Methodologies for probing the metatranscriptome of grassland soil

Aaron Garoutte  
*Michigan State University*

Erick Cardenas  
*University of British Columbia*

James Tiedje  
*Michigan State University*

Adina Howe  
*Iowa State University, adina@iastate.edu*

Follow this and additional works at: [http://lib.dr.iastate.edu/abe_eng_pubs](http://lib.dr.iastate.edu/abe_eng_pubs)  
Part of the Agriculture Commons, Bioresource and Agricultural Engineering Commons, and the Environmental Microbiology and Microbial Ecology Commons

The complete bibliographic information for this item can be found at [http://lib.dr.iastate.edu/abe_eng_pubs/784](http://lib.dr.iastate.edu/abe_eng_pubs/784). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Methodologies for probing the metatranscriptome of grassland soil

Abstract
Metatranscriptomics provides an opportunity to identify active microbes and expressed genes in complex soil communities in response to particular conditions. Currently, there are a limited number of soil metatranscriptome studies to provide guidance for using this approach in this challenging matrix. Hence, we evaluated the technical challenges of applying soil metatranscriptomics to a highly diverse, low activity natural system. We used a non-targeted rRNA removal approach, duplex nuclease specific (DSN) normalization, to generate a metatranscriptomic library from field collected soil supporting a perennial grass, Miscanthus x giganteus (a biofuel crop), and evaluated its ability to provide insight into its active community members and their expressed protein-coding genes. We also evaluated various bioinformatics approaches for analyzing our soil metatranscriptome, including annotation of unassembled transcripts, de novo assembly, and aligning reads to known genomes. Further, we evaluated various databases for their ability to provide annotations for our metatranscriptome. Overall, our results emphasize that low activity, highly genetically diverse and relatively stable microbiomes, like soil, requires very deep sequencing to sample the transcriptome beyond the common core functions. We identified several key areas that metatranscriptomic analyses will benefit from including increased rRNA removal, assembly of short read transcripts, and more relevant reference bases while providing a priority set of expressed genes for functional assessment.

Keywords
Metatranscriptome, Metagenome, Switchgrass, Short read assembly

Disciplines
Agriculture | Bioresource and Agricultural Engineering | Environmental Microbiology and Microbial Ecology

Comments
This is the accepted manuscript of an article published in Journal of Microbiological Methods, 131 (December 2016): 122-129, http://dx.doi.org/10.1016/j.mimet.2016.10.018.

Rights
© 2016. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.
Methodologies for probing the metatranscriptome of grassland soil
Aaron Garoutte*,1, Erick Cardenas2, James Tiedje1,3, & Adina Howe1,4

1 Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, United States of America
2 Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada
3 Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan, United States of America
4 Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa, United States of America
* Corresponding Author, Email: garoutte@msu.edu (AG)

Abstract
Metatranscriptomics provides an opportunity to identify active microbes and expressed genes in complex soil communities in response to particular conditions. Currently, there are a limited number of soil metatranscriptome studies to provide guidance for using this approach in this challenging matrix. Hence, we evaluated the technical challenges of applying soil metatranscriptomics to a highly diverse, low activity natural system. We used a non-targeted rRNA removal approach, duplex nuclease specific (DSN) normalization, to generate a metatranscriptomic library from field collected soil supporting a perennial grass, Miscanthus x giganteus (a biofuel crop), and evaluated its ability to provide insight into its active community members and their expressed protein-coding genes. We also evaluated various bioinformatics approaches for analyzing our soil metatranscriptome, including annotation of unassembled transcripts, de novo assembly, and aligning reads to known genomes. Further, we evaluated various databases for their ability to provide annotations for our metatranscriptome. Overall, our results emphasize that low activity, highly genetically diverse and relatively stable microbiomes, like soil, requires very deep sequencing to sample the transcriptome beyond the common core functions. We identified several key areas that metatranscriptomic analyses will benefit from including increased rRNA removal, assembly of short read transcripts, and more relevant reference bases while providing a priority set of expressed genes for functional assessment.
**Introduction**

Metatranscriptomics holds promise for providing insight into which organisms are active and which gene subsets are expressed within microbial communities, but its use is particularly challenging in complex systems, especially soil. Metatranscriptomics has been most prevalently used in marine ecology studies, where, as examples, it has helped identify key nutrient transformations in hydrothermal plumes [1]; patterns of niche diversification in coastal waters [2]; seasonal and diurnal patterns of gene expression in the English Channel [3] and patterns of diazotroph diversity along salinity and nutrient gradients [4]. In contrast, the application of metatranscriptomics in terrestrial environments has been limited, mostly either targeting specific genes (e.g., phylogenetic markers or functional genes), [5,6] experimentally enriched soil communities [7,8] or in greenhouse pot-based experiments [9]. In forest soils, fungal-targeted metatranscriptomics has been used to identify novel hydrolase enzymes [5], and a targeted approach (16S rRNA, ITS, and cellobiohydrolase) has shown that low-abundance species play an important role in carbon decomposition [7]. Metatranscriptomics has also been used to contrast expression in pristine soils and those contaminated with polycyclic aromatic hydrocarbons [10] and domain-level changes in the rhizosphere of potted plants [11]. While these examples demonstrate the feasibility and usefulness of soil transcriptomics, the application of non-targeted metatranscriptomics to field collected agricultural soils, e.g., croplands and pastures, has yet to be demonstrated; these soils comprise over 40% of global land use [12] and are essential to food production and ecosystem services.

Soil metatranscriptomics presents several obstacles. First, soil microbial communities are incredibly diverse; one gram of soil is estimated to contain nearly one million distinct genomes [13], magnitudes higher than aquatic and host-associated habitats [14]. Second, reference genomes from soil are limited, making sequence annotation difficult. Third, RNA, especially mRNA, is in low abundance because of the primarily dormant or starved state of the community members, with few perturbations to induce expression. For example, turnover rates of soil microbes has been calculated to be 30- to 300-fold slower than that of microbes in the ocean [15]. Overall, the mRNA comprises only about 4% of total RNA [16], highlighting the challenge of isolating or enriching the mRNA prior to sequencing to achieve greater sequence depth. A common approach for mRNA enrichment is to remove rRNA through subtractive hybridization [17]. This approach presents challenges of its own in that it is hindered by the difficulty of obtaining intact rRNA through soil RNA extraction methods. Finally, soil metatranscriptomics is challenged by the high temporal and spatial diversity in soil populations due to habitat complexity at small scales (<1 mm), and various stochastic perturbations (e.g., rainfall, plant litter introduction, micro and mesofauna movements). Consequently, capturing appropriate snapshots of targeted activity in soil requires sampling high biodiversity within complex
and often unknown and unpredictable dynamics. Furthermore, the lack of the soil metatranscriptome reference datasets makes it difficult to evaluate appropriate sampling and experimental strategies and for insight into common system responses.

In this study, we evaluated mRNA enrichment as well as various bioinformatic approaches to analyze soil metatranscriptomes, and present recommendations for such studies. Our metatranscriptome originated from bulk soil associated with a *Miscanthus x giganteus* crop, an important bioenergy crop due to its perenniability and its high biomass yield compared to other crops [18]. Soils were sampled at the period of most active plant growth (early August), at midday when photosynthesis was maximum, and after a rainfall period so that soil water was not limiting to maximize soil microbial community response to potentially new substrates. Our objective was to determine our ability to access and identify actively transcribed genes in this soil microbial community from the soil’s metatranscriptome.

To address the challenge of low concentration of mRNA, we enriched for mRNA by removing rRNA using duplex nuclease specific (DSN) normalization. DSN is a non-targeted approach which has several advantages over subtractive hybridization including less stringent RNA quality requirements, a lower required amount of RNA (100 ng compared to 1 ug), and an increased rRNA removal efficiency [17]. An alternative to DSN is subtractive hybridization, which targets the removal of specific rRNAs based on genomic primer targets. In contrast to subtractive hybridization, which removes rRNAs with bias, DSN does not target specific rRNAs and consequently the remaining rRNAs are more likely to reflect their original distributions. To assess the gene content of our metatranscriptome, we annotated against several gene or genome reference databases, including MG-RAST M5NR, the Carbohydrate Active Enzyme (CAZy) database, and a soil genome dataset termed RefSoi, and three *de novo* assembled metagenomes obtained from the same plot during the Spring of 2009. Additionally, we assessed the value of assembly of longer sequences, or contigs, from the metatranscriptome for improved insight into community activity.

**Methods**

Various methods that can be used for soil metatranscriptome analysis including (A) direct annotation of short read sequences (B) assembly and annotation and (C) alignment of reads to existing reference genomes are summarized in Figure 1 along with the advantages and disadvantages of each. We analyzed our dataset set using all three of these methods. We compared the results of each method to the others to determine best practices and to identify the advantages and disadvantages of each method.

**Fig. 1: Metatranscriptome data analysis workflow.** Various methods for metatranscriptome data analysis are shown. (a) Direct annotation of short reads. (b)
Assembly of short reads into longer contigs and subsequent annotation. (c) Short read mapping to genomes compiled in the RefSoil database.

Metatranscriptome sample collection and library preparation
The bulk soil samples for metatranscriptomics were obtained from a four-year-old stand of Miscanthus (*Miscanthus x giganteus*), plot G6R1 at the Bioenergy Cropping Systems Experiment (BCSE; also known as the Intensive Site) (42°23'47"N, 85°22'26"W) at the W.K. Kellogg Biological Station in Southwest, Michigan, USA. Samples were collected midday on August 1st, 2012. The mean air temperature for the previous week was 24 °C, and there had been 117 mm of rain in the week preceding the sampling with 2 days of no rain directly prior to sampling; the soil was still moist. A composite sample comprised of three soil samples was taken from random points in the plot. The soil was quickly sieved (4 mm) to remove roots, and frozen on dry ice to prevent mRNA degradation. Samples were stored at -80 °C until RNA extraction. RNA was extracted from 2 g of soil using the PowerSoil RNA kit (MoBio, Carlsbad, CA), and DNA was then removed by DNase treatment (Invitrogen, Carlsbad, CA). RNA (100 ng) was converted to cDNA and treated with duplex specific nuclease (DSN) to reduce the abundance of rRNA as described in [17]. Samples were sequenced with the Illumina HiSeq sequencing platform at the Research Technology Support Facility, Michigan State University, East Lansing, MI, USA, generating 100 base pair (bp) reads.

**MG-RAST databases used for annotation of unassembled raw reads**

Ribosomal RNA sequences were identified using riboPicker [21] (Figure 1a), Rfam [22] databases and MG-RAST [23]. The resulting non-rRNA sequences were submitted to MG-RAST (v 3.3.7.3) using the M5NR[24] for gene annotation. Many reads were annotated as *Enterobacteria phage phiX174*, which is commonly used as a control in sequencing facilities. Hence, the sequences were mapped to *Enterobacteria phage phiX174 sensu lato* genome (NC_001422.1) using Bowtie 2 (v2.0.0-beta6, [21]) and removed from the analysis as they were likely the result of contamination. Additionally rRNA sequences within the unassembled database were annotated using the MG-RAST M5RNA database (MGRAST IDs 4554103.3, Unassembled Metatranscriptome). Annotations were identified using the following preset quality filter parameters: Max. e-value cutoff 1e-5, Min. percent identity cutoff 60% and Min. Alignment length cutoff of 15.

**Metatranscriptome assembly & annotation**

Sequences were filtered using digital normalization (flags: -C 20, -k 20, N 4, -x 2e9) as described in [27–29]. Normalized reads were assembled using Velvet (v 1.2.10) [30] with odd numbered k-mers from length 19 to 59 (Figure 1B). Assemblies produced from
different k-mer lengths were merged using AMOS (v 3.1.0) [31] and CD-HIT (v 4.5.7) [32]. Resulting assembled contigs with lengths greater than 200 bp were annotated with MG-RAST (v 3.3.7.3) [24] (MGRAST IDs 4532564.3 Assembled Metatranscriptome), the CAZy database (date accessed: July 13, 2008) [33] which contains enzymes involved in carbon compound synthesis and decomposition, and the Rfam database which contains non-coding RNAs.

**Previous metagenomes used for comparison to soil metatranscriptome**

The reference metagenomes used in this study were obtained from one bulk and two rhizosphere soil samples collected from the same Miscanthus bioenergy plot in October 2009. DNA was extracted from 2.5 g soil as described in [19]. The high molecular weight DNA was then gel purified, electroeluted, and concentrated using methods described in [20]. Samples were sequenced with both the Illumina GAII and 454 sequencing platforms at the Joint Genome Institute (Walnut Creek, CA) generating 100-base reads.

**Metagenome assembly & annotation**

Sequences were trimmed using a quality score of 20. Reads were assembled using SOAPdenovo[34] with a k-mer range of 21, 23, 25, 27, 29 and 31. All of the default settings were used for the SOAPdenovo assembly (flags –d1 and –R). Contigs were then merged using SGA [35] with all default parameters. Contigs greater than 500 bp were annotated by MG-RAST (v3, 2011-02-22) (MGRAST IDs 4465947.3, Bulk MetaG, 4465942.3, Rhizo MetaG1, and 4465943.3; Rhizo MetaG2).

**Estimation of abundance of assembled contigs or reference sequences**

The abundance of assembled contigs and reference sequences (e.g., soil genomes) was estimated as the median base pair coverage of all transcript alignments to contigs (assembled metatranscriptome and reference metagenomes) or genomes with the RefSoil database. Mapping of unassembled metatranscriptome reads to contigs or genomes was performed using Bowtie2 (v2.0.0-beta6, [25]) with the following default parameters: end-to-end alignment, minimum score threshold for 100 bp reads was -60.6, –D 100, distinct alignments for each read. Base pair coverage was estimated using BedTools (v 2.17.0) [26]. For metagenomic, metatranscriptomic, and soil reference annotated genes, coverage was estimated on the genic region rather than the complete originating contig (which may contain both genic and intergenic sequences).
Curation of soil reference genome database, RefSoil

A manually curated database of soil bacterial genomes was built to provide a soil-specific reference set. Strains with completely sequenced genomes were selected from the Gold Database (http://genomesonline.org) on August 19th, 2011. The inclusion criteria involved both information on isolation of the sequenced organism and literature searches regarding the ecology of the species. e.g. Erwinia amylovora CFBP 1430 was selected even when it was originally isolated from a Crataegus plant because it is commonly detected in soils. Obligate human pathogens and non-soil relevant extremophiles were excluded. If redundant genomes were found at the species-level, only two per species were kept to reduce the database bias. A total of 492 organisms, representing 19 different phyla and contributing a total of 1,031 replicons (chromosomes and plasmids) formed the database. Complete GenBank accessions were downloaded and parsed to extract whole genome sequences and features (gene coordinates, and annotations) (Supp Table 1). A complete list of genomes and accession numbers used in the RefSoil database is in Supp Table 1.

Unassembled read mapping to metagenomes and RefSoil genomes

Sequences were mapped to the three metagenome assemblies and the genomes within the RefSoil database using Bowtie 2 (v2.0.0-beta6, [25]) with the following default parameters: end-to-end alignment, minimum score threshold for 100 bp reads was -60.6, -D 100, distinct alignments for each read (Figure 1C). Coverage of annotated regions was estimated using BedTools (v 2.17.0) [26]. Only reads with a minimum alignment length of 100 bp (to references) and contigs (or genes/genomes) with at least two mapped reads were considered.

Results

Characterization of sequences in the unassembled soil metatranscriptome

Genes identified in transcripts include sequences associated with both rRNA and mRNA genes, informative of the active community structure and function. The large majority of transcripts, 169 million reads (82.8%), shared similarity to known rRNA genes. The justification of using the DSN approach for mRNA enrichment over probe-based rRNA removal was the unbiased removal of rRNA gene fragments. Using the DSN metatranscriptome library preparation, the remaining rRNA sequences were evaluated to determine the taxonomic composition of the active community members, resulting in nearest matches to over 22,000 species in our metatranscriptome. The most abundant “active” bacterial phyla were Actinobacteria and Proteobacteria (Figure 2, blue bar), and sequences associated with Ascomycota were the most abundant fungal phylum.
Fig. 2: **Phylogenetic distribution of sequence annotations identified in unassembled and assembled metatranscriptome and associated soil metagenomes.** Phylogeny of rRNA from the unassembled metatranscriptome compared to the phylogeny of MG-RAST’s best-hit classification of protein-coding genes for the assembled metatranscriptome and the reference metagenomes.

* The pink bar in the Assembled MetaT (mRNA) Firmicutes represents the proportion of misannotation in the sample (explained below).

In unassembled transcripts sharing sequence similarity with known genes, the most abundant protein coding annotations from the SEED database were associated with hypothetical (3.7%) or housekeeping proteins such as GroEL (2.9%), DNA-directed RNA-polymerase (1.9%), and the translation elongation factor-Tu (1.8%). These non-rRNA sequences account for 179,088 reads, representing 8,345 protein coding genes (Table 1). Similar functional profiles were also obtained when annotating the unassembled metatranscriptome against other databases in MG-RAST including GenBank, KEGG, and RefSeq (Table 1, Supp Table 2-5). Overall, the reads comprising the unassembled metatranscriptome were associated with a few dominant annotations, where the five most abundant annotations represented 12% of the total abundance of annotations our metatranscriptome.

**Table 1. Summary of sequence annotations of the unassembled and assembled soil metatranscriptome against various reference databases.** Results of annotation by MG-RAST.
Note that annotations are defined within the MG-RAST M5NR database, where distinct annotations may be represented by multiple features. Features may be associated with a specific gene in a reference genome.

To further explore both the taxonomic and functional content of the metatranscriptome, transcripts were also compared against the RefSoil database, which resulted in 94 million reads aligning to RefSoil genomes, the large majority of which were associated with rRNA gene annotations (Table 2). Similar to SEED-associated annotations, the most represented functions of the metatranscriptome in the RefSoil database were associated with hypothetical proteins, ribosomal structure, and housekeeping genes.

Overall, the most abundantly represented RefSoil genomes in the soil metatranscriptome included *Syntrophus aciditrophicus* SB, *Methylococcus capsulatus* str. Bath, and *Novosphingobium* sp. PP1Y (Supp Table 6). The most genes (e.g., presence rather than abundance) were identified in genomes of *Nocardioides sp.* JS614, *Bradyrhizobium japonicum* USDA 110, and *Streptomyces scabiei* 87.22 (Supp Table 7).

**Table 2: Summary of transcript mapping.** Transcripts mapped to reference assemblies (available in MG-RAST with IDs indicated) or genomes with proportion of reads identified as similar to rRNA genes and mapping uniquely to a specific reference assembly.
Characterization of sequences in the assembled soil metatranscriptome

Assembly of the metatranscriptome was highly successful and incorporated 73.8% of the reads into 116,556 contigs totaling 32.4 Mbp. In contrast to the unassembled metatranscriptome, the majority of assembled contigs (78.3%) were not associated with rRNA genes. Overall, a total of 15,032 (13.3%) contigs shared sequence similarity with known proteins in the SEED database (Figure 3, Supp Table 8-11). To estimate abundance of assembled sequences, unassembled reads were aligned to contigs, and the median base pair coverage of each contig was calculated. The most abundant gene functions identified within the soil metatranscriptome assembly included those associated with hypothetical proteins or with functions associated with RNA and protein metabolism (Figure 4 dark blue and red lines only). Comparing the number of annotations identified with the unassembled and assembled metatranscriptome datasets; we found a larger number of annotations in the unassembled dataset (Figure 5). However, the five most abundant subsystems in the SEED annotations were shared between the assembled and unassembled metatranscriptome datasets though with differing rank abundances (Figure 4).

Fig. 3: Distribution of assembled metatranscriptome annotations. Proportion of assembled metatranscriptome sequences associated with known rRNA, gene function (SEED), or non-coding sequences (RFAM).
Fig. 4: Comparison of functional profiles of metatranscriptomes and metagenomes. Annotations were identified in the assembled and unassembled metatranscriptome datasets as well as the three metagenome assemblies against the MG-RAST SEED database.
Fig. 5: Comparison of the total number of gene annotations identified in the unassembled and assembled metatranscriptomes. Results were generated using the MG-RAST Metagenome Analysis page.

Community composition represented by the assembled metatranscriptome was identified by comparing the contigs to the taxonomic origins of proteins in the MG-RAST M5NR, resulting in the identification of 2,200 species. Similar to the rRNA in the unassembled metatranscriptome dataset, the dominant phyla represented were *Actinobacteria* and *Proteobacteria*. In contrast to the unassembled metatranscriptome, *Firmicutes* also represented a large portion of protein annotations in the assembled dataset (Figure 2, red bar), but this was mainly due to hypothetical proteins associated with
Heliobacterium modesticaldum, Lactobacillus rhamnosus, and Staphylococcus aureus. The unique detection of abundant Firmicutes in the assembled dataset and its absence in unassembled transcripts suggest the presence of a database bias in rRNA gene sequences within the MG-RAST M5NR, and hence this likely annotation error (as noted by the different shading of the red bar in Figure 1). This is likely as the three previously mentioned organisms are associated with human disease and therefore comprise a larger portion of available genomes.

The large majority of contigs within the soil metatranscriptome (greater than 65%) could not be annotated with any of the reference databases used in this study (MG-RAST, RefSoil, CAZy, or associated metagenomes) (Figure 3). To evaluate the possible presence of non-coding RNAs, sequences were compared to known non-coding RNAs in the RFam database, resulting in a total of 3,036 contigs (2.6%) sharing similarity to RNA genes, regulatory RNAs, or self-splicing RNAs. The major RNA families identified included RNAs associated with transcription and translation (5/5.8S, tmRNA, and RNaseP), signal recognition particles, and riboswitches (Supp Table 12).

Further, longer sequence lengths of assembled contigs significantly improved annotations, doubling the median alignment lengths to known proteins (Figure 6). To assess the impacts of sequence length, we evaluated the influence of varying similarity thresholds. Stricter criteria for alignment scores (e.g., decreased minimum E-value cutoff) reduced the abundance and total number of unique features in the unassembled dataset. Overall, confidence in annotations (e.g., median E-value scores) for the assembled annotations were much higher (lower E-value) than for the unassembled (Figure 7), and variations in the E-value thresholds did not have as pronounced an effect on the total number of annotations nor the number of unique features. Importantly, assembled contigs provides longer sequence lengths for annotation (62 aa vs. 31 in the unassembled set), allowing for improved annotations (e.g., similarity comparisons to CAZy enzymes) (Figure 6). In total, assembly resulted in 688 contigs, comprising 194,985 bp, which could be classified into five enzyme categories including glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases (PL), and carbohydrate esterases (CE). The large majority of these sequences (572 contigs, 83%) were associated with GH, GT, and CBM-containing enzymes. The most frequent CAZy gene families were GT2, GH36, CBM13, and GH18 (Supp Table 12). Overall, these CAZy-associated contigs were present at relatively low abundances within the metatranscriptome, averaging 4.1-fold coverage. The most abundant enzymes classes included GH19, GH17, and CBM14 with 137, 47, and 19-fold coverage, respectively (Supp Table 13).

**Fig. 6: Comparison of annotation alignment lengths of the assembled and unassembled datasets.** Amino acid alignment lengths of SEED subsystem annotations for the assembled and unassembled datasets. The minimum alignment length is set to the MG-RAST default of 15 amino acids.
Fig. 7: Comparison of annotation E-values of the assembled and unassembled datasets. E-value of SEED Subsystem annotations of the assembled and unassembled datasets. The minimum e-value is set to the MG-RAST default of 1e-5.
Comparison of the metatranscriptome datasets to metagenomes

To provide further insight into the active subset of microbial communities, we evaluated the membership identified in the soil metatranscriptome (gene expression) and compared this to membership identified in the soil metagenomes (gene potential). Assemblies of the reference metagenomes produced 1.3 million contigs and represent over one billion bases (Table 3). Approximately 30 to 40 million unassembled metatranscriptome reads mapped to these metagenomes, with the majority of these reads being rRNA (98%, Table 2). The remaining non-rRNA transcripts mapped to a total of 147 genes, the most abundant were related to housekeeping functions, e.g., RNA polymerase, chaperone proteins, and translation elongation factors. The functional profiles of the assembled metatranscriptome and metagenomes from the same site revealed that the metatranscriptome was greatly enriched in genes related to RNA and protein metabolism. In contrast, the metagenomes were enriched in genes related to carbohydrate, amino acid (and derivatives), and DNA metabolism (Figure 4). The overlap of functional annotations between the assembled metagenomes and the assembled metatranscriptome (e.g., at the functional level) was comprised of 2,413 annotations (62% of the metatranscriptome). Comparing taxonomic profiles of the metatranscriptomes to those of the metagenomes, we found that sequences associated with Proteobacteria were enriched in the metagenomes, while sequences associated with Actinobacteria were enriched in both the assembled and unassembled metatranscriptomes (Figure 2).
Table 3: Summary of assembled metagenomes in number of assembled contigs and total base pairs represented in the assembly. Results of short read assembly of metagenome samples.

<table>
<thead>
<tr>
<th>Assembled metagenomes</th>
<th>Contigs</th>
<th>Base Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaG Bulk</td>
<td>617,602</td>
<td>457,810,820</td>
</tr>
<tr>
<td>MetaG Rhizo1</td>
<td>303,353</td>
<td>216,957,151</td>
</tr>
<tr>
<td>MetaG Rhizo2</td>
<td>453,481</td>
<td>360,952,806</td>
</tr>
<tr>
<td>Total</td>
<td>1,374,436</td>
<td>1,035,720,777</td>
</tr>
</tbody>
</table>

Discussion

Our aim was to use metatranscriptomics to assess biological information of a (normally) marginally active soil microbiome and to understanding the technical and methodological challenges of this approach. Towards this end, we assessed approaches to generate a soil metatranscriptome library (e.g., mRNA enrichment), analysis approaches (e.g., de novo assembly), and the gene content of the dataset. Overall, we identified multiple expressed genes in our soil metatranscriptome (Figure 4), though it was largely dominated by ribosomal rRNA genes as well as sequences of unknown origin and function (Table 2, Figure 3). To increase the information gleaned from soil metatranscriptomics in the future, we identify below several areas for improvement.

The abundance of rRNA in metatranscriptomes must be further reduced in order to improve sampling of mRNA encoding protein-coding genes. The large proportion of rRNA within our soil metatranscriptome library compromised our ability to sample deeply and consequently access more protein-coding genes. Although DSN normalization was expected to remove diverse rRNA, it did not with 83% rRNA remaining in our metatranscriptome library. This fraction is comparable to rRNA removal efficiency in a human gut metatranscriptome following subtractive hybridization [37] and in a sandy soil metatranscriptome with no rRNA removal [17]. Direct comparison of RNA extraction efficiencies in the two soils may not be appropriate because of different soil characteristics and the sampling season; their much lighter textured (sandy) soil was sampled in winter, and our medium textured (loamy) soil was sampled in late summer [38]. In general, reports on rRNA remaining based on multiple approaches and environments vary from 50 to 85% [2,37–39], evidence that rRNA removal in metatranscriptomes remains inefficient for complex communities, regardless of extraction methodology. Though DSN normalization has improved performance compared to subtractive hybridization in pure cultures, its effectiveness in high diversity soil systems remains unclear. An alternative approach is to bypass rRNA removal and to sequence more deeply and computationally remove rRNA reads. This approach is more feasible as sequencing prices decrease.
A useful result of the presence of rRNA in our metatranscriptome was that it did allow us to make taxonomic inferences about active members of the community. Unlike samples prepared using subtractive hybridization, the DSN normalization preserves the relative abundance of sequences within the sample [17]. Since the relative abundance of sequences is preserved taxonomic annotations associated with the remaining rRNA sequences (in unassembled reads) is reflective of the relative abundances in the original sample. Notably, ribosomal RNA sequences from the assembled metatranscriptome dataset were not used because assemblers typically cannot assemble highly conserved sequences like 16S rRNA genes. Therefore, as a proxy, the taxonomic classification of the most similar known homologous protein was used for community analysis of the assembled metatranscriptome. Taxonomic annotation of both metatranscriptome datasets ( unassembled rRNA and assembled protein coding contigs) suggests that they share a similar taxonomic profile that contrasts with those observed in the metagenomes, highlighting the increased activity of sequences associated with Actinobacteria and diminished activity of sequences associated with Proteobacteria. This result is consistent with other findings that indicate Actinobacteria are more abundant and active in bulk soils while Proteobacteria tend to be more abundant in the rhizosphere [41,42].

The curation and availability of the RefSoil database allows for the evaluation of sequencing datasets in the context of cultivated soil organisms. Despite the diversity of soil microbial communities and the difficulty of cultivating microbial representatives, this database was surprisingly represented within our soil metatranscriptome. Many transcripts could be aligned to RefSoil genomes, although most were associated with rRNA genes. This result suggests that the RefSoil database captures a large amount of the SSU rRNA (taxonomic) diversity in our sample. The functions contained with RefSoil were not nearly as well represented in the soil metatranscriptome, suggesting that although this database may capture many of the genus-to-species level of diversity, the genetic diversity within those groups is still very large.

We found de novo assembly of this soil metatranscriptome to be an important step towards providing improved references for soil sequencing approaches, evidenced by longer sequence lengths, data reduction, improved confidence in annotation, and the development of reference sequences that do not rely on a priori information. Previously, the high diversity of soil communities has resulted in only a fraction of sequences being assembled in metagenomic studies [28]. For this soil metatranscriptome, 73.8% of the reads in the unassembled dataset mapped to our assembly, suggesting that the diversity of soil metatranscriptomes is significantly less than that of metagenomes. As a consequence, if rRNA can be efficiently removed prior to sequencing, metatranscriptomic efforts may require less sequencing depth than previously suggested by soil metagenomes. The longer sequence lengths provided by the assembly also provide higher confidence in annotations as well as the identification of multiple novel and abundant sequences. Importantly, our metatranscriptome assembly provides a specific set of genomic references that can be used
for comparative soil studies. The presence of shared (highly) expressed sequences in multiple datasets can be used to prioritize encoded genes for characterization. An indirect advantage to soil metatranscriptome assembly is also that it discards many rRNA-associated sequences because these sequences are difficult to assemble, allowing it to be used as a method for rRNA removal that does not rely on having known references. As datasets continue to grow in volume, assembly may become an increasingly efficient method for both improving gene annotation and removing rRNA.

We evaluated the novel information gained through our metatranscriptome by comparing our soil metatranscriptome to available metagenomes from the same plot. The majority of genes were unshared between these datasets though the majority of encoded functions were similar. Within functional annotations, 62% of the metatranscriptome annotations were shared with the metagenomes. However, many of these were associated with rRNA genes; relatively few non-rRNA transcripts (~2.7 million) were aligned to the metagenomes, with only 3.7% of non-rRNA annotations shared. This result indicates that the metatranscriptome is functionally similar to the metagenomes but is composed of distinct genotypes with high levels of functional redundancy between members. Previous studies have also shown little overlap between metagenome and metatranscriptome libraries [42,43]. A possible explanation for this observation is a change in the microbial communities over the time (2 years) between sampling the soil metatranscriptome and metagenomes. While changes during this time are very likely, we expect that these changes are relatively small in the metagenomes as soil microbial populations are thought to have turnover times ranging from 6.8 to 0.24 years [15,44,45]. Another possible explanation for the low overlap between samples is the soil subhabitat (metatranscriptome of bulk vs. metagenome of rhizosphere). Rhizosphere soils, generally, contain more active communities compared to bulk soils [10]. Differences in the sequencing depth of the metatranscriptomic and metagenomic efforts may have also contributed to differences in these datasets. A final explanation for the distinct communities identified between these sequencing efforts is that the biologically active communities in the soil may not be represented in metagenomes due to under sampling or spatial differences. In this case, metagenomic libraries may be most useful for generating gene references reflecting possible soil diversity, while metatranscriptomics may be most appropriate for targeting active communities.

Overall, our metatranscriptome was dominated by sequences that could not be associated with genes that have previously been studied and for which no function is known (e.g., hypothetical proteins). Abundant hypothetical proteins are observed in other metatranscriptomes [40,44]. Insight into these sequences (these “known unknowns”) is necessary to determine if they generally play an important role in function. Additionally, as increasing numbers of metatranscriptomes become available, the development of novel approaches that use unsupervised classification methods to identify patterns of codon
usage across microbial communities [47] or co-occurrence of sequences [48–50] within multiple datasets should prove useful in characterizing these sequences.

Conclusion

Based on our evaluation of this Miscanthus soil sample, soil metatranscriptomics holds promise for identifying actively transcribed genes in the soil. The methods for leveraging this technology still require much development to reach genes important to ecological fitness or ecosystem functions. From this relatively small sample (20 Gbp), we were able to produce an assembly that captured the majority of reads in the sequenced dataset. The resulting assembly allowed us to identify, with high confidence, several sequences similar to known genes and soil genomes that are actively transcribed and of interest to carbon cycling. The development of a soil specific database was helpful for analyzing our soil metatranscriptome, but a large majority of the assembled sequences still lack references in databases. This does show, however, the value of expanding the soil isolate genome and physiology database. Overall this study illustrates that metatranscriptomic sequencing can be preformed on samples of field collected soil.

Acknowledgements

We thank Tamara Cole for helpful discussions regarding this manuscript and Jeff Landgraf for troubleshooting the duplex specific nuclease normalization method. This work was funded by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494).

References


