SCIENTIFIC DATA

OPEN

Data Descriptor: A metagenomic survey of forest soil microbial communities more than a decade after timber harvesting

Received: 15 February 2017
Accepted: 15 May 2017
Published: 25 July 2017

Roland C. Wilhelm[1,*], Erick Cardenas[1,*], Hilary Leung[1], Kendra Maas[1], Martin Hartmann[2], Aria Hahn[1], Steven Hallam[1] & William W. Mohn[1,†]

The scarcity of long-term data on soil microbial communities in the decades following timber harvesting limits current understanding of the ecological problems associated with maintaining the productivity of managed forests. The high complexity of soil communities and the heterogeneity of forest and soil necessitates a comprehensive approach to understand the role of microbial processes in managed forest ecosystems. Here, we describe a curated collection of well replicated, multi-faceted data from eighteen reforested sites in six different North American ecozones within the Long-term Soil Productivity (LTSP) Study, without detailed analysis of results or discussion. The experiments were designed to contrast microbial community composition and function among forest soils from harvested treatment plots with varying intensities of organic matter removal. The collection includes 724 bacterial (16S) and 658 fungal (ITS2) amplicon libraries, 133 shotgun metagenomic libraries as well as stable isotope probing amplicon libraries capturing the effects of harvesting on hemicellulolytic and cellulolytic populations. This collection serves as a foundation for the LTSP Study and other studies of the ecology of forest soil and forest disturbance.

Design Type(s)	observation design • longitudinal data collection method
Measurement Type(s)	GeneSequence • whole genome sequencing • metagenomics analysis
Technology Type(s)	amplicon sequencing • Whole-Genome Shotgun Sequencing • radioisotope labeled DNA probe
Factor Type(s)	soil • temperature • hydrological precipitation process • environmental factor • experimental condition • isotopic labelling
Sample Characteristic(s)	soil metagenome • Province of Ontario • forest soil • State of California • State of Texas • Province of British Columbia

[1]Department of Microbiology and Immunology, Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3. [2]Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf CH-8903, Switzerland. *These authors contributed equally to this work. †Present address: Department of Microbiology and Immunology, Life Sciences Institute, 2350 Health Sciences Mall, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z1. Correspondence and requests for materials should be addressed to W.W.M. (email: wmohn@mail.ubc.ca).

Background & Summary

The Long-term Soil Productivity Study (LTSP) was initiated in 1989 to track changes in soil quality and productivity of managed forests. Foresters lacked data on the impact of growing trends in intensified forest management, such as shorter crop cycles, greater biomass extraction, and use of highly mechanized equipment[1]. Partners in the LTSP Study have since collected longitudinal data from a range of North America's most-productive forests (>100 sites) to assess the long-term effects of soil compaction and organic matter (OM) removal. The objective is to develop indices for monitoring soil quality based in physical, chemical, and biological properties of soil. Each LTSP site has replicated experimental plots for harvested treatments with three intensities of OM removal as well as unharvested reference plots. OM removal intensity corresponded to common harvesting strategies, such as debranching *in situ* (OM1) or removal of both trunks and branches (OM2), or the less common, and extreme, practice of removing trunks, branches and top soil (OM3). In the ensuing years, these treatments produced a consistent gradient in soil properties according to harvesting intensity, such as decreasing amounts of total carbon and nitrogen as well as increased dryness, mean daily temperature and temperature fluctuation. While the latest LTSP studies show that the effects of varying degrees of OM removal appear minor on net primary productivity[2–4], the metagenomic datasets presented here show consistent changes in bacterial and fungal populations that reflected harvesting intensity[5–9].

We collected data on soil microbial communities from all treatment plots at eighteen LTSP sites in six different ecozones between 2008 and 2014 when reforested stands were 11 to 17 years old (Fig. 1; Table 1). To capture the extent of harvesting impacts throughout the soil profile, corresponding samples from organic and mineral layers were included in all datasets. The goal was to identify changes in community composition, diversity, and functional potential resulting from the intensity of OM removal. We surveyed soil microbial communities in treatment plots using amplicon sequencing (Table 2 (available online only)) of bacteria (Data Citation 1) and fungi (Data Citation 2), along with shotgun metagenomes from all treatment plots at a single site within each ecozone (Data Citation 3; Table 3 (available online only)). Shotgun metagenomes revealed impacts on the functional potential of communities, such as decomposers involved in carbon cycling[5], which led to further targeted studies of hemicellulolytic[7] (Data Citations 4,5) and cellulolytic[8] populations (Data Citation 6) using ^{13}C-stable isotope probing (Table 4 (available online only)). Stable isotope probing (SIP) can be used to track populations that assimilate ^{13}C-label into their biomass by recovering and sequencing the resultant 'heavier'^{13}C-enriched DNA[10].

Here, we provide an overview of the data collection, without detailed analysis of results or discussion, to draw attention to its unparalleled comprehensive and multi-faceted view into forest soil microbial communities. The collection is the first high-throughput sequencing-based survey of LTSP sites, and, as such, provides base-line data for future LTSP investigations at an important, stage of forest regeneration just prior to canopy closure. Researchers can revisit this collection as our understanding of the ecology of forest soils advances. This will be important given the substantial proportion of unknown taxa in these collections affected by timber harvesting[9]. SIP data offers unique insights into the effects of timber harvesting on decomposer populations, including detailed information on uncultured taxa provided by the ten partial genomes recovered from SIP-cellulose shotgun metagenomes (Data Citation 6). The consistency in experimental design, sequencing methodology and sample sources ensures the value of this collection for on-going studies of forest soil microbial communities, in particular those pertaining to biogeography, soil strata and forest disturbance.

Methods

Experimental design

Soil samples were collected from reforested experimental plots within the Long-Term Soil Productivity (LTSP) Study from eighteen different sites across six conifer-dominated North American ecozones named after the predominant tree species (Fig. 1; Table 1): IDF_{BC} (interior Douglas-fir), SBS_{BC} (sub-boreal spruce), PP_{CA} (ponderosa pine), BS_{ON} (black spruce), JP_{ON} (jack pine) and LP_{TX} (loblolly pine). Ecozones were chosen to exemplify a broad range of climates and regions in North America where forestry is a major industry. These differed by several factors, including soil type, mean annual temperature and precipitation, tree species and bulk soil chemistry, such as carbon and nitrogen content and pH (Table 1). Each ecozone contained three sites with four treatments: REF (or OM0), a neighbouring unharvested reference plot; and three harvested treatments: OM1, where only tree boles (stems) were removed and woody debris was left in place; OM2, where whole trees including branches were removed and; OM3, where whole trees were removed and the upper organic layer of forest floor scraped away (Fig. 2). Compaction was controlled and, in all cases, plots with minimum compaction ('C0') were sampled. Moderate (C1) and severe (C2) compaction treatments were included in 16S rRNA gene and ITS pyrotag libraries in SBS_{BC} and IDF_{BC} (Table 2 (available online only)). Similarly, additional amplicon libraries were prepared from soils in JP_{ON} and LP_{TX} which had been exposed to glyphosate (Table 2 (available online only)). Samples were collected from triplicate plots at each of the three sites in BS_{ON} and JP_{ON}, while at sites in the other four ecozones triplicate samples were collected from a single, larger plot. Each sample corresponds to a composite from between three to five sampling points (consistent within a given ecozone) in a plot which helped account for soil heterogeneity. Organic layer

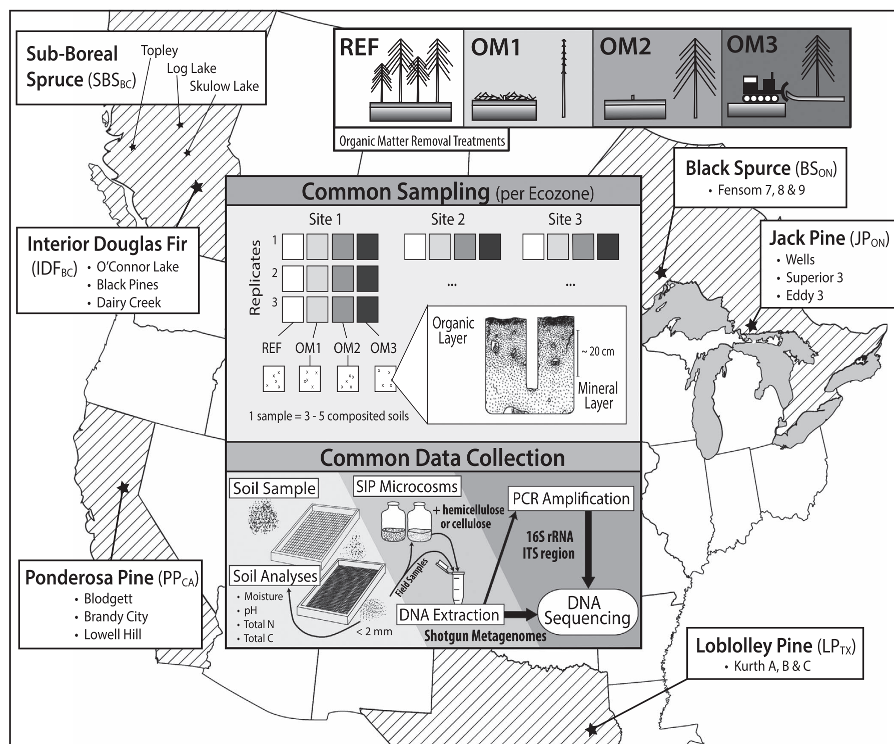

Figure 1. Locations of soil sampling and descriptions of data collection conducted for this study. The locations and names of eighteen North American sampling sites are shown grouped into six ecozones. The term 'ecozone' is used to refer to the distinct local assemblages of organisms and climatic factors between groupings of sites. An overview of the design for soil sampling along with data collection are superimposed on the map. OM removal treatments are shaded according to the intensity of OM removal.

samples (O-horizons) were first sampled with a trowel and then the top 20 cm of mineral layer soil (A and upper B-horizon) were collected using a Stoney auger (5 cm diameter). For several 16S rRNA gene amplicon and whole shotgun sequencing libraries from Skulow Lake (denoted in Tables 2 and 3 (available online only)), samples were collected from five soil horizons: LFH and mineral horizons (Ahe, Ae, AB and Bt), distinguished using criteria from the Canadian System of Soil Classification. Samples were stored at 4 °C during transport and until each sample was sieved through 2-mm mesh to remove roots, then stored at −80 °C until DNA was extracted within three months of sampling. Soil samples used in metatranscriptomics were flash frozen in liquid nitrogen, transported on dry ice and subsequently stored at −80 °C until DNA and RNA was extracted within 12 months of sampling.

Amplicon and shotgun metagenome sequencing

DNA was extracted from field samples (0.5 g of soil) using the manufacturer's recommended protocol for the FastDNA Spin Kit for Soil (MPBio, Santa Ana, CA). Amplicon libraries were prepared for the 16S rRNA gene (V1-V3 regions) and fungal internal transcribed spacer region (ITS2) according to the procedure of Hartmann *et al.*[5]. The region spanning V1–V3 was amplified using barcoded universal primers: 27F (5′-AGA GTT TGA TCM TGG CTC AG–3′) and 519R (5′-GWA TTA CCG CGG CKG CTG–3′)[11,12] and the fungal internal transcribed spacer (ITS2) region was amplified using barcoded primers: ITS3 (5′-GCA TCG ATG AAG AAC GCA GC–3′) and ITS4 (5′-TCC TCC GCT TAT TGA TAT GC–3′)[13]. Amplicons were generated via polymerase chain reaction (PCR) in triplicate for each sample and pooled prior to purification and quantification. DNA quantitation was performed using Pico-Green fluorescent dye (ThermoFisher, MA, USA). Samples were sequenced using the Roche 454 Titanium platform at the McGill University and at the Genome Québec Innovation Centre with a maximum of 40 samples multiplexed on each quarter plate. Table 2 (available online only) contains information on all amplicon libraries created from field soil samples, including 16S rRNA gene amplicon (Data Citation 1) and ITS amplicon libraries (Data Citation 2). This includes a second set of samples from Skulow Lake with a narrower focus on five soil depths in only REF and OM3 (Table 2 (available online only)). These amplicon libraries were made from primers targeting the V6-V8 region of the 16S rRNA gene and were amplified according to Gies *et al.*[14] using barcoded universal primers: 926F (5′-CC TAT CCC CTG TGT GCC TTG GCA GTC TCA GAA ACT YAA AKG AAT TGR CGG-3′) and

Site Name	Ecozone code	Site code	Region	Latitude	Longitude	Elevation (m)	Soil classification	Tree cover	Climatic zone	Annual mean temp (°C)	Precipitation—warmest quarter (mm)	Year established	Sample collection date	Country
Fensom	BS_{ON}	A7	Ontario	49.07	− 89.41	445	Orthic Dystric Brunisol	Black Spruce	Dfb, Humid Continental warm summer	2.4	266	1995	7/3/2011	Canada
Fensom	BS_{ON}	A8	Ontario	49.08	− 89.38	450	Orthic Dystric Brunisol	Black Spruce	Dfb, Humid Continental warm summer	1.8	266	1995	7/4/2011	Canada
Fensom	BS_{ON}	A9	Ontario	49.07	− 89.39	442	Gleyed Dystric Brunisol	Black Spruce	Dfb, Humid Continental warm summer	1.5	266	1995	7/5/2011	Canada
Brandy City	PP_{CA}	BR	California	39.55	− 121.04	1135	Mesic Ultic Haploxeralfs	Ponderosa pine, sugar pine, white fir, giant sequoia	Csa, Mediterranean hot summer	11.2	55	1995	6/22/2011	USA
Blodgett	PP_{CA}	BL	California	38.88	− 120.64	1350	Mesic Ultic Haploxeralfs	Ponderosa pine, sugar pine, white fir, g iant sequoia	Csa, Mediterranean hot summer	11.2	55	1995	9/16/2011	USA
Lowell Hill	PP_{CA}	LH	California	39.26	− 120.78	1268	Mesic Ultic Haploxeralfs	Ponderosa pine, sugar pine, white fir, giant sequoia	Csa, Mediterranean hot summer	11.2	55	1995	9/16/2011	USA
Wells	JP_{ON}	JW	Ontario	46.42	− 83.37	228	Orthic Humo-Ferric Podzol	Jack Pine, Black Spruce, Red Pine	Dfb, Humid Continental cool summer	4.4	248	1993–1994	7/7/2011	Canada
Superior	JP_{ON}	JS	Ontario	47.57	−82.85	426	Orthic Dystric Brunisol	Jack Pine, Black Spruce	Dfb, Humid Continental cool summer	1.7	250	1993–1994	8/4/2011	Canada
Eddy	JP_{ON}	JE	Ontario	46.75	−82.25	490	NA	Jack Pine, Balsam fir, White birch	Dfb, Humid Continental cool summer	2.8	242	1993–1994	8/3/2011	Canada
Kurth	LP_{TX}	TXA	Texas	31.11	−95.15	88	Aquic Glossudalfs	Loblolly Pine, Beautyberry, Yaupon, Sweetgum, Oaks, Wax Myrtle	Cfa, Humid subtropical	19.0	253	1997	3/12/2012	USA
Kurth	LP_{TX}	TXB	Texas	31.11	−95.15	88	Aquic Glossudalfs	Loblolly Pine, Beautyberry, Yaupon, Sweetgum, Oaks, Wax Myrtle	Cfa, Humid subtropical	19.0	253	1997	3/12/2012	USA
Kurth	LP_{TX}	TXC	Texas	31.11	−95.15	88	Aquic Glossudalfs	Loblolly Pine, Beautyberry, Yaupon, Sweetgum, Oaks, Wax Myrtle	Cfa, Humid subtropical	19.0	253	1997	3/12/2012	USA
O'Connor Lake	IDF_{BC}	OC	British Columbia	50.88	−120.35	1075	Brunisolic Gray Luvisol	Douglas fir	Dfb, Humid Continental warm summer	2.5	300	1999	6/26/2010	Canada
Black Pines	IDF_{BC}	BP	British Columbia	50.93	−120.28	1180	Brunisolic Gray Luvisol	Douglas fir, Lodgepole pine	Dfb, Humid Continental warm summer	2.5	300	1999	6/22/2010	Canada
Dairy Creek	IDF_{BC}	DC	British Columbia	50.85	−120.42	1150	Brunisolic Gray Luvisol	Douglas fir, Subalpine fir, Lodgepole pine	Dfb, Humid Continental warm summer	2.5	300	1999	6/25/2010	Canada
Log Lake	SBS_{BC}	LL	British Columbia	54.35	−122.61	780	Orthic Humo-Ferric Podzol, Gleyed Eluviated Dystric Brunisol	Subalpine fir, Douglas fir, Interior Spruce	Dfc, Boreal cool summer	3.3	146–193	1994	7/9/2008	Canada
Topley	SBS_{BC}	TO	British Columbia	52.32	−126.31	1100	Orthic Gray Luvisol, Gleyed Gray Luvisol	Lodgepole pine, Subalpine fir, Interior spruce	Dfc, Boreal cool summer	1.7	146–193	1994	7/11/2008	Canada
Skulow Lake	SBS_{BC}	SL	British Columbia	52.32	−121.92	1050	Orthic Gray Luvisol	Lodgepole pine, Interior spruce	Dfc, Boreal cool summer	3.8	146–193	1995	8/14/2009	Canada

Table 1. Site information for all sampling locations utilized in this study.

1392R (5′-CGT ATC GCC TCC CTC GCG CCA TCA GAC GGG CGG TGT GTR C-3′). PCR product was pooled, purified, quantified and sequenced as for other libraries.

Whole shotgun metagenomes were generated for a single site in each of the six ecozones resulting in three replicates for each treatment and each soil horizon for each ecozone. At Skulow Lake (SBS_{BC}),

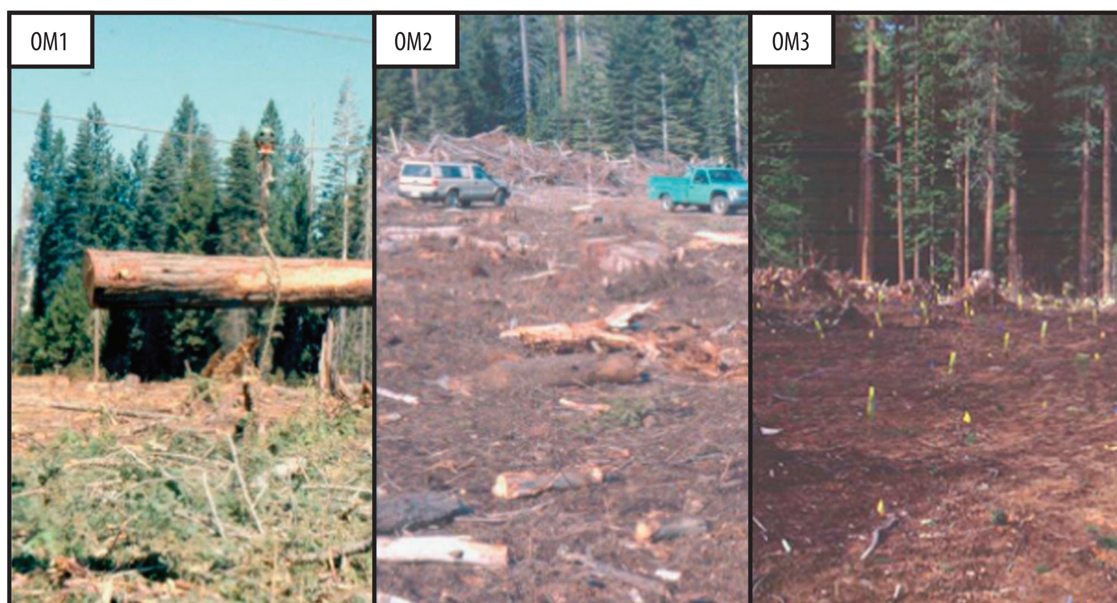

Figure 2. Plot conditions immediately following harvesting (year zero) in the PP_{CA} ecozone. These photographs capture the initial variation in the amount of organic matter (OM) removal at plots which were sampled 11–17 years for this study. Plots like these were replicated at three sites within every ecozone [photo credit: Dr Matt Busse; mbusse@fs.fed.us.

sampling for metagenomes focused on changes along the soil profile and, thus, did not cover all four OM harvesting treatments, only REF versus OM3. Unlike in pyrotag libraries, a second complete set of shotgun metagenomes from Skulow Lake does not exist. Insufficient organic layer soil was available from OM3 to prepare shotgun metagenomes for BS_{ON}, IDF_{BC}, JP_{ON}, and PP_{CA}. The same soil samples were used for shotgun and amplicon metagenomes, but from separate DNA extractions. After quantification, triplicate samples were multiplexed in each Illumina HiSeq lane for sequencing. Samples from ecozones BS_{ON}, JP_{ON}, PP_{CA} and LP_{TX} were sequenced at the US DOE Joint Genome Institute (Walnut Creek, CA) producing paired-end, 150-bp Illumina libraries while samples for the IDF_{BC} and SBS_{BC} ecozones were sequenced at the Michael Smith Genome Sciences Centre (Vancouver, Canada) resulting in paired-end, 75-bp and 100-bp Illumina libraries, respectively.

Stable isotope probing amplicon and shotgun metagenome sequencing

Microcosms were prepared by adding between 0.75 and 2.0 g of 2 mm sieved organic or mineral layer soil to 30-mL serum vials. Larger quantities of mineral soil were necessary because of lower microbial biomass, requiring two DNA extractions per mineral soil to obtain the necessary 5 μg of DNA. Moisture content was standardized to 60% w/v for mineral soil, due to lower absorptive capacity, and 125% w/v for organic soil and pre-incubating at 20 °C for one week. Microcosms were then amended with either 1.0% w/w of ^{13}C-labeled hemicellulose (97 atom %; IsoLife; U-10509, Lot: 0901-0273) or custom prepared bacterial cellulose (99 atom % ^{13}C). Each soil sample was paired with a microcosm containing the same amount of unlabelled (~1.1 atom % ^{13}C) substrate. Separation between ^{13}C-enriched DNA and unlabelled DNA is never complete, in part, due to variation in GC content[15]. Thus, ^{13}C-libraries are always compared to identically prepared sequencing libraries from samples amended with unlabelled substrate. Following either 2-day (hemicellulose), 11-day (cellulose—organic layer) or 14-day (cellulose—mineral layer) incubations, soil was lyophilized and stored at −80 °C until DNA was extracted as previously described. DNA extracts from replicates within each site were pooled in equal amounts and unlabelled controls were processed identically. Harvested treatment OM2 was not included in SIP-hemicellulose libraries.

High purity ^{13}C-labled cellulose was necessary to ensure that only organisms possessing the necessary metabolic capability could assimilate the ^{13}C. Bacterial cellulose was utilized due to irremediable impurity in plant-derived cellulose from IsoLife (58% glucose+4.4% lignin+remainder of sugars from hemicellulose). Bacterial cellulose was produced by cultivating *Gluconacetobacter xylinus* str. KCCM 10,100 with either ^{13}C-labeled glucose (99 atom % ^{13}C, Cambridge Isotope Laboratories, MA, USA), or unlabelled glucose, as sole carbon source in Yamanaka medium[16]. Cellulose pellicules were purified by boiling in sodium hydroxide per previously described methods[17], with the addition of a second boiling step and an increase in boiling time to 4 h. While IsoLife hemicellulose was also impure (53%

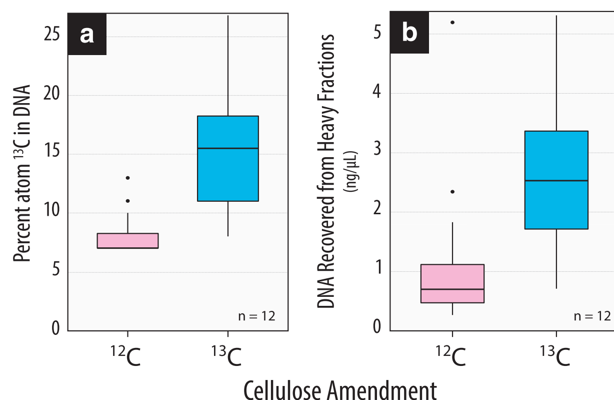

Figure 3. The successful enrichment and recovery of ^{13}C-enriched DNA in SIP experiments. The assimilation of ^{13}C by functional guilds during stable isotope probing experiments was evident in (**a**) the total ^{13}C-enrichment of soil DNA extract and (**b**) recovery of DNA from the densest CsCl gradient fractions (F_1-F_7). These differences were evident in comparing soils amended with either ^{12}C- (unlabelled) or ^{13}C-cellulose. These trends were apparent in both cellulose and hemicellulose SIP experiments. Boxplots depict the average (centre line) and spread (from 25th to 75th percentile) of data, while the whiskers extend to the extrema. A total of twelve samples were averaged for each factor.

hemicellulose sugars), the sources of impurity were more recalcitrant forms of carbon, such as cellulose and lignin, which were not substantially degraded during the 2-day incubation.

^{13}C-enriched DNA was separated and recovered using cesium chloride density gradient ultracentrifugation[10]. The atom % ^{13}C was measured before and after density separation using UHPLC-MS/MS[18]. Amplicon libraries were prepared and sequenced from ^{13}C-enriched DNA for SIP-hemicellulose (Data Citations 4,5) and SIP-cellulose (Data Citation 6) experiments targeting both bacterial and fungal phylogenetic markers as previously described (Table 4 (available online only)). Four shotgun metagenome libraries were generated from ^{13}C-enriched DNA from SIP-cellulose incubations for REF, OM1 and OM3 treatments and one unlabelled control sample (REF) from PP_{CA}. Sufficient DNA for metagenome library preparation was achieved by pooling the corresponding DNA extracts from mineral layer samples at all three sites in PP_{CA}. Shotgun metagenomes were prepared from 40–50 ng of enriched DNA using the Nextera DNA Sample Preparation Kit (Illumina Inc., CA, USA) and were multiplexed on two lanes of Illumina HiSeq (2 × 100-bp), yielding 285 million paired-end reads (Data Citation 6). There was insufficient ^{13}C-enriched DNA to generate metagenomes from the organic layer samples. Raw sequences were quality-controlled, assembled and binned into partial genomes (Table 4 (available online only); Data Citation 6) according to methods described in Wilhelm *et al.*[8].

Data Records

The raw pyrosequencing output (~400-bp reads) for all 16S rRNA gene ($n = 724$ samples) and ITS ($n = 658$ samples) amplicon libraries that were generated from field soil samples averaged 8,900 and 8,400 reads per library (post-QC), respectively, and are archived at the European Sequencing Archive (Table 2 (available online only)). The raw SFF files for 16S rRNA amplicon libraries can be found with the study accession PRJEB8599 (Data Citation 1), except for all libraries from British Columbia which were archived in study accession PRJEB12501 in ‘fastq’ format (Data Citation 2). The latter contains the entire collection of ITS amplicon libraries in ‘fastq’ format. All libraries were extracted from raw SFF files using the mothur command ‘sffinfo’ and either uploaded in standard flowgram format (SFF) or converted to ‘fastq’ format[19] using ‘sffinfo’ to produce paired ‘fasta’ and ‘qual’ files, which were merged into ‘fastq’ format using ‘PairedFastaQualIterator’ from the SeqIO module in BioPython[20].

Raw shotgun metagenomic data for all ecozones can be downloaded in ‘fastq’ format with the study accession PRJEB8420 from the European Nucleotide Archive (Table 3 (available online only), Data Citation 3). After our quality filtering, libraries had a median count of 59.3 million sequences for 150-bp read libraries, while shorter read libraries (75-bp) had higher median counts (115 million). These numbers are provided as an estimate of the number of high quality reads obtainable from our raw data, but will change depending on the parameters selected during quality processing.

Due to cost of ^{13}C-labeled materials and additional labour required to process SIP samples, only a subset of ecozones were selected for SIP-hemicellulose (PP_{CA} & IDF_{BC}) and SIP-Cellulose (PP_{CA}) characterizations. The raw pyrosequencing output (~400-bp reads) for SIP-hemicellulose 16S rRNA (Data Citation 4) and ITS (Data Citation 5) amplicon libraries, which averaged 4,200 and 3,800 reads per library (post-QC), respectively, are available in SFF format from the European Nucleotide Archive (Table 4 (available online only)). These datasets include both ^{13}C- and ^{12}C-libraries for 16S rRNA genes, $n = 35$ and 15, respectively, and for ITS, 18 and 2, respectively. Similarly, the raw sequencing data for all

SIP-cellulose 16S rRNA amplicon and ITS amplicon libraries, averaging 11,800 and 10,300 reads per library, respectively, are available in 'fastq' format from the European Nucleotide Archive (Table 4 (available online only)). The ^{13}C- and ^{12}C-libraries from these data are more balanced, with a total of 24 16S rRNA gene libraries for both, and 24 and 23, respectively, for ITS libraries. Raw Illumina, paired-end, 100-bp shotgun metagenome sequencing libraries for pooled, SIP-cellulose mineral soil incubations were archived in 'fastq' format (Data Citation 6; Table 4 (available online only)), along with 10 partial genomes in 'fasta' format comprised of assembled scaffolds (Data Citation 6; Sample accessions: ERZ288956—ERZ288966).

Technical Validation

The recovery of ^{13}C-enriched DNA was validated by quantifying ^{13}C-content of nucleic acids[18] (Fig. 3). The completeness of draft genomes recovered from SIP-Cellulose metagenomes was assessed by scanning for essential single-copy, house-keeping genes with hidden Markov models[21].

Usage Notes

Metagenomes for all ecozones, except IDF_{BC}, can be also found at the DOE JGI portal (http://genome.jgi.doe.gov/) under proposal ID 543. This site provides the raw sequencing data and annotation for the assembled and unassembled metagenomes.

References

1. Powers, R. F. *et al.* The North American long-term soil productivity experiment: Findings from the first decade of research. *Forest Ecol. Manag.* **220,** 31–50 (2005).
2. Thiffault, E. *et al.* Effects of forest biomass harvesting on soil productivity in boreal and temperate forests—A review. *Environ. Rev.* **19,** 278–309 (2011).
3. Ponder, F. *et al.* Effects of organic matter removal, soil compaction and vegetation control on 10th year biomass and foliar nutrition: LTSP continent-wide comparisons. *Forest. Ecol. Manag.* **278,** 35–54 (2012).
4. Holub, S. M., Terry, T. A., Harrington, C. A., Harrison, R. B. & Meade, R. Tree growth ten years after residual biomass removal, soil compaction, tillage, and competing vegetation control in a highly-productive Douglas-fir plantation. *Forest Ecol. Manag.* **305,** 60–66 (2013).
5. Cardenas, E. *et al.* Forest harvesting reduces the soil metagenomic potential for biomass decomposition. *ISME J.* **9,** 1–12 (2015).
6. Hartmann, M. *et al.* Significant and persistent impact of timber harvesting on soil microbial communities in Northern coniferous forests. *ISME J.* **6,** 2199–2218 (2012).
7. Leung, H. T., Maas, K. R., Wilhelm, R. C. & Mohn, W. W. Long-term effects of timber harvesting on hemicellulolytic microbial populations in coniferous forest soils. *ISME J.* **10,** 363–375 (2016).
8. Wilhelm, R. *et al.* Long-term Enrichment of Stress-tolerant Cellulolytic Soil Populations following Timber Harvesting Revealed by Multi-omic Stable Isotope Probing. *Front. Microbiol* **8,** 537 (2017).
9. Wilhelm, R. *et al.* Biogeography and Organic Matter Retention Shape Effects of Timber Harvesting on Forest Soil Microbial Communities in Long-term Soil Productivity Study. *ISME J.* doi:10.1038/ismej.2017.109 (2017).
10. Neufeld, J. D. *et al.* DNA stable-isotope probing. *Nat. Protoc.* **2,** 860–866 (2007).
11. Lane, D. J. in *Nucleic Acid Techniques in Bacterial Systematics* (eds Stackebrandt, E. & Goodfellow, M.) 115–175 (John Wiley and Sons, 1991).
12. Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol. Rev.* **59,** 143–169 (1995).
13. White, T. J., Bruns, T., Lee, S., & Taylor, J. in *PCR Protocols: a Guide to Methods and Applications* (eds Innis, M. A., Gelfand, D. H., Sninsko, J. J. & White, T. J.) 315–322 (Academic Press, 1990).
14. Gies, E. A., Konwar, K. M., Beatty, T. & Hallam, S. J. Illuminating Microbial Dark Matter in Meromictic Sakinaw Lake. *Appl. Env. Microbiol* **80,** 6807–6818 (2014).
15. Youngblut, N. D. & Buckley, D. H. Intra-genomic variation in G+C content and its implications for DNA stable isotope probing. *Environ. Microbiol. Rep* **6,** 767–775 (2014).
16. Ruka, D. R., Simon, G. P. & Dean, K. M. Altering the growth conditions of *Gluconacetobacter xylinus* to maximize the yield of bacterial cellulose. *Carbohyd. Polym.* **89,** 613–622 (2012).
17. Pinnell, L. J., Dunford, E. A., Ronan, P., Hausner, M. & Neufeld, J. Recovering glycoside hydrolase genes from active tundra cellulolytic bacteria. *Can. J. Microbiol.* **60,** 469–476 (2014).
18. Wilhelm, R., Szeitz, A., Klassen, T. L. & Mohn, W. W. Sensitive, Efficient Quantitation of ^{13}C-Enriched Nucleic Acids via Ultrahigh-Performance Liquid Chromatography–Tandem Mass Spectrometry for Applications in Stable Isotope Probing. *Appl. Environ. Microbiol.* **80,** 7206–7211 (2014).
19. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
20. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
21. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–538 (2013).

Data Citations

1. *European Nucleotide Archive* PRJEB8599 (2015).
2. *European Nucleotide Archive* PRJEB12501 (2016).
3. *European Nucleotide Archive* PRJEB8420 (2015).
4. *European Nucleotide Archive* PRJEB9181 (2015).
5. *European Nucleotide Archive* PRJEB9182 (2015).
6. *European Nucleotide Archive* PRJEB9761 (2016).

Acknowledgements

This study was supported by grants from Genome Canada and Genome BC as well as an NSERC Strategic Project Grant. Sampling and data collection were contributed by collaborators at the Canadian

Forestry Services (Drs Paul Hazlett, Kara Webster, Robert L. Fleming), the Ontario Ministry of Natural Resources (Dr David Morris), BC Ministry of Forests, Lands and Natural Resources Operations (Drs Shannon Berch, Chuck Bulmer, Bill Chapman, Stephane Dubé, Graeme Hope and Marty Kranabetter) and the U.S. Forest Service (Drs Matt Busse and Andy Scott). Drs E.C., M.H. and K.M. were supported by postdoctoral fellowships from the Tula Foundation and both R.C.W. and A.H. by NSERC graduate scholarships. Sequencing conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We also acknowledge the contribution of scientists at the McGill University and Génome Québec Innovation Center, Montréal, Canada.

Author Contributions

R.C.W. contributed to field sampling, processing of soil samples and DNA extractions, all data corresponding to the SIP-cellulose experiment, data curation and data submission, and was the principle author of this manuscript. E.C. contributed shotgun metagenomes, data curation, and data submission. H.L. contributed to field sampling, processing of soil samples and DNA extraction and produced all data corresponding to SIP-hemicellulose experiment. K.M. performed PCR for amplicon libraries from PP_{CA}, LL_{TX}, JP_{ON} and BS_{ON} field soils and helped to consolidate metadata. A.H. contributed amplicon and shotgun metagenomes at various soil depths from Skulow Lake. M.H. contributed to field sampling and all amplicon data from SBS_{BC} and IDF_{BC}. S.H. designed and initiated the sampling program with W.W. M. and supervised sample processing and sequencing for the project. W.W.M. conceived and supervised the project, contributed to sampling and edited the manuscript.

Additional Information

Tables 2–4 are only available in the online version of this paper.

Competing interests: The authors declare no competing financial interests.

How to cite this article: Wilhelm, R. C. *et al.* A metagenomic survey of forest soil microbial communities more than a decade after timber harvesting. *Sci. Data* 4:170092 doi: 10.1038/sdata.2017.92 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files made available in this article.

© The Author(s) 2017