobacteria and Planctomycete proteins. Their prevalence in multiple soils suggest that amongst diverse functions and memberships, these genes and bacteria are critical for C-cycling.

Another interesting observation within the core was the presence of ubiquitous and abundant sequences associated with CE, in particular CE10. In general, CE genes act on plant polysaccharides to degrade acetylated plant hemicelluloses [46] and are commonly clustered with GHs in operons or regulons and are co-expressed to decompose esterified polysaccharides of plant cell walls. Little is known about CE10, which is associated with colinesterase type enzyme. Soil substrates associated with CE, α- and β-glycerophosphates and choline-P, have been identified as degradation products of phospholipids of cellular membranes during NMR analysis [47–49] and have been shown to cycle and accumulate in agricultural soils [50]. The observed CE10 presence in the core supports the premise that cell membranes may be an important soil substrate, and CE hydrolases may be a potential target for quantifying the importance of microbial cell wall turnover, which is a pressing question given the putative importance of microbial necromass to soil C storage [51,52].

At a broad level, the functions encoded by the core are observed in global soils, ranging from those originating from agriculture to deserts. However, at the strain level (e.g., sequence variation), our core was significantly more represented in soils from similar land-use and geography. Within the same large-scale field experiment, we found that independent of crop selection or management practices (corn and unfertilized prairie), multiple soil metagenomes shared a large majority of core sequences (97%). In soils originating from Iowa but located about 60 miles SE from the FP site, core sequences were less prevalent compared to local samples (33% in prairie and 68% in corn). In contrasting soils, such as deserts and forests, significantly fewer FP-CAZy core sequences could be identified, suggesting that the soil environment, which can be geographically specific, is important for defining a soil core and its functional potential. This result has been reported previously where soil type was the critical driver of microbial community compositional differences [53,54]. For metagenomic

studies, this observation has implications for the genomic or functional level that should be compared between studies. We observe that genes encoding for functions (e.g., enzyme classes) are largely similar in global soils but find that core sequences are not broadly representative, at least at the current sequencing depths being used to study oil microbiomes.

Sequencing-based approaches for studying the soil microbial communities continue to increase in volume and in complexity (e.g., metaproteomics and metabolomics). Our results present a challenge that is confronting the study of soil ecosystems. At a local scale (soils originating from the same experimental plot), greater than average sequencing depth for current soil studies, and focusing on genes encoding for carbon cycling functions, we are able to identify less than a 1% signal of sequences being shared among multiple metagenomes. In contrast, the gut microbiome identified that 40% of genes was shared within at least half of the 124 individuals studied [12]. These results reaffirm that we are still only beginning to sample the immense diversity in these soils and are far from identifying a minimal soil core microbiome, at least at the gene level. Our results also emphasize the need to consider varying scales of characterization when comparing soil microbial communities. Given the unique nature of soil, we have a strong need to evaluate new methods for binning soil sequences include protein clustering (e.g., operational protein units analogous to operational taxonomic units from sequenced 16S rRNA amplicons), co-occurrence networks and interactions, and improved hierarchy in functional annotations. Continued efforts to us our identified targets to expand what is known about the minimal genes encoding functions in soil will be helpful to identify critical community processes within complex metagenomes and can serve as the framework with which to deconstruct and understand the high diversity of microbial soil communities.

Supporting Information

S1 Fig. Phylogenetic tree of taxonomic origins of CAZy proteins most similar to core genes (using available 16S rRNA genes). The diamonds on the maximum likelihood phylogenetic tree indicate branches with bootstrap values greater than 80. CAZY gene abundance was calculated across four whole-soil replicates and the error bars represent 95% confidence intervals.
(EPS)

S2 Fig. Abundance of phyla represented in enzyme families GH, GT, CB, and CE associated with the fertilized prairie core metagenome.
(EPS)

S3 Fig. Relative abundance of CAZy enzyme families and classes in fertilized prairie metagenome core and the cumulative fertilized prairie metagenome.
(EPS)

S4 Fig. Abundance of phyla associated with enzymes in fertilized prairie metagenome core and the cumulative fertilized prairie metagenome.
(EPS)

S1 Table. Sequencing summary of fertilized prairie whole soil metagenomes.
(DOCX)

S2 Table. Number of shared core sequences among various soil metagenomes. Number of replicates is one unless otherwise indicated.
(DOCX)

Acknowledgments

We thank Elizabeth Bach for sample collection, Sarah Hargreaves for DNA extraction, and Stephanie Moorman and Sarah Owens for preparing sequencing libraries. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number SC0010775. Fieldwork was supported by the USDA National Institute of Food and Agriculture Carbon Cycle Science Program grant number 2011–01033. We acknowledge the support and infrastructure of Magellan, the Argonne Cloud Computing Platform, and their team, especially Ryan Aydelott and Scott Devoid. We are also grateful to the late Dave Sundberg, whose attentive site management made this research possible.

Author Contributions

Conceptualization: AH FY RW KH.

Data curation: AH FY RW.

Formal analysis: AH FY RW.

Investigation: AH FY RW.

Methodology: AH FY RW.

Project administration: AH KH.

Resources: AH FM KH.

Software: AH FY RW.

Supervision: AH KH.

Validation: AH FY RW.

Visualization: AH FY RW.

Writing – original draft: AH FY RW KH.

Writing – review & editing: AH FY RW FM KH.

References

1. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008; 320: 1034–1039. doi: 10.1126/science.1153213 PMID: 18497287
2. Morales SE, Holben WE. Linking bacterial identities and ecosystem processes: Can "omic" analyses be more than the sum of their parts? FEMS Microbiol Ecol. 2011; 75: 2–16. doi: 10.1111/j.1574-6941.2010.00938.x PMID: 20662931
3. Schmidt MWI, Torn MS, Abiven S, Dittmar T, Guggenberger G, Janssens I a., et al. Persistence of soil organic matter as an ecosystem property. Nature. 2011; 478: 49–56. doi: 10.1038/nature10386 PMID: 21979045
4. Leff JW, Jones SE, Prober SM, Barberán A, Borer ET, Firn JL, et al. Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. Proc Natl Acad Sci. National Academy of Sciences; 2015; 112: 10967–10972. doi: 10.1073/pnas.1508382112 PMID: 26283343
5. Xue K, Yuan M M., Shi Z J., Qin Y, Deng Y, Cheng L, et al. Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming. Nat Clim Chang. Nature Research; 2016; 6: 595–600. doi: 10.1038/nclimate2940
6. Cardenas E, Kranabetter JM, Hope G, Maas KR, Hallam S, Mohn WW. Forest harvesting reduces the soil metagenomic potential for biomass decomposition. ISME J. Nature Publishing Group; 2015; 9: 2465–2476. doi: 10.1038/ismej.2015.57 PMID: 25909978

7. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci U S A. 2014; 111: 4904–9. doi: 10.1073/pnas.1402564111 PMID: 24632729
8. Luo C, Rodriguez-R LM, Johnston ER, Wu L, Cheng L, Xue K, et al. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. Appl Environ Microbiol. 2014; 80: 1777–1786. doi: 10.1128/AEM.03712-13 PMID: 24375144
9. Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, et al. Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. Science. 2013; 342: 621–4. doi: 10.1126/science.1243768 PMID: 24179225
10. Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N. Temporal variability in soil microbial communities across land-use types. ISME J. Nature Publishing Group; 2013; 7: 1641–1650. doi: 10.1038/ismej.2013.50 PMID: 23552625
11. Baldrian P, Kolařík M, Štursová M, Kopecký J, Valášková V, Větrovský T, et al. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J. 2012; 6: 248–258. doi: 10.1038/ismej.2011.95 PMID: 21776033
12. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464: 59–65. doi: 10.1038/nature08821 PMID: 20203603
13. Rodriguez-R LM, Konstantinidis KT. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics. Oxford University Press; 2014; 30: 629–35. doi: 10.1093/bioinformatics/btt584 PMID: 24123672
14. Correa-Galeote D, Marco DE, Tortosa G, Bru D, Philippot L, Bedmar EJ. Spatial distribution of N-cycling microbial communities showed complex patterns in constructed wetland sediments. FEMS Microbiol Ecol. 2013; 83: 340–351. doi: 10.1111/j.1574-6941.2012.01479.x PMID: 22928965
15. Franklin RB, Mills AL. Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia agricultural field. FEMS Microbiol Ecol. 2003; 44: 335–346. doi: 10.1016/S0168-6496(03)00074-6 PMID: 12830827
16. Keil D, Meyer A, Berner D, Poll C, Schützenmeister A, Piepho H-P, et al. Influence of land-use intensity on the spatial distribution of N-cycling microorganisms in grassland soils. FEMS Microbiol Ecol. 2011; 77: 95–106. doi: 10.1111/j.1574-6941.2011.01091.x PMID: 21410493
17. Bach EM, Hofmockel KS. Soil aggregate isolation method affects measures of intra-aggregate extracellular enzyme activity. Soil Biol Biochem. Elsevier Ltd; 2014; 69: 54–62. doi: 10.1016/j.soilbio.2013.10.033
18. Bach EM, Hofmockel KS. A time for every season: soil aggregate turnover stimulates decomposition and reduces carbon loss in grasslands managed for bioenergy. GCB Bioenergy. 2016; 8: 588–599. doi: 10.1111/gcbb.12267
19. Bach EM, Hofmockel KS. Coupled Carbon and Nitrogen Inputs Increase Microbial Biomass and Activity in Prairie Bioenergy Systems. Ecosystems. Springer US; 2015; 18: 417–427. doi: 10.1007/s10021-014-9835-8
20. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. Oxford University Press; 2014; 30: 2114–2120.
21. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014; 42: D633–42. doi: 10.1093/nar/gkt1244 PMID: 24288368
22. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research. 2015; 4. doi: 10.12688/f1000research.6924.1 PMID: 26535114
23. Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT, Marçais G, et al. These Are Not the K-mers You Are Looking For: Efficient Online K-mer Counting Using a Probabilistic Data Structure. Zhu D, editor. PLoS One. Public Library of Science; 2014; 9: e101271. doi: 10.1371/journal.pone.0101271 PMID: 25062443
24. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv. 2012;1203.4802: 1–18. Available: http://arxiv.org/abs/1203.4802
25. Pell J, Hintze a., Canino-Koning R, Howe a., Tiedje JM, Brown CT. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci. 2012; 109: 13272–13277. doi: 10.1073/pnas.1121464109 PMID: 22847406
26. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. Cold Spring Harbor Lab; 2008; 18: 821–829.

27. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28: 3150–3152. doi: 10.1093/bioinformatics/bts565 PMID: 23060610

28. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22: 1658–1659. doi: 10.1093/bioinformatics/btl158 PMID: 16731699

29. Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics. BioMed Central Ltd; 2007; 8: 64.

30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. Nature Publishing Group; 2012; 9: 357–359.

31. Glass EM, Meyer F. The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes. Handb Mol Microb Ecol I Metagenomics Complement Approaches. John Wiley and Sons; 2011; 9: 325–331. PMID: 18803844

32. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014; 42: D490–5. doi: 10.1093/nar/gkt1178 PMID: 24270786

33. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009; 37: D141—5. doi: 10.1093/nar/gkn879 PMID: 19004872

34. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011; 28: 2731–2739. doi: 10.1093/molbev/msr121 PMID: 21546353

35. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008; 320: 1034–9. doi: 10.1126/science.1153213 PMID: 18497287

36. Schmidt MWI, Torn MS, Abiven S, Dittmar T, Guggenberger G, Janssens I a, et al. Persistence of soil organic matter as an ecosystem property. Nature. 2011; 478: 49–56. doi: 10.1038/nature10386 PMID: 21979045

37. Treseder KK, Balser TC, Bradford M a., Brodie EL, Dubinsky E a., Eviner VT, et al. Integrating microbial ecology into ecosystem models: Challenges and priorities. Biogeochemistry. 2012; 109: 7–18. doi: 10.1007/s10533-011-9636-5

38. Keenleyside WJ, Clarke AJ, Whitfield C. Identification of residues involved in catalytic activity of the inverting glycosyl transferase WbbE from Salmonella enterica serovar borreze. J Bacteriol. 2001; 183: 77–85. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC94852/pdf/jb000077.pdf doi: 10.1128/JB.183.1.77-85.2001 PMID: 11114903

39. Yin B, Crowley D, Sparovek G, De Melo WJ, Borneman J. Bacterial Functional Redundancy along a Soil Reclamation Gradient. Appl Environ Microbiol. American Society for Microbiology; 2000; 66: 4361–4365. doi: 10.1128/AEM.66.10.4361-4365.2000

40. Rousk J, Brookes PC, Baath E. Contrasting Soil pH Effects on Fungal and Bacterial Growth Suggest Functional Redundancy in Carbon Mineralization. Appl Environ Microbiol. American Society for Microbiology; 2009; 75: 1589–1596. doi: 10.1128/AEM.02775-08 PMID: 19151179

41. Barns SM, Cain EC, Sommerville L, Kuske CR. Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. Appl Environ Microbiol. 2007; 73: 3113–3116. doi: 10.1128/AEM.02012-06 PMID: 17337544

42. Gaby JC, Buckley DH. A comprehensive evaluation of PCR primers to amplify the nifH gene of nitrogenase. PLoS One. Public Library of Science; 2012; 7: e42149. doi: 10.1371/journal.pone.0042149 PMID: 22848735

43. Gittel A, Bárta J, Kohoutova I, Schnecker J, Wild B, Čapek P, et al. Site-and horizon-specific patterns of microbial community structure and enzyme activities in permafrost-affected soils of Greenland. Front Microbiol. 2014; 5.

44. Tveit A, Schwacke R, Svenning MM, Urich T. Organic carbon transformations in high-Arctic peat soils: key functions and microorganisms. ISME J. Nature Publishing Group; 2012; 7: 299–311. doi: 10.1038/ismej.2012.99 PMID: 22955232

45. Schellenberger S, Kolb S, Drake HL. Metabolic responses of novel cellulolytic and saccharolytic agricultural soil Bacteria to oxygen. Environ Microbiol. Blackwell Publishing Ltd; 2010; 12: 845–861. doi: 10.1111/j.1462-2920.2009.02128.x PMID: 20050868

46. Biely P. Microbial carbohydrate esterases deacetylating plant polysaccharides. Biotechnol Adv. 2012; 30: 1575–1588. Available: http://ac.els-cdn.com/S0734975012000869/1-s2.0-S0734975012000869-main.pdf?_tid=28be4c16-84ef-11e6-a050-00000aacb362&acdnat=1475007519_96af62cbcdba244999a390ce297a7221 doi: 10.1016/j.biotechadv.2012.04.010 PMID: 22580218

47. Doolette AL, Smernik RJ, Dougherty WJ. Overestimation of the importance of phytate in NaOH–EDTA soil extracts as assessed by 31 P NMR analyses. Org Geochem. 2011; 42: 955–964.

48. He Z, Olk DC, Cade-Menun BJ. Forms and lability of phosphorus in humic acid fractions of Hord silt loam soil. Soil Sci Soc Am J. 2011; 75: 1712–1722.

49. Young EO, Ross DS, Cade-Menun BJ, Liu CW. Phosphorus speciation in riparian soils: A phosphorus-31 nuclear magnetic resonance spectroscopy and enzyme hydrolysis study. Soil Sci Soc Am J. 2013; 77: 1636–1647.

50. Abdi D, Cade-Menun BJ, Ziadi N, Parent L-É. Long-term impact of tillage practices and phosphorus fertilization on soil phosphorus forms as determined by P nuclear magnetic resonance spectroscopy. J Environ Qual. 2014; 43: 1431–1441. doi: 10.2134/jeq2013.10.0424 PMID: 25603090

51. Miltner A, Bombach P, Schmidt-Brücken B, Kästner M. SOM genesis: microbial biomass as a significant source. Biogeochemistry. Springer Netherlands; 2012; 111: 41–55. doi: 10.1007/s10533-011-9658-z

52. Simpson AJ, Simpson MJ, Smith E, Kelleher BP. Microbially Derived Inputs to Soil Organic Matter: Are Current Estimates Too Low? Environ Sci Technol. American Chemical Society; 2007; 41: 8070–8076. doi: 10.1021/es071217x

53. da Jesus EC, Susilawati E, Smith SL, Wang Q, Chai B, Farris R, et al. Bacterial communities in the rhizosphere of biofuel crops grown on marginal lands as evaluated by 16S rRNA gene pyrosequences. Bioenergy Res. 2010; 3: 20–27. doi: 10.1007/s12155-009-9073-7

54. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci U S A. 2012; 109: 21390–5. www.pnas.org/cgi/doi/10.1073/pnas.1215210110 PMID: 23236140