Supplementary material

Title: The tale of a neglected energy source: Elevated hydrogen exposure affects both microbial diversity and function in soil

Authors: Mondher Khdhiri, Sarah Piché-Choquette, Julien Tremblay, Susannah G. Tringe and Philippe Constant

Table of content

Method S1. Description of the hypothesized paths constituting the models to predict direct and indirect effects of H_2 exposure on microbial activities 3

Method S2. Dynamic microcosm chamber unit 8

Method S3. Metagenomic analysis pipeline 9

Table S1. Absence of CH_4 production in soil microcosms. 11

Table S2. List of genome bins for which the distribution was influenced by H_2 exposure. 12

Table S3. Multiple linear regression equations defining the relative abundance of (A) genome bins outlined in the PCA and the corresponding (B) OTUs (rRNA marker gene amplicon sequencing) according to biotic and abiotic parameters. 13

Table S4. Genome bins of potential HOB containing hydrogenase gene based on search in metagenomic annotation databases (1 gene) and [NiFe]-hydrogenase HMM (7 genes). 14

Figure S1. Community-level carbon substrates utilization profiling. 15

Figure S2. Overview of the taxonomic composition of bacterial and fungal communities in soils. 16

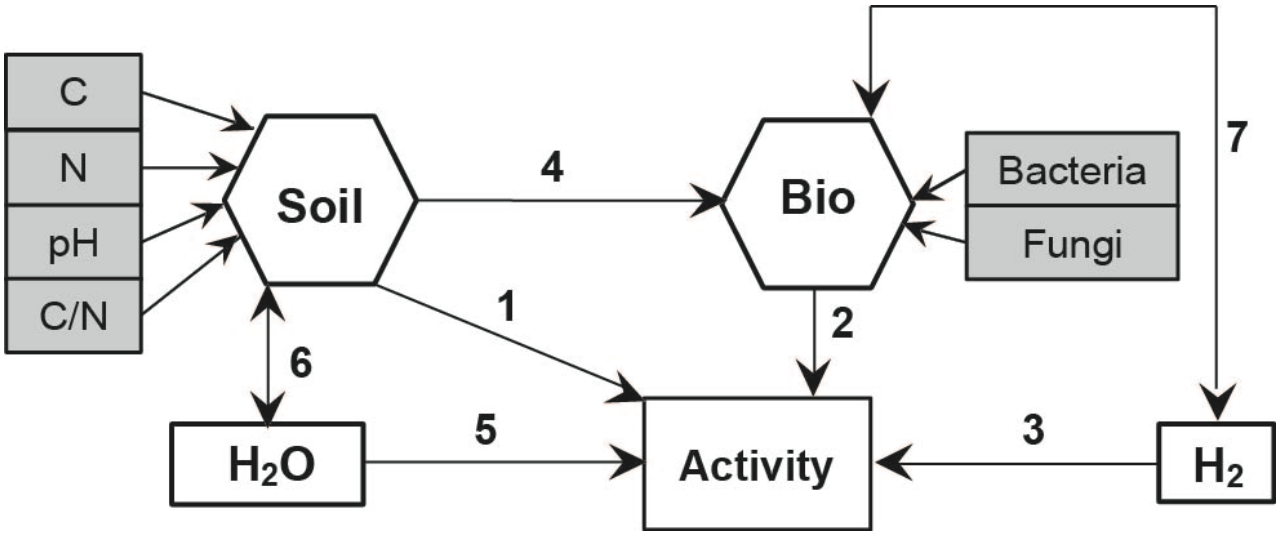
Figure S3. UPGMA agglomerative clustering of soil microcosms according to a Euclidean distance matrix calculated with a Hellinger distance matrix of annotated gene abundance profiles 17

Figure S4. Bubble-chart representation of the OTUs associated with the 4 genome bins of interest 18

References 19

Method S1. Description of hypothesized paths constituting the models predicting direct and indirect effects of H_2 exposure on measured microbial processes

A structural equation model (SEM) was computed based on the extensive procedure described by Grace (2006) to test the hypothesis that idiosyncratic impact of H_2 exposure on soil microbial community structure and function is explained by soil biotic and abiotic features acting as ecological filters for microbial species and functional groups. The overall model was comprised of seven paths involving two composites variables (SOIL and BIO) and two variables (H_2O and H_2). Composite variables were computed by linear combination of observed indicators, of which C, N, C\N and pH were combined as a composite variable, named SOIL, and species richness (*i.e.* Shannon index) of bacterial and fungal communities as the composite variable named BIO:



The hypothetical model was based on a number of assumptions supported by literature to predict trace gas turnover and carbon utilization profiles. These assumptions are described in the following table:

Path	Description	Potential mechanism
1	Soil physicochemical effects on microbial activity.	The metabolic activity of microorganisms is influenced by the abiotic conditions of their environment.
2	Microbial diversity effects on microbial activity.	According to the biodiversity-ecosystem functioning theory, microbial processes are favored by the presence of a high diversity level (herein expressed as species richness).
3	Direct and indirect effects of H_2 exposure on microorganisms catalyzing a specific process.	The results of this study have demonstrated the impact of H_2 on microbial metabolism. Direct effects on H_2 oxidation rate and indirect effects on CH_4 oxidation rate and carbon mineralization potential were observed. The underlying mechanisms are unknown, but should involve alteration of microbial community structure (abundance and interactions) induced by HOB favored by H_2 exposure.
4	Impact of soil physicochemical properties on microbial diversity.	Abiotic conditions in soil are known drivers of microbial community structure.
5	Soil water content effect on microbial activity.	Soil water content influences microbial community structure and function. Under water stress conditions, drought-tolerant microorganisms are favored. Gas diffusion is also a limiting factor for H_2 and CH_4 oxidation rates measured in soil. Our inability to maintain the exact same water content in soils (*i.e.* lower water content in farmland soil under eH_2 treatment when compared to aH_2 treatment) justified the addition of this variable in the model (see Table 1).
6	Covariation between soil physicochemical properties and soil water content.	Soil water holding capacity measured in the soil samples representative of the three land-use types defined the amount of water to be added to soil microcosms.
7	Covariation between microbial diversity and H_2 treatment	Different biodiversity levels were expected to occur at different H_2 exposure levels.

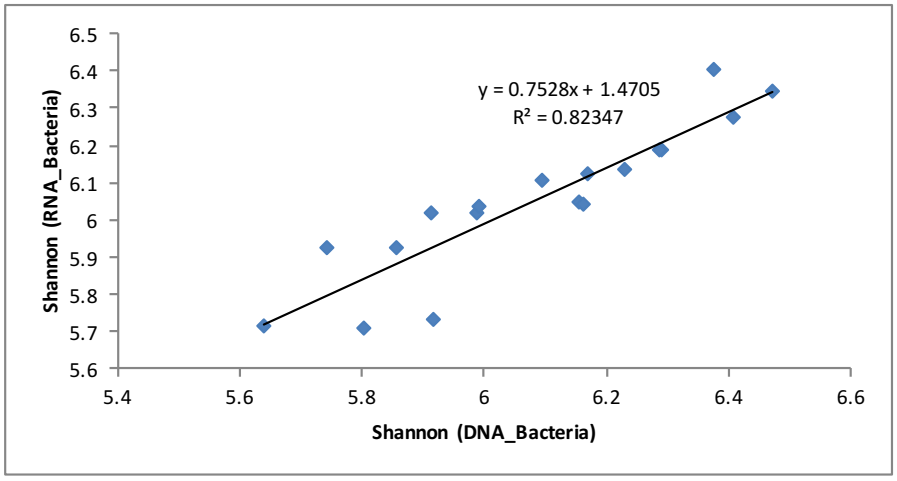
Composites variables were estimated using multiple linear regression. The weighting for the relationship between measured and composite variables for the four modeled microbial processes (high

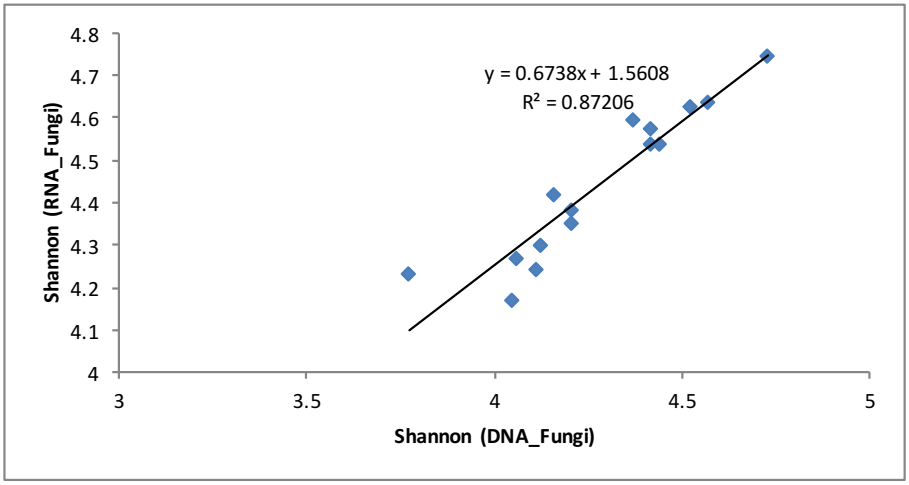
affinity H_2 oxidation rate, low affinity H_2 oxidation rate, CH_4 oxidation rate and Ecoplates profile) are shown in the table below. Values reported in the table were used to calculate composite variables estimates that were then included in SEM calculation:

Measured variables	**Composite variables**	
	SOIL	**BIO**
High affinity H_2 oxidation		
C	-216	
N	-165	
pH	3797	
C/N	1171	
Bacteria species richness (Shannon)		-511
Fungi species richness (Shannon)		-694
Low affinity H_2 oxidation		
C	129	
N	-855	
pH	-423	
C/N	-154	
Bacteria species richness (Shannon)		210
Fungi species richness (Shannon)		-40
CH_4 oxidation		
C	-10	
N	81	
pH	406	
C/N	78	
Bacteria species richness (Shannon)		-38
Fungi species richness (Shannon)		53
Ecoplate		

C	0.237	
N	-2.295	
pH	-0.381	
C/N	-0.215	
Bacteria species richness (Shannon)		-0.013
Fungi species richness (Shannon)		0.059

Due to the lack of two bacterial and four fungal ribotying profiles, missing Shannon indices were derived from RNA-based ribotyping profiles performed on soil samples collected from the microcosms exposed to eH_2 and aH_2 treatments (RNA-based analyses were obtained from the same incubations reported in the current study, but are unpublished as of yet). Regression analyses between Shannon indices obtained from DNA-based ribotyping profile (independent variable) and Shannon indices obtained from RNA-based ribotyping profile (dependent variable) are shown in the following figures:



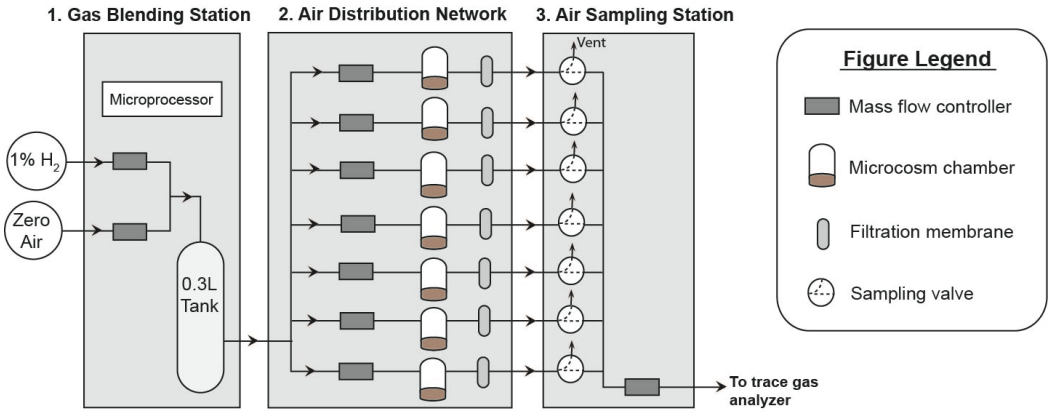


Derived equations were used to estimate the five Shannon index values presented in the following table (ultimately, Shannon index value for microcosm LP(c) (fungi) is still missing since sequencing of both DNA and RNA libraries were unsuccessful):

Sample	**Shannon index obtained from RNA-based ribotyping profile**	**Estimated Shannon index used in SEM**
HP(b) - Bacteria	6.34	6.47
LP(c) - Bacteria	5.71	5.64
HA(b) - Fungi	4.75	4.73
LP(a) - Fungi	4.64	4.57
LP(b) - Fungi	4.54	4.42

Method S2. Dynamic microcosm chamber unit

The term "dynamic" holds for the continuous air stream circulating in and out of the microcosms, resulting in a user-specified hydraulic residence time and chemical composition of the headspace. The system is comprised of three units: a gas blending station, a gas distribution network and an air sampling station. The gas blending station has been designed to control the dilution of a commercial gas mixture, comprising of 1% H_2 in synthetic air (GST-Welco, Pennsylvania, U.S.A), with gas supplied from a Balston zero air generator (Parker Hannifin Corp., QC, Canada). It comprises of two Brooks® Delta II Smart mass flow controllers (0-700 Sccm, accuracy of ±1.0%). The air sampling networks consists of seven Brooks® Delta II Smart mass flow controllers (0-100 Sccm, accuracy of ±1.0%) delivering gas mixture to independent soil microcosms. Microcosms are also vented to the atmosphere through the air sampling station. This station has an option allowing to measure, in real-time, H_2 and CO levels when the Trace Analytical ta3000R (Reduction Gas Detector) is connected in series (that option was not used in the present study). Operation of the Dynamic microcosm chamber unit is under the control of a programmable logic controller (Allen-Bradley® PanelView Plus 600) comprising of a digital display of system status and alarms for defaults in flowrate, gas supply and power.



Method S3. Metagenomic analysis pipeline

Metagenomic libraries were prepared and sequenced on an Illumina HiSeq2000 system on a 2x150 configuration. Sequencing raw data (256 GB for metagenome) were processed through our metagenomics bioinformatics pipeline. Read count summaries are provided for metagenome sequencing libraries (Data set 5A). Sequencing adapters were removed from each read (Trimmomatic v0.32) [Bolger et al., 2014] to generate quality controlled (QC) reads. QC-passed reads from each sample were assembled into a large metagenome assembly using Ray software v2.3.1 [Boisvert et al., 2012] with a kmer size of 31 (Data set 5B). Gene prediction of obtained contigs was performed by calling genes on each assembled contig using MetageneMark v1.0 (Tang et al., 2013). Genes were annotated following the JGI's guidelines (Huntemann et al., 2016): 1) RPSBLAST (v2.2.29+) (Camacho et al., 2009) against COG database (e.g. CDD v3.11); 2) RPSBLAST (v2.2.29+) against KOG database (e.g. CDD v3.11). The best hit having at an e-value ≥ 1e-02 was kept for each query; 3) HMMSCAN (v3.1b1) [Eddy, 2011] against PFAM-A (v27.0) database (Finn et al., 2013) best hit having at least an e-value ≥ 1e-02 was kept for each query; 4) TIGRFAM database (v15.0) best hit having at least an e-value ≥ 1e-02 was kept for each queries; 5) BLASTP (v2.2.29+) against KEGG database v71.0, and 6) BLASTN (v2.2.29+) against NCBI`s nucleotide (nt) database (version of May 16th 2013). Contigs (and not genes) sequences were blasted against NCBI's nt database as well for taxonomic assignment. For each of these databases comparisons, the best hit having at least an e-value ≥ 1e-02 was kept for each query. QC-passed reads were mapped (BWA mem v0.7.10) (unpublished - http://bio-bwa.sourceforge.net) against contigs to assess quality of metagenome assembly and to obtain contigs abundance profiles. Alignment files in bam format were sorted by read coordinates using samtools v1.1 and only properly aligned read pairs were kept for downstream steps. Each bam file (containing properly aligned paired-reads only) was analyzed for coverage of called genes and contigs using bedtools (v2.17.0) [Quinlan & Hall, 2010] using a custom bed file representing gene coordinates

of each contig. Only paired-reads both overlapping their contigs or genes were considered for gene counts. Coverage profiles of each sample were merged together to generate an abundance matrix (rows = contig, columns = samples) for which a corresponding CPM (Counts Per Million) abundance matrix (edgeR v3.10.2) [Robinson et al., 2010] was generated as well.

Taxonomy of each contig was assigned using the NCBI taxonomy database [Benson et al., 2009; Sayers et al., 2009] (ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz) (as downloaded on June 5th 2015). Each GIs resulting from BLASTN against nt were used to retrieve full taxonomic lineages (when available) from the NCBI taxonomy database. Taxonomic lineages were integrated to the contig abundance of read counts matrix to generate an OTU table format file (with contigs replacing OTUs as rows). Taxonomic summaries were performed using a combination of in-house Perl and R scripts and Qiime v.1.9.0 [Caporaso et al., 2010]. Genome bins abundance tables along with their taxonomic lineages are included in Data set 5C.

Binning was done using Metabat (v0.26.1) [Kang et al., 2015] using an abundance matrix generated using the jgi_summarize_bam_contig_depths software (Kang et al., 2015) with --minContigLength 1000 --minContigDepth 2 and --minContigIdentity 95 parameters. Genome bins obtained from Metabat were further processed/decontaminated by splitting each bin into three sub-bins based on the assigned taxonomic lineage at the Order level, as each bin typically had a significant amount of contigs associated with the same Order taxon. For instance, in our dataset, the bin labeled "1" had 607 contigs assigned to the *Rhizobiales* Order level, 136 contigs to the *Burkholderiales* and 189 more contigs to an undefined taxonomy value. Consequently, 3 sub-bins were generated and labeled Bin-1 (*Rhizobiales*), Bin-1 (*Burkholderiales*) and Bin-1 (NULL) respectively. All genome bins were characterized using CheckM v1.0.4 (Parks et al., 2015) and detailed in Data set 5D.

Table S1. Absence of CH_4 production in soil microcosms. The absence of methane production in soil microcosms was tested in poplar and farmland soils (soils where difference in CH_4 oxidation rates were most prominent). For that purpose, microcosms were flushed for 15 minutes with a synthetic gas mixture (Zero Air, Praxair distribution Inc. PA, USA) containing non detectable CH_4 (detection limit "DL" of the Agilent GC system below 0.2 ppmv CH_4). CH_4 level was then monitored for more than 10 hours (compared to 8 hours for CH_4 oxidation measurements reported in the article). As no CH_4 was detected, we also monitored CO_2 as an analytical control of the flame ionization detector (FID) of the gas chromatograph. Average values of three independent replicates are shown with standard deviations.

	aH_2 microcosms				**eH_2 microcosms**				
Concentrations (ppmv)	**CO_2**		**CH_4**			**CO_2**		**CH_4**	
Time (min)	**Poplar**	**Farmland**	**Poplar**	**Farmland**	**Time (min)**	**Poplar**	**Farmland**	**Poplar**	**Farmland**
0	18 (10)	18 (4)	< DL	< DL	**0**	4 (1)	4 (1)	< DL	< DL
66	238 (38)	239 (74)	< DL	< DL	**55**	49 (4)	40 (2)	< DL	< DL
132	464 (71)	421 (130)	< DL	< DL	**110**	96 (5)	75 (3)	< DL	< DL
802	2416 (361)	1842 (510)	< DL	< DL	**630**	504 (28)	398 (21)	< DL	< DL

Table S2. List of genome bins for which the distribution was influenced by H_2 exposure. Log(FC) = Log-fold-change, number of times this OTU is more/less abundant in H_2 treatment. Log(CPM) = Log-fold-change in counts per million. LR = Likelihood ratio or how many times more likely the OTU is more/less abundant in a treatment than the other. P-value = significance level (only p-values < 0.05 were kept). = most abundant in 10,000 ppmv H_2 treatment. = most abundant in 0.5 ppmv H_2 treatment.

FARMLAND	Log(FC)	Log(CPM)	LR	P-value	Response to H_2
Bin-30 (*Burkholderiales*)	0.72	12	5	0.03	
Bin-6 (NULL)	0.94	13	8	0.004	
Bin-3 (*Xanthomonadales*)	1.1	16	11	0.0007	
Bin-6 (*Xanthomonadales*)	1.2	16	13	0.0003	
Bin-2 (*Burkholderiales*)	2.0	12	32	1.84E-08	
Bin-2 (*Xanthomonadales*)	2.1	17	41	1.24E-10	
LARCH	Log(FC)	Log(CPM)	LR	P-value	Response to H_2
Bin-21 (NULL)	-3.9	11	17	4.19E-05	
POPLAR	Log(FC)	Log(CPM)	LR	P-value	Response to H_2
Bin-3 (*Sphingomonadales*)	-1.7	12	4	0.048	
Bin-2 (*Xanthomonadales*)	-0.5	15	5	0.03	
Bin-7 (*Xanthomonadales*)	-0.7	16	10	0.002	
Bin-6 (*Burkholderiales*)	-1.4	13	22	2.10E-06	
Bin-3 (NULL)	-1.4	14	24	1.18E-06	
Bin-6 (*Xanthomonadales*)	-1.6	16	41	1.71E-10	
Bin-3 (*Xanthomonadales*)	-1.6	16	44	3.11E-11	

Table S3. Multiple linear regression equations defining the relative abundance of (A) genome bins outlined in the PCA and the corresponding (B) OTUs (rRNA marker gene amplicon sequencing) according to biotic and abiotic parameters. The interaction between the variables "H_2 treatment" and "net CO_2 production" was considered due to the ability of HOB to fix CO_2, resulting in a significant decrease in net CO_2 production rates in eH_2 treatments (CO_2 alone displayed non-significant explanatory power).

A.

Equations	R^2 (p-value)	Residual standard error (ε)
Bin-1 (Rhizobiales) = -13×10^3 ($\pm1.7\times10^3$) pH + 77×10^3 ($\pm8.5\times10^3$) + ε	0.78 (0.0001)	1379
Bin-2 (Xanthomonadales) = 2.7×10^4 ($\pm0.65\times10^4$) pH – 2.1×10^{-2} ($\pm5.9\times10^{-3}$) CO_2 x Treatment -1.3×10^5 ($\pm3.2\times10^4$) + ε	0.78 (0.0001)	5091
Bin-3 (Xanthomonadales) = -1.6×10^3 ($\pm2.7\times10^2$) C/N – 1.1×10^{-2} ($\pm1.1\times10^{-3}$) CO_2 x Treatment + 1.9×10^4 ($\pm3.1\times10^3$) + ε	0.95 (0.0001)	977
Bin-6 (Xanthomonadales) = 1.6×10^3 ($\pm3.0\times10^2$) C/N – 9.7×10^{-3} ($\pm1.3\times10^{-3}$) CO_2 x Treatment + 1.9×10^4 ($\pm3.4\times10^3$) + ε	0.92 (0.0001)	1072

B.

Equations	R^2 (p-value)	Residual standard error (ε)
Bin-1 (OTU *Bradyrhizobium*) = -0.041 (±0.003) pH + 0.245 (0.016) + ε	0.93 (2.8×10^{-9})	2.5×10^{-3}
Bin-1 (OTU *Rhodoplanes*) = -0.033 (±0.003) pH + 0.190 (0.013) + ε	0.92 (4.6×10^{-9})	2.1×10^{-3}
Bin-2 (OTU *Luteimonas*) = 7.62×10^{-2} ($\pm1.6\times10^{-2}$) pH – 8.7×10^{-8} ($\pm1.6\times10^{-8}$) CO_2 x Treatment – 3.9×10^{-1} ($\pm7.8\times10^{-2}$) + ε	0.85 (1.6×10^{-4})	1.2×10^{-2}
Bin-3 (OTU *Lysobacter*) = -4.0×10^{-3} ($\pm1.2\times10^{-3}$) C/N – 5.3×10^{-8} ($\pm5.7\times10^{-9}$) CO_2 x Treatment + 2.1×10^{-2} ($\pm1.3\times10^{-2}$) + ε	0.98 (6.6×10^{-9})	4.1×10^{-3}
Bin-6 (OTU Xanthomonadaceae) = -5.3×10^{-3} ($\pm1.5\times10^{-3}$) C/N – 3.7×10^{-8} ($\pm7.2\times10^{-9}$) CO_2 x Treatment + 5.0×10^{-2} ($\pm1.7\times10^{-2}$) + ε	0.89 (3.9×10^{-5})	5.3×10^{-3}

Table S4. Genome bins of potential HOB containing hydrogenase gene based on search in metagenomic annotation databases (1 gene) and [NiFe]-hydrogenase HMM (2 genes). The table shows the relative abundance of genome bins (mean and standard deviation) for triplicate microcosms exposed to aH_2 and eH_2 treatments. The asterisks (*) represent genome bins displaying a significant difference ($\alpha < 0.05$) between the two H_2 treatments (Wilcoxon-Mann-Whitney test).

Genome bins	Soil microcosms					
	Poplar monoculture		Agricultural		Larch monoculture	
	eH_2-P	aH_2-P	eH_2-A	aH_2-A	eH_2-L	aH_2-L
Bin-1 (Rhizobiales)[A]	11371 (665)	11773 (871)	10824 (1061)	12593 (1032)	17021 (1920)	16856 (1901) *
Bin-4 (NULL)	1102 (28) *	1273 (110) *	1205 (92)	1316 (102)	906 (35)	865 (27)
Bin-25 (NULL)	3770 (75) *	3576 (129) *	3014 (31) *	3175 (40) *	3722 (68) *	4011 (105) *

[A]A gene fragment encoding for the small subunit of a group 1 [NiFe]-hydrogenase was found in metagenomic annotation database.

Figure S1. Community-level carbon substrates utilization profiling. This heatmap reflects the utilization of 31 carbon sources divided into 6 subcategories. The upper part of the graph consists of an UPGMA agglomerative clustering of the profiles, showing a clear dichotomy between both H_2 treatments.

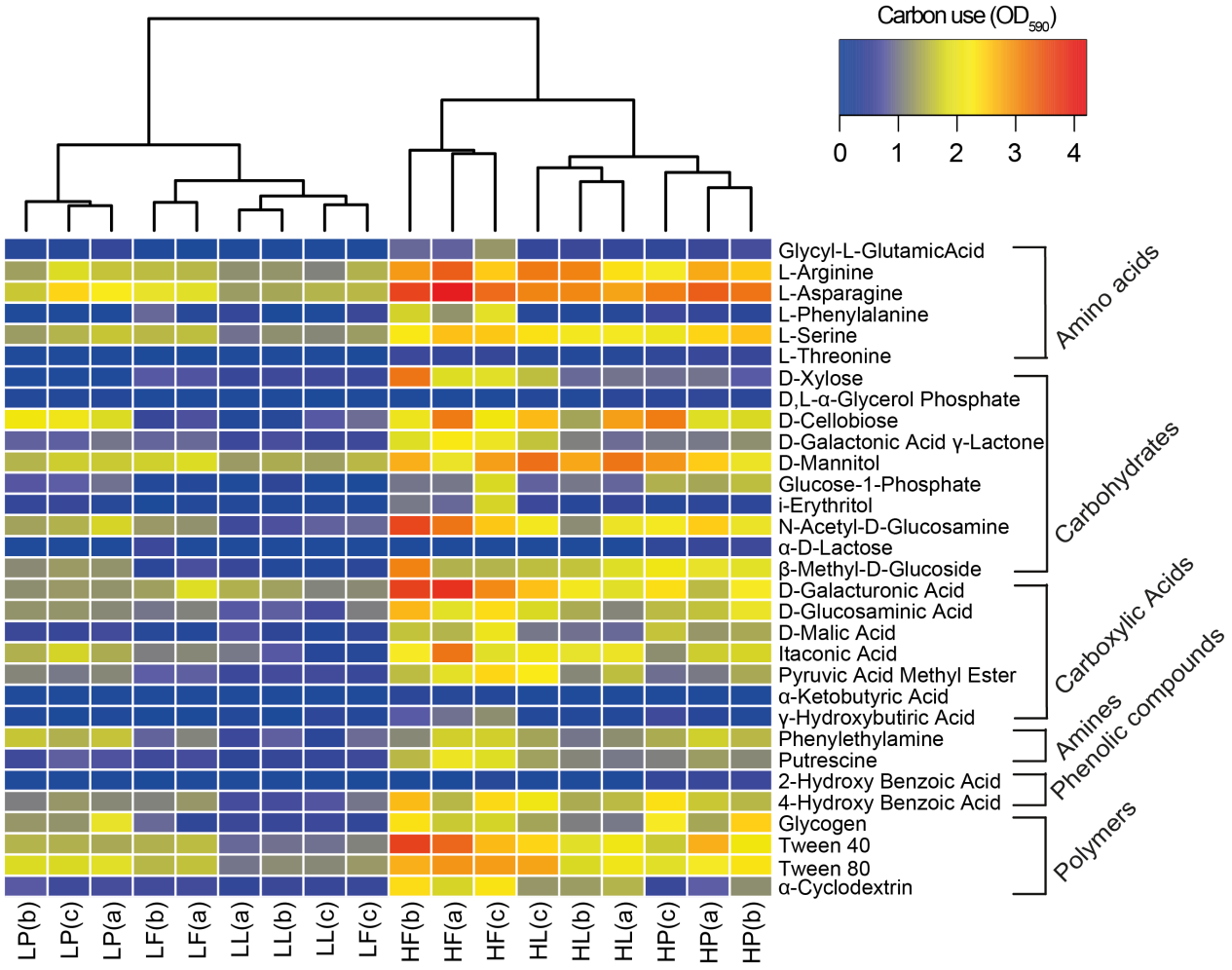


Figure S2. Overview of the taxonomic composition of bacterial and fungal communities in soils. (A) Relative abundance of bacteria and fungi at the phylum level. (B) Venn diagram showing the distribution of bacteria and fungi OTUs in the three soils.

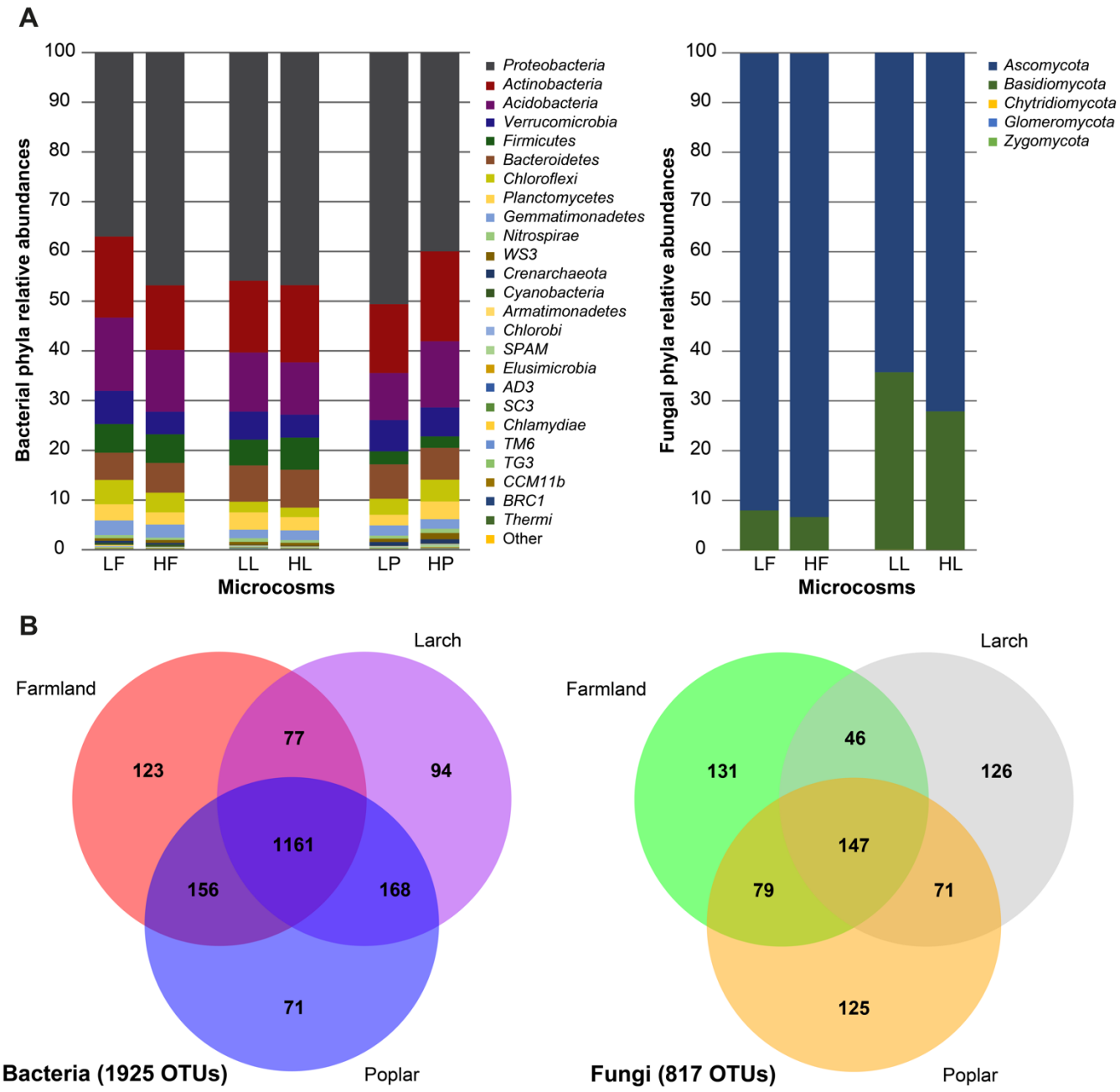


Figure S3. Metagenomic analysis using MG-RAST pipeline. (A) UPGMA agglomerative clustering of soil microcosms according to a Euclidean distance matrix calculated with a Hellinger distance matrix of annotated gene abundance profiles. Gene annotation (COG database) was performed using unassembled sequenced reads with the pipeline MG-RAST. Land-use types are distinguished by three different symbols (square; farmland, circle; larch and triangle; poplar). Black symbols indicate soil microcosms exposed to eH_2 treatment and white symbols indicate soil microcosms exposed to aH_2 treatment. The Barplots show distribution of sequences (CPM values) in COG gene categories for (B) farmland, (C) larch and (D) poplar soil exposed to eH_2 and aH_2.

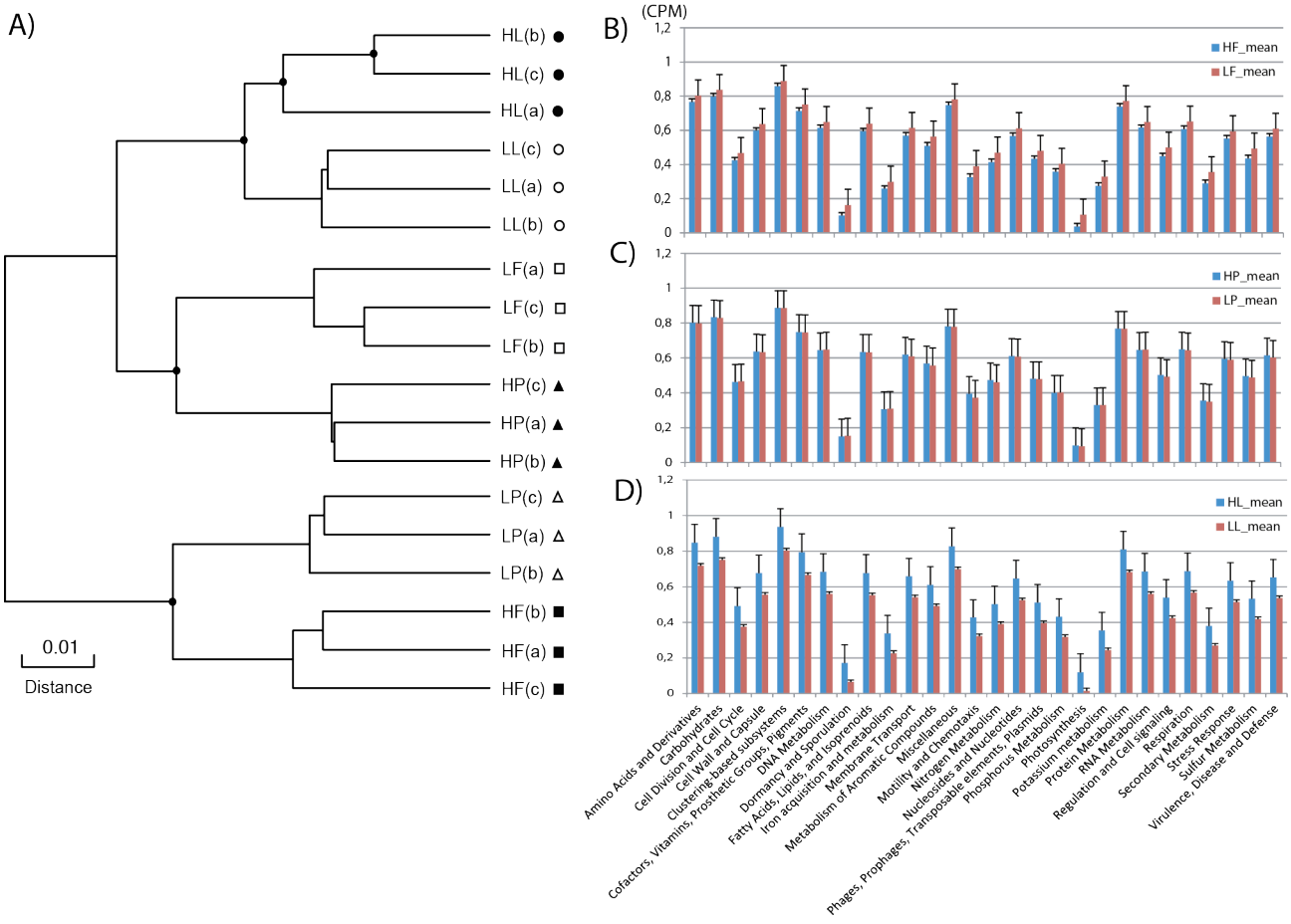
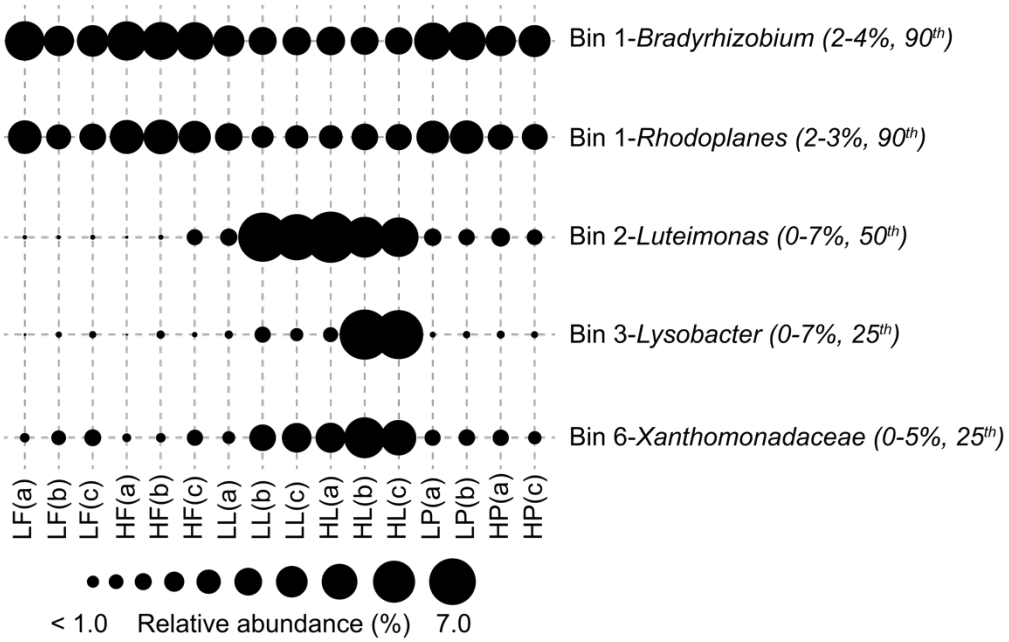


Figure S4. Bubble-chart representation of OTUs associated with the 4 genome bins of interest. Correspondence between genome bins and OTUs was achieved through correlation network analysis. Dots size is proportional to the relative abundance of OTUs in the 16S rRNA gene amplicon sequencing. Genome bins are presented along with taxonomic affiliation retrieved from the 16S rRNA gene amplicon sequencing analysis. The relative abundance and the rank of the connectivity of each OTU are written in parentheses.



References

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. (2009) GenBank. Nucleic Acids Res. 37:D26–31.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. (2012) Ray Meta: scalable *de novo* metagenome assembly and profiling. Genome Biol. 13:R122.

Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–20.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods. 7:335–6.

Eddy SR. (2011) Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

Finn RD, Alex B, Jody C, Penelope C, Eberhardt RY, Eddy SR, et al. (2013) Pfam: the protein families database. Nucleic Acids Res. 42:D222–30.

Grace JB. (2006) Structural equation modeling and natural systems. Cambridge University Press.

Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, et al. (2016) The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). Stand. Genomic Sci. 11:17.

Kang DD, Froula J, Egan R, & Wang Z (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25:1043–55.

Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–2.

Robinson MD, McCarthy DJ, Smyth GK. (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26:139–40.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37:D5–15.

Tang S, Shiyuyun T, Mark B. (2013) *Ab Initio* Gene Identification in Metagenomic Sequences. Encyclopedia of Metagenomics. p. 1–8.