Bioinformatics and Microbiomes:

Metatranscriptomics and analysis of microbial activity

Petr Baldrian

Institute of Microbiology of the CAS
baldrian@biomed.cas.cz

Samples after sampling

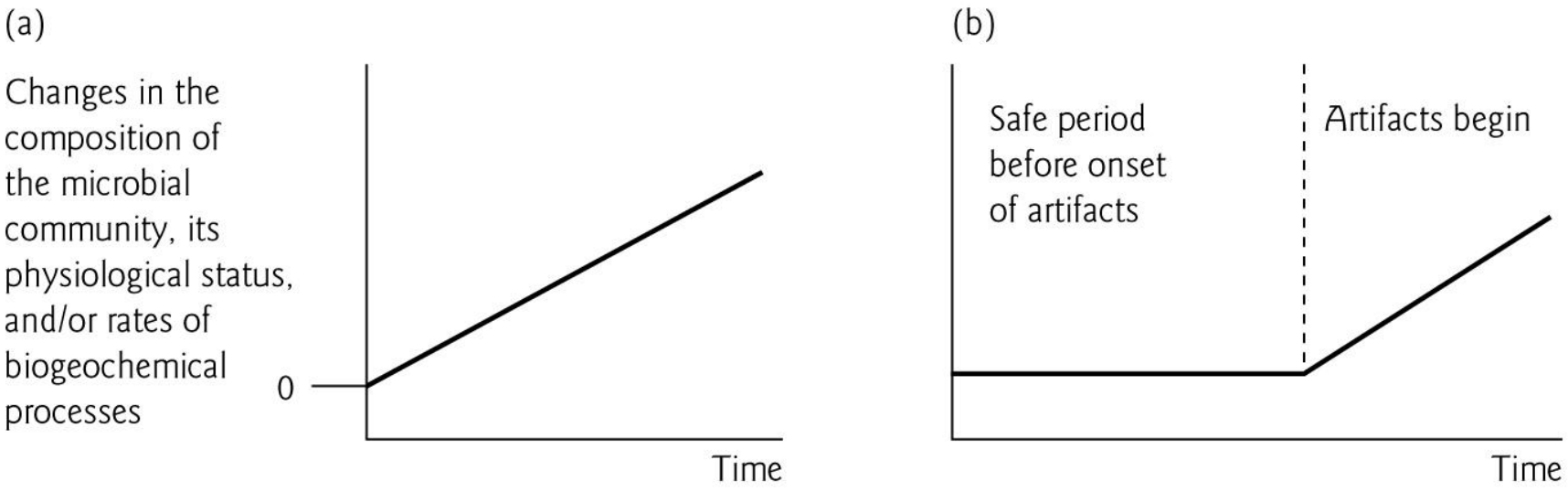


Figure 6.3 Uncertainties in seeking data on in situ biogeochemical processes from samples removed from the field and incubated in the laboratory. The two graphs describe the quantitative and/or qualitative influence of sampling and incubation on biogeochemical processes of interest. (a) Changes in environmental samples may begin the instant they are disturbed in a field site, or (b) after some uncertain "safe period" during which valid measurements may theoretically be completed. (From Madsen, E.L. 1996. A critical analysis of methods for determining the composition and biogeochemical activities of soil microbial communities *in situ*. *In*: G. Stotzky and J.-M. Bollag (eds) *Soil Biochemistry*, Vol. 9, pp. 287–370. Copyright 1996. Reproduced by permission of Taylor and Francis Group, a division of Informa plc.)

Analysis of environmental processes in the real environment

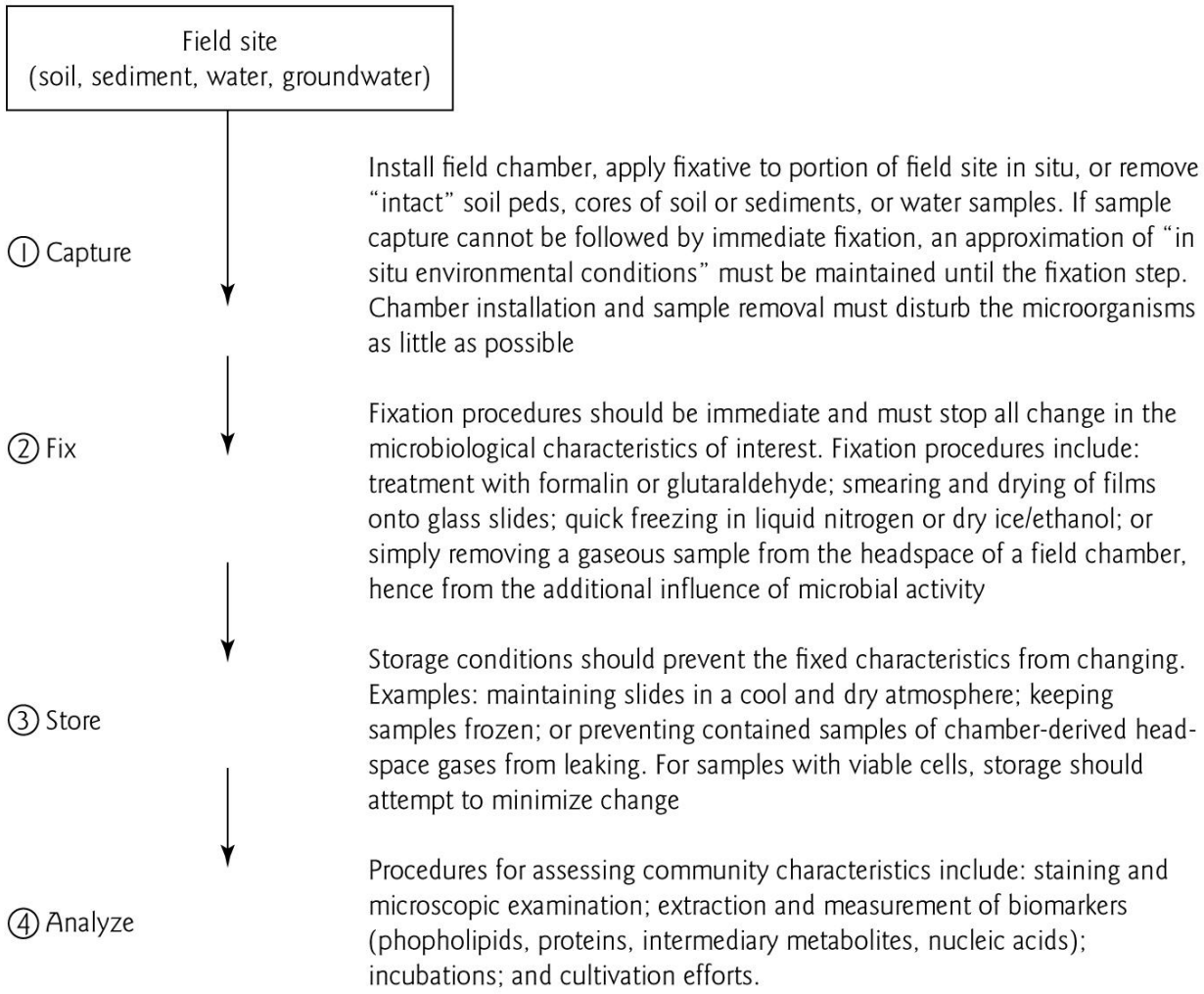


Figure 6.4 Four-step methodological scheme for sampling and processing and generating information from microbial communities in nature. (From Madsen, E.L. 1996. A critical analysis of methods for determining the composition and biogeochemical activities of soil microbial communities *in situ*. *In*: G. Stotzky and J.-M. Bollag (eds) *Soil Biochemistry*, Vol. 9, pp. 287–370.

Exploring ecosystem functioning - metagenomics

Opportunities:

- can indicate genetic potential of the community
- DNA community relatively stable over time (good representation of the ecosystem)
- can theoretically reveal co-occurrence of genes (when longer contigs of DNA covering several genes are assebled)
- sometimes gives exact identity of the gene source (when the gene sequence and 16S co-occur on an assebled molecule)
- powerfull for exploration of bacteria / archaea
- long read sequencing can recover whole bacterial genomes

Limitations:

- eukaryota (e.g. fungi) contain much of noncoding DNA (up to over 90%)
- eukaryotic genes contain introns
- due to the above, short reads of eukaryotic genes can give no information (noncoding DNA) or be difficult to assign (contains both exons and introns)
- need for assembly
- potential is not the function, presence is not activity (extracellular DNA, pseudogenes, levels of expression…)
- for diverse ecosystems, high depth of sequencing required

Metatranscriptomics - opportunities

- can indicate real activity in the studied ecosystem; fast response to disturbance / experimental treatment
- little danger of „ancient“ RNA from dead cells – such RNA decomposes rapidly
- avoids the problem with noncoding DNA
- for gene-coding sequences, functional and taxonomic assignment is more simple than for DNA, even for shorter reads
- powerfull for exploration of both prokaryota and eukaryota
- in the case eukaryota, ease of isolation and purification by „fishing“ the poly-A tails of mRNA molecules (but does not always work)
- metatranscriptomes much less complex than metagenomes - results in easier assembly (higher coverage with the same number of reads)
- with sufficient depth of sequencing, relative importance of individual processes can be analysed to some depth by comparison of transcription level
- while metagenomics tell which genes **may be involved**, metatranscriptomics tell which genes actually **are involved** (expressed)

Metatranscriptomics - limitations

- expression is highly regulated and corresponds to „actual“ conditions, not „mean“ conditions of the site; for example, transcription increases by orders of magnitude when dry soil is moistened
- mRNA is short-lived so the metatranscriptome reveals what happened within last tens of minutes
- the amount of extracted RNA usually makes amplification necessary; PCR amplification of cDNA brings bias
- extracted RNA contains much rRNA that can be difficult to remove
- genes with low level of expression are difficult to recover
- there is little (if any useful) information on mRNA stability in time and translation rate and thus the amount of protein molecules synthesized per mRNA molecule in its lifetime
- metagenomics can theoretically deliver long contigs - chromosome fragments with multiple genes that are from the same genome; this is impossible for metatranscriptomics

Workflow

Sampling
- sample collection
- stabilization
- storage

Library preparation
- RNA isolation
- RNA purification
- DNA removal
- rRNA removal (to recover all mRNA) or capture of eukaryotic mRNA (poly A)
- RNA fragmentation
- cDNA synthesis
- RNA removal
- adapter ligation (with barcodes when multiple samples will be sequenced)
- amplification
- selection of molecules of approppriate size

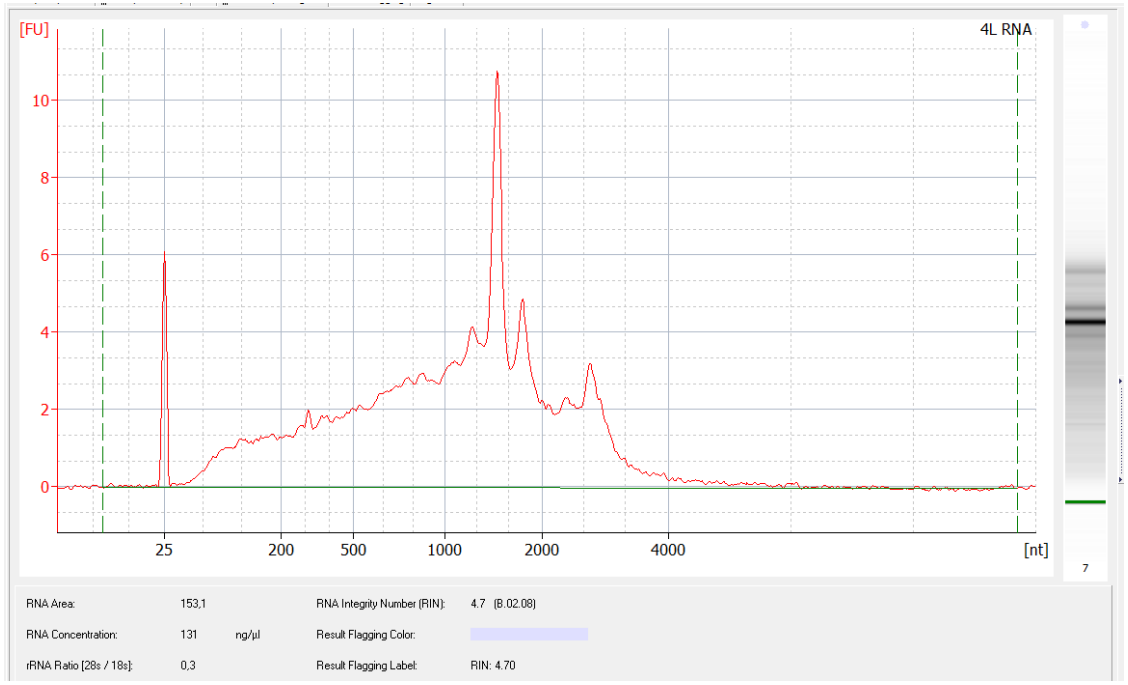
Sequencing

Data analysis
- long sequences – direct annotation (function, binning to microbial taxa)
- short sequences – sequence assembly and annotation of contigs (scaffolds)


USE LIQUID NITROGEN ONLY
USE LIQUID NITROGEN ONLY

RNA isolation and purification

- sample homogenization by mortar and pestle in liquid nitrogen
- MoBio RNA PowerSoil Kit used for isolation in combination with Zymo Research OneStep PCR Inhibitor Removal Kit (PVPP removal of humic and fulvic acids)
- DNA removal (DNAse)
- verification of DNA removal (no PCR amplification of 16S)
- check of RNA yield (Qubit)
- check of RNA quality (BioAnalyzer)
- storage of isolated total RNA (-80 °C)

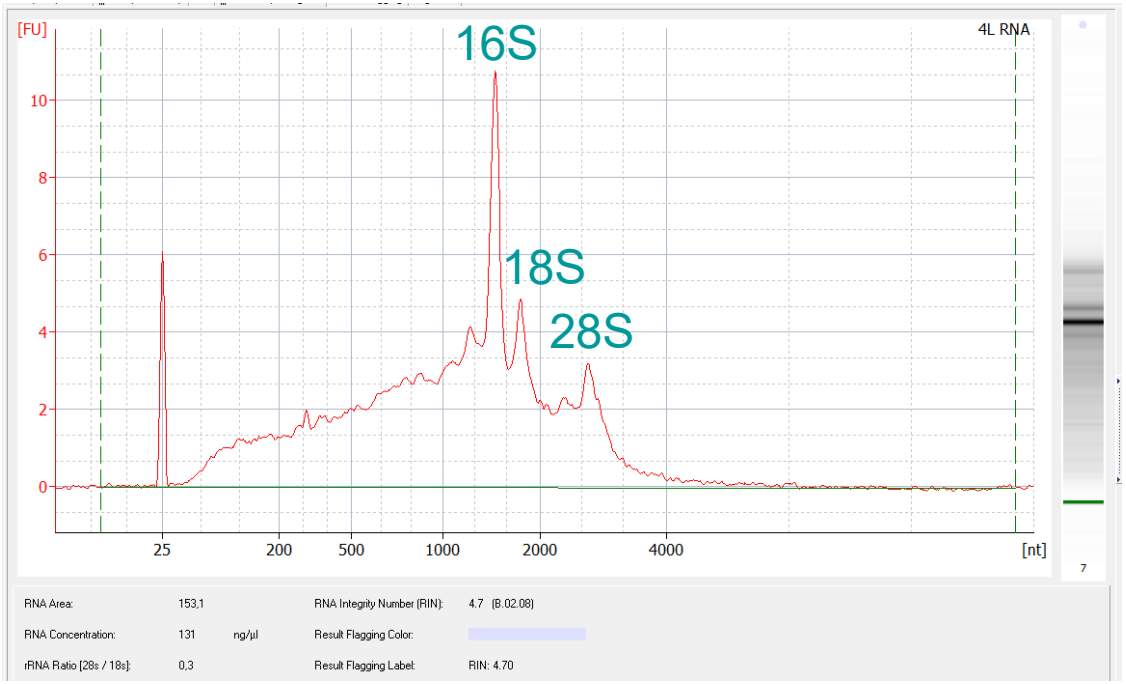


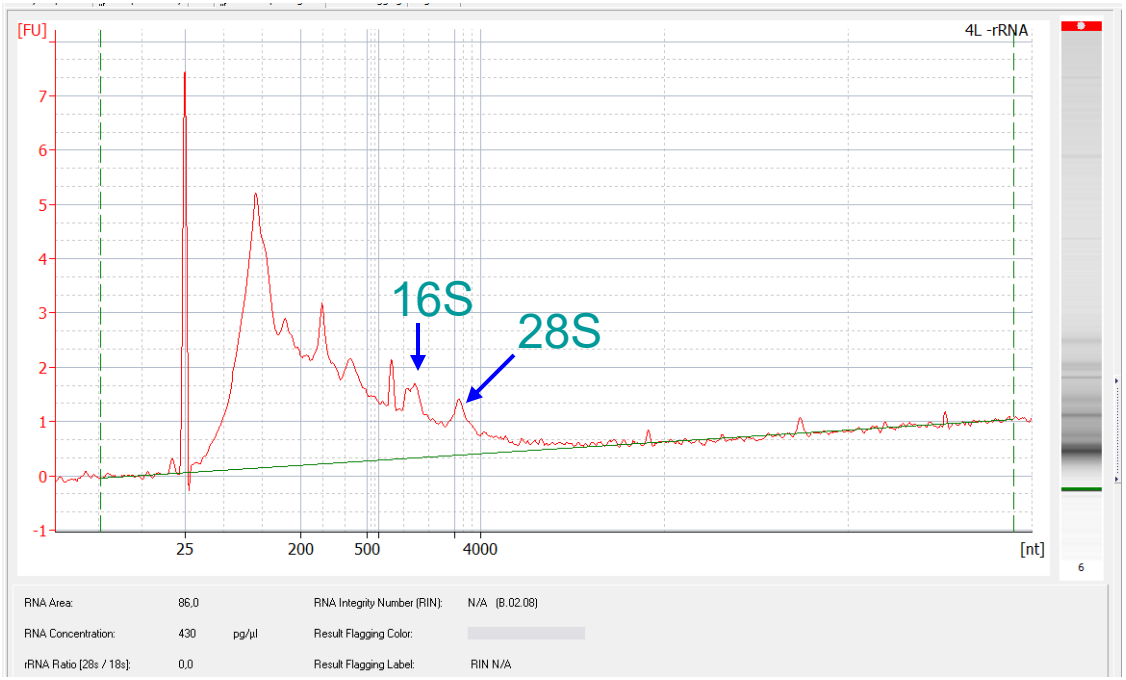
Removal of ribosomal RNA

- use of Epicentre RiboZero kits

- to remove both bacterial and eukaryotic rRNA, RiboZero Metabacterial and RiboZero Human/Mouse/Rat kits were combined

- efficiency of rRNA removal verified on Bioanalyzer





Library preparation

- RNA fragmentation to get desired size distribution of molecules

- cDNA synthesis (reverse transcription); success check by amplification of 16S from cDNA

- RNA removal (RNAse treatment)

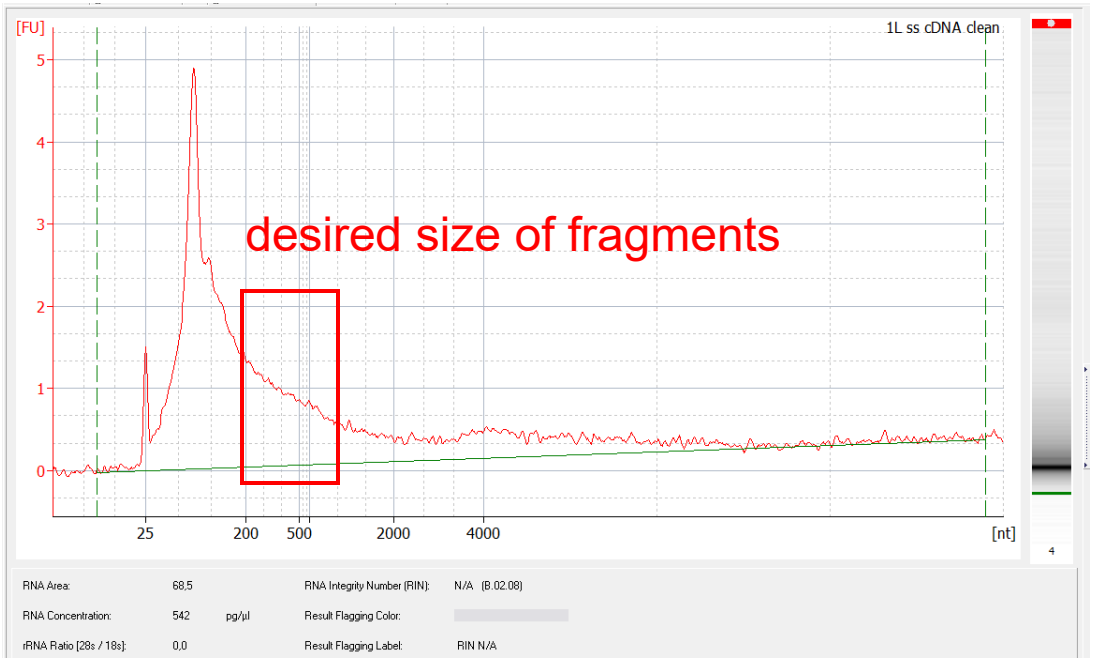
- Illumina adaptor ligation to cDNA

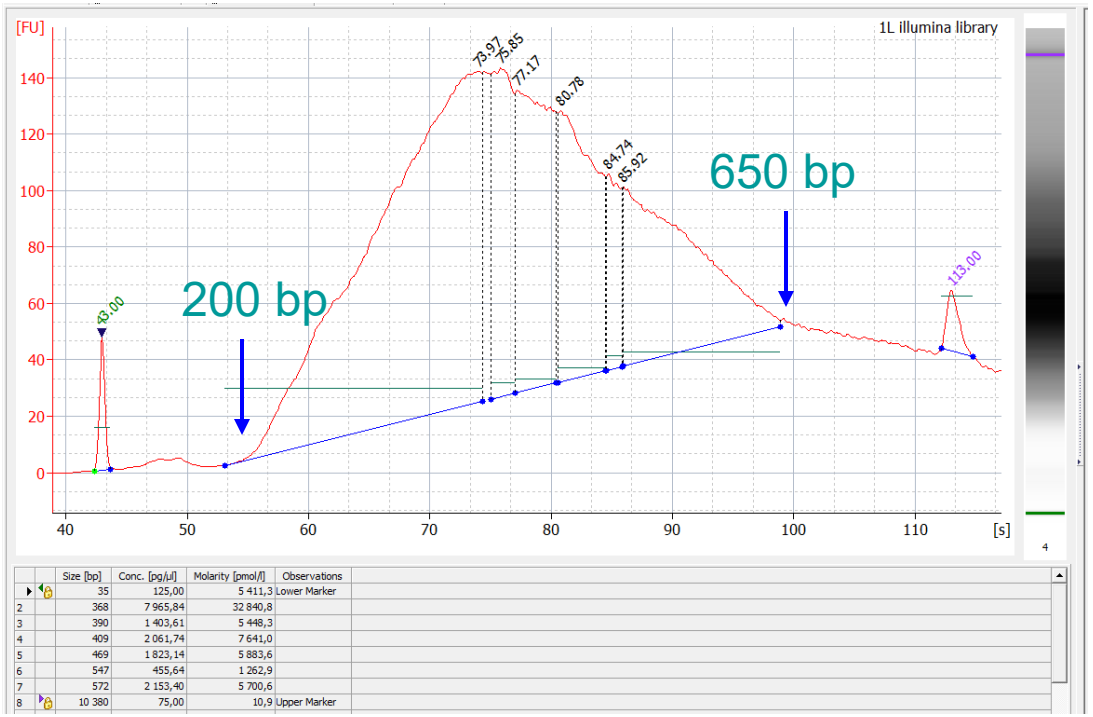
- amplification from Illumina adaptors with barcodes for each sample (10-15 cycles)

- check of DNA yield and size distribution

- removal of short fragments (Agencourt AMPure beads) and long fragments (cut from gel)

- equimolar combination of samples into one common library





Sequencing and data analysis (largely applies to metagenomics as well)

Samples are typically sequenced on Illumina HiSeq/NextSeq (2 x 150 or 2 x 250 bases pair-end reads)

Theoretically, one lane should deliver up to 350 millions pair-end sequences. For highly diverse microbiomes, 10-12 metatranscriptome or metagenome samples fit per one lane.

Reads can be annotated directly or after assembly.

Long read sequencing platforms can now read large part of genes or even whole operons

Assembly of reads into contigs

By overlap of reads, short sequences (150-base **reads**) are assembled into longer ones (contigs, 300-4 000 bases for metatranscriptome, 300-100 000 bases for metagenome).

Assembly rules will be presented elsewhere

What to remember about assemblies:

- assembly is done by overlap at each position, there is a minimal number of overlapping reads needed to generate contig (such as 5) => **anything unique, present in less than five copies will not generate contig; low expression genes are not assembled**
- assemblers do not understand same genes in different contexts (between different neighbouring genes) => **identical sequences with multiple copies per genome (rRNA including 16S, ITS) are not assebled; this for sure also happens to some genes**
- furthermore: assemblers do not like same **short** sequences within different context => **conserved regions of highly abundant genes are likely discarded as chimeras** (noone knows how frequently this happens, vulnerable genes: those common to all organisms)
- **chimeras** can be created but **can not be recognised** (noone knows how frequent this is, (small) „mock community“ tests (mixing genomes of a few microbes) tell that probably not often (???)

Annotation of sequences

Direct annotation (we annotate 150-base or 250-base **reads**):

- short sequences – low likelihood to find similar sequences, reads often miss variable parts
- all reads may be annotated (theoretically, some short sequences are not found in databases)

Annotation of contigs (we annotate 300-4000 base **contigs**):

- longer sequences – higher probability of finding functionally important features and specific, variable parts of genes that distinguish it from others and allow fine annotation

- annotation using HMMER profiles (sequence motives of active site) – higher probability of HMM recovery

- only those reads that were assembled into contigs (typically 40-80%) may be annotated

- **the longer the sequence, the better the annotation; best is to annotate whole genes**

Annotation of sequences

Direct annotation of reads (we annotate 150-base or 250-base **reads**):

- short sequences – low likelihood to find similar sequences, reads often miss variable parts
- all reads may be annotated (theoretically, some short sequences are not found in databases)

Annotation of contigs (we annotate 300-4000 base **contigs**):

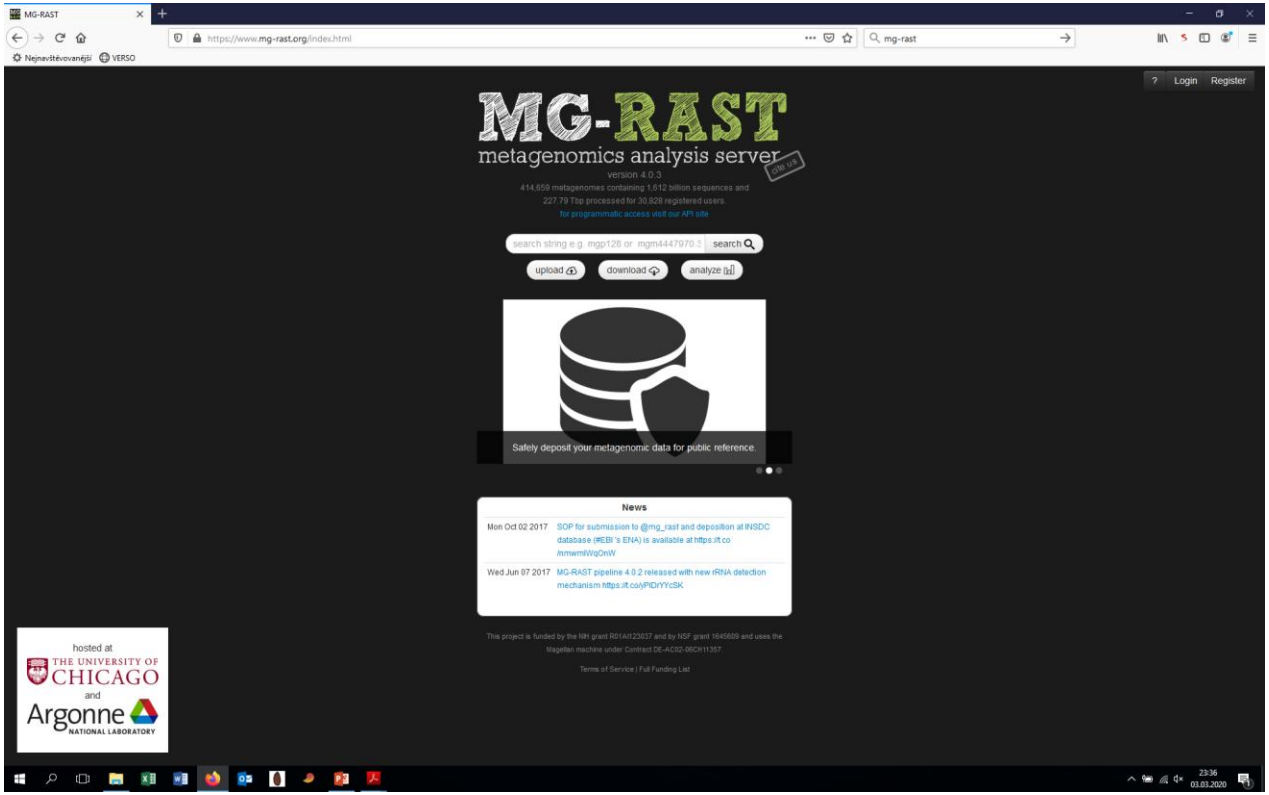
- longer sequences – higher probability of finding functionally important features and specific, variable parts of genes that distinguish it from others and allow fine annotation
- annotation using HMMER profiles (sequence motives of active site) – higher probability of HMM recovery
- only those reads that were assembled into contigs (typically 40-80%) may be annotated
- **the longer the sequence, the better the annotation; best is to annotate whole genes**

Annotation

- made online, e.g. in **MG RAST** (our example), IMG Gold (JGI for their projects) or in house

MG RAST

- Example of search (BLAST)-based annotation
- Any user data can be uploaded; runs **against various databases**



Annotation

- made online, e.g. in **MG RAST** (our example), IMG Gold (JGI for their projects) or in house
- **neither MG RAST nor IMG Gold provides any useful annotation of fungal genes**
- each **nucleotide** sequence is translated in all six reading frames (three forward, three reverse) into **peptide sequences – gene calling**
- too short sequences (many stop codons) are discarded, best reading frame and best (longest) peptide sequence is retained for annotation – gene calling results
- for longer contigs, whole genes are sometimes recovered
- transcripts do not contain introns (…for most of their lifetime), so gene calling of **eukaryotic sequences** is <u>much better</u> for meta**transcriptomes** than for metagenomes
- there are some tools to remove introns (one designed by our group)
- gene calling for **prokaryotic sequences** is <u>slightly better</u> for meta**genomes**, since most genes are recovered as complete

Annotation

- annotation considers **peptide (protein) sequences,** due to nucleotide variation, use of nucleotide sequences would lead to too little recovery

Annotation is based on either:

(1) finding the **best match sequence** in certain database
(2) finding some **specific feature** in the sequence

Finding the **best match sequence** in certain database

- finding of the longest available sequnece with the highest similarity, i.e., sequence with the **lowest E-value** (E: probability that the sequences do not have anything in common) – BLASTp or DIAMOND BLAST , identification of **best hit sequence**
- **This approach is used to generically annotate all sequences**

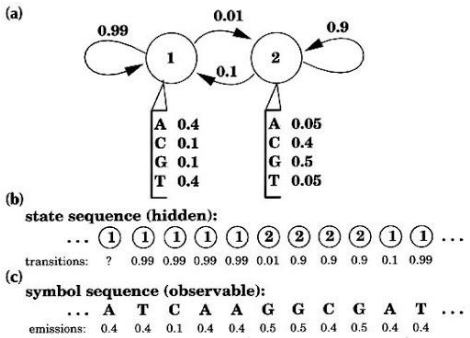
- function of the gene (description) is retriever from the database used, e.g. as the hierarchical function classification (KEGG – see later)

- the identity of the microbe is retrieved from the **best hit sequence** or as a „consensus“ the **last common ancestor** of top 10, 100 hits (e.g., **MEGAN6 pipeline**)

Annotation

Finding some **specific feature** in the sequence

- finding some functional feature, such as the sequence of the active center of an enzyme
- functional features are often described as **Hidden Markov Models (HMM)**

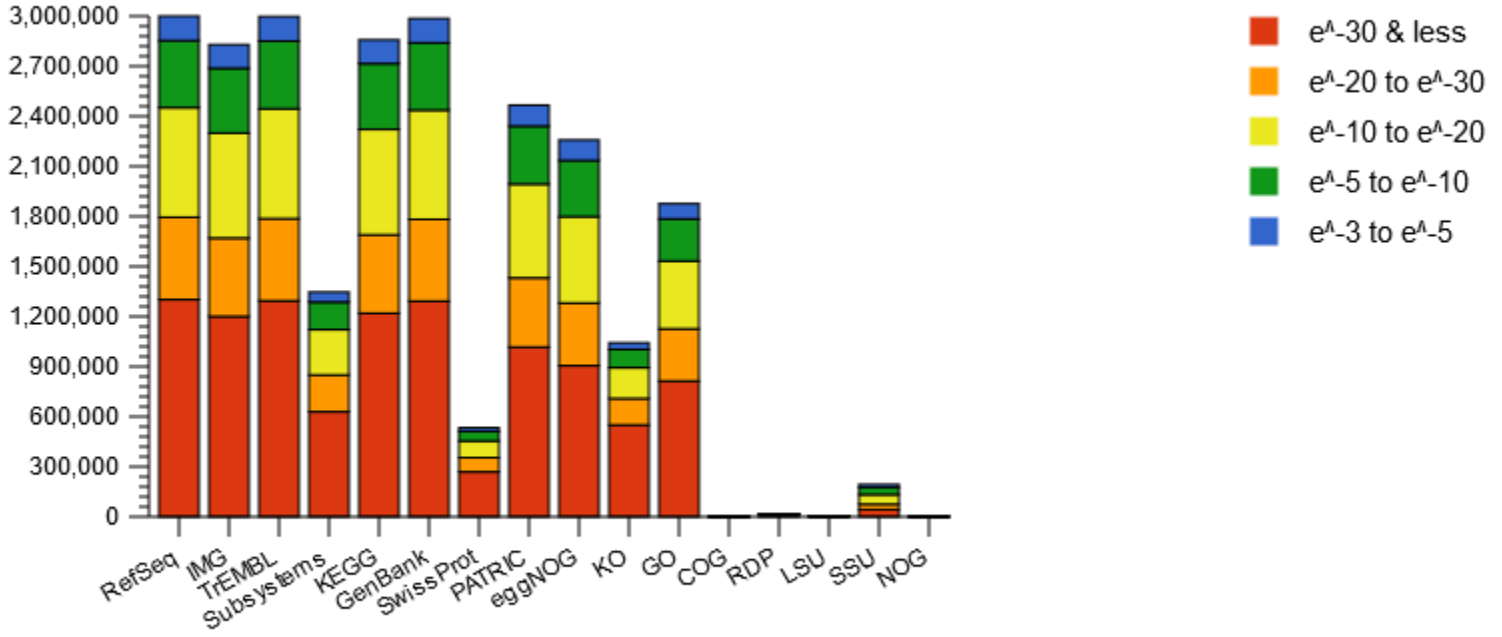


- **This approach can be used only for some sequences: for those, where the HMM exists**

- available for Carbohydrate active genes (www.cazy.org)
- some other genes (e.g. N-cycling) Functional Ontology Assignments for Metagenomes (https://omictools.com/foam-tool)
- not available for other proteins (such as, e.g., for ribosomal proteins)

- **some sequences can be annotated by BLAST and HMM (what is best)?**
- **HMM annotation does NOT (typically) provide taxonomy information**

Annotation – how good is good?



Can we test the quality of annotation (quality = probability that annotation is correct)

For taxonomic annotation
For functional annotation

For taxonomy: annotation of known genomes, as if they are unknown

Function frequency – per base coverage

Each contig in assembly is created from different numbers of reads

Frequency of occurrence (metagenome) or of transcription (metatranscriptome):

Per base coverage = number of reads mapping to contig * read length / contig length

Done by mapping of reads back to contigs

Short read versus long read sequencing

HMW DNA isolation
(3 samples)

Fragmentation

Library preparation

Assembly

Read mapping

Results

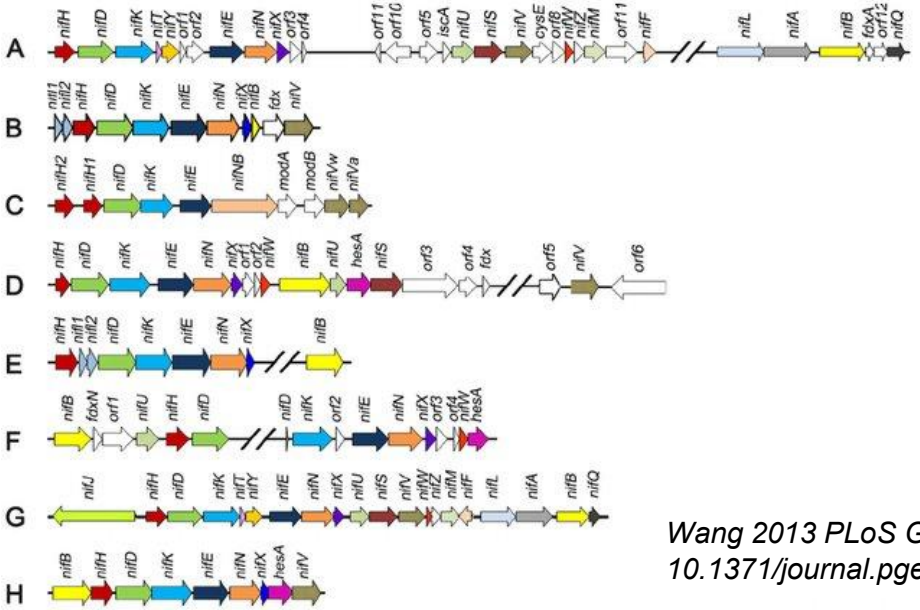
Size selection

Library preparation

Results
(99%, 99.9% precision)

Comparison

Short read versus long read sequencing – operon completeness



Wang 2013 PLoS Genetics
10.1371/journal.pgen.1003865

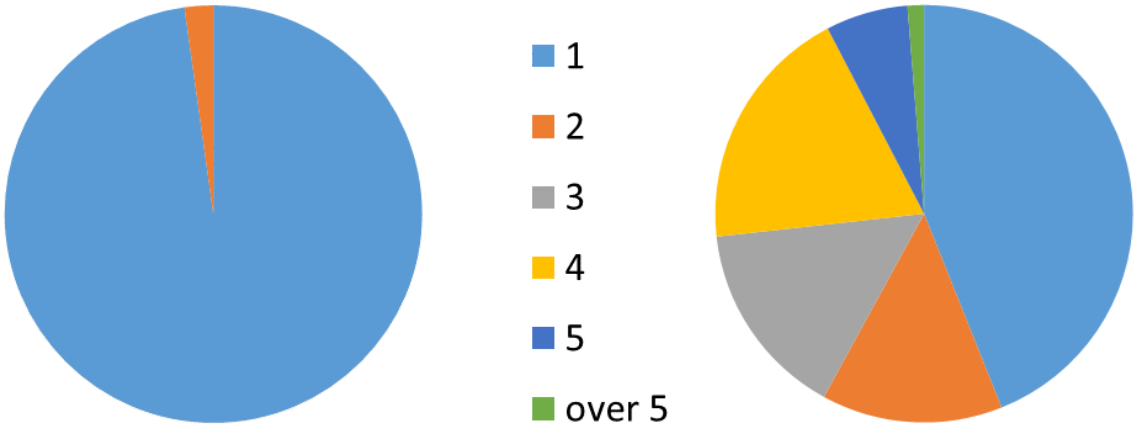
Contigs encoding for nitrogen fixation genes contain multiple nif *genes, typically four or more*

Multiple genes are needed for full functionality

*In assembled **Illumina** data, only two* nif *genes are found in one contig.*

*In **PacBio** data, more than one half of contigs contain two and more* nif *genes*

*Most **Illumina** genes are incomplete (<600 nucleotides).*



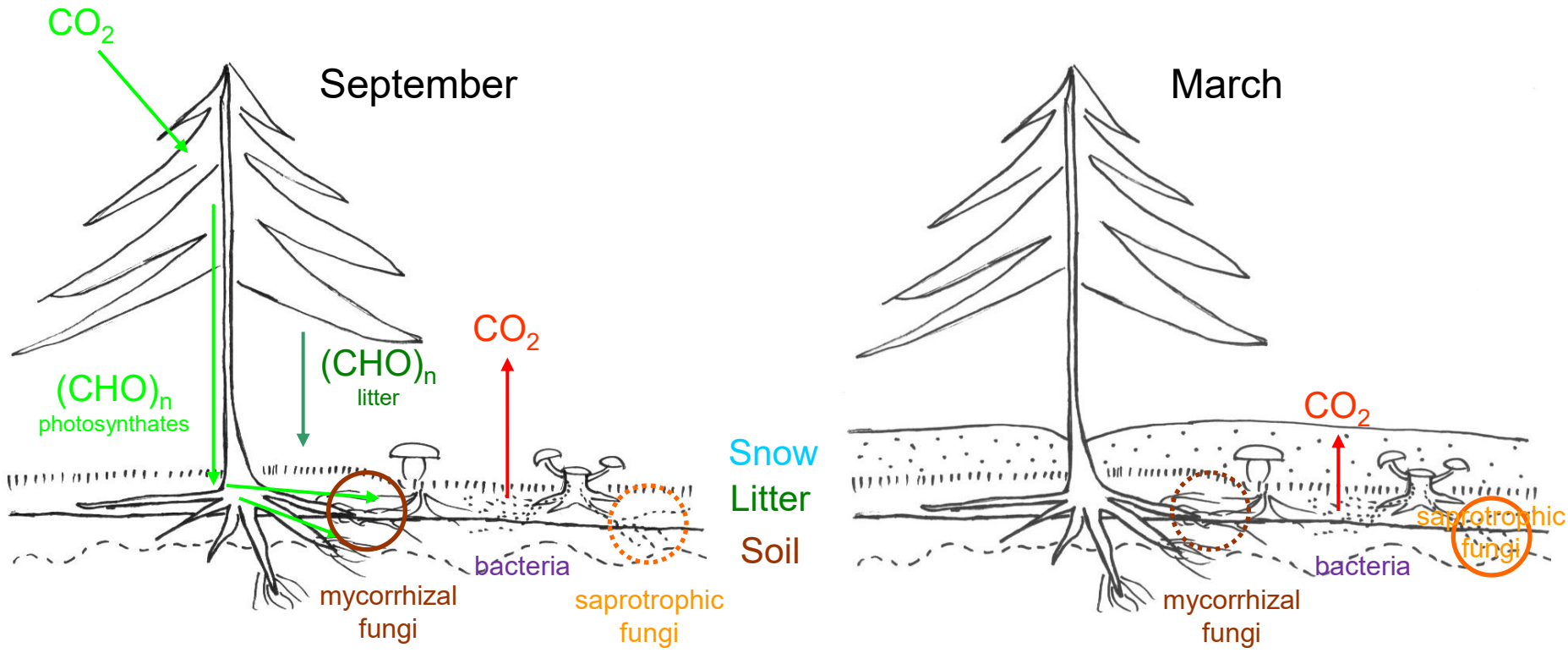
Short reads (Illumina) versus long reads in metagenomics

Oxford Nanopore Promethion Solo 2

- ***PacBio instruments** deliver data of **sufficient quality** for reads of lengths up to 20 kilobases*
- ***Oxford Nanopore Promethion solo 2** gives **cheap data** but considerably **lower quality***
- ***No assembly bias** if long reads are directly annotated. Repeats, introns or highly similar sequences (including bacterial 16S, fungal ITS) are sequenced*
- *Complete genes are more reliably annotated. Long reads may show different taxonomic profiles – more Eukaryota*
- *Long reads recover **sequences of rare biosphere taxa**. Illumina had hits to 4000 genomes, PacBio to 5200 different genomes.*

- *PacBio sequencing is considerably **more expensive***
- ***Lower quality** in Nanopore PS2 – assembly needed*
- *It can be **challenging to get long DNA/RNA** from some samples*

Seasonality of forest soils



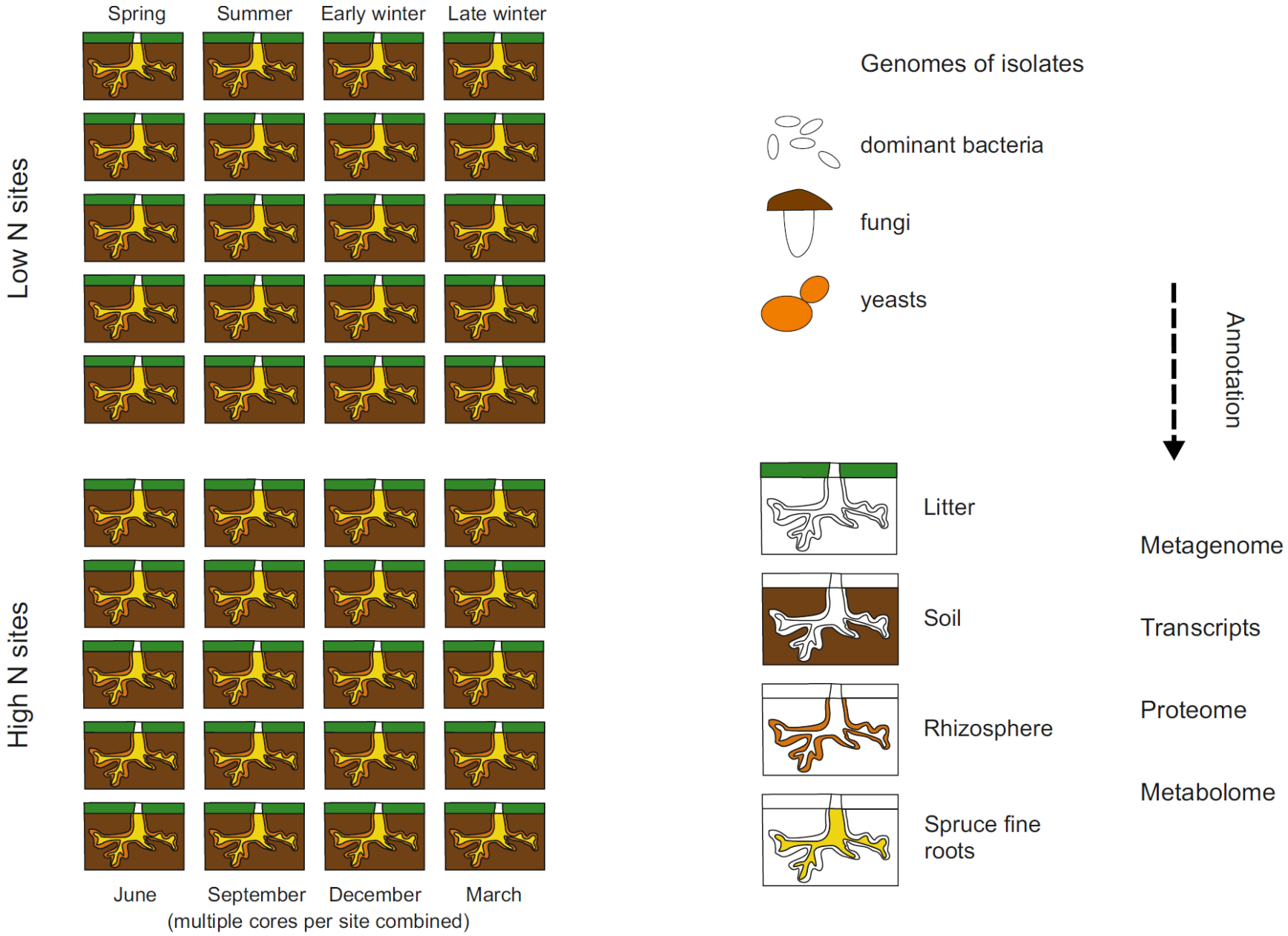
- Seasonal differences in rhizodeposition affect the nutrient availability in soil
- Summer – high input of simple C compounds through tree roots
- Winter – nutrient limitation due to absence of simple C, decomposition of organic matter
- Activity of root-associated microbes, such as mycorrhizal fungi, expected to decrease in winter

Seasonality of forest soils

September - soil temperature 15°C

March - soil temperature 2°C

Seasonality of forest soils



Seasonality of forest soils

Sampling

5 sites x 4 compartments (litter, soil, rhizosphere, root) x 4 seasons (March, June, September, December) = 96 samples

Sampling site: *10 m^2*

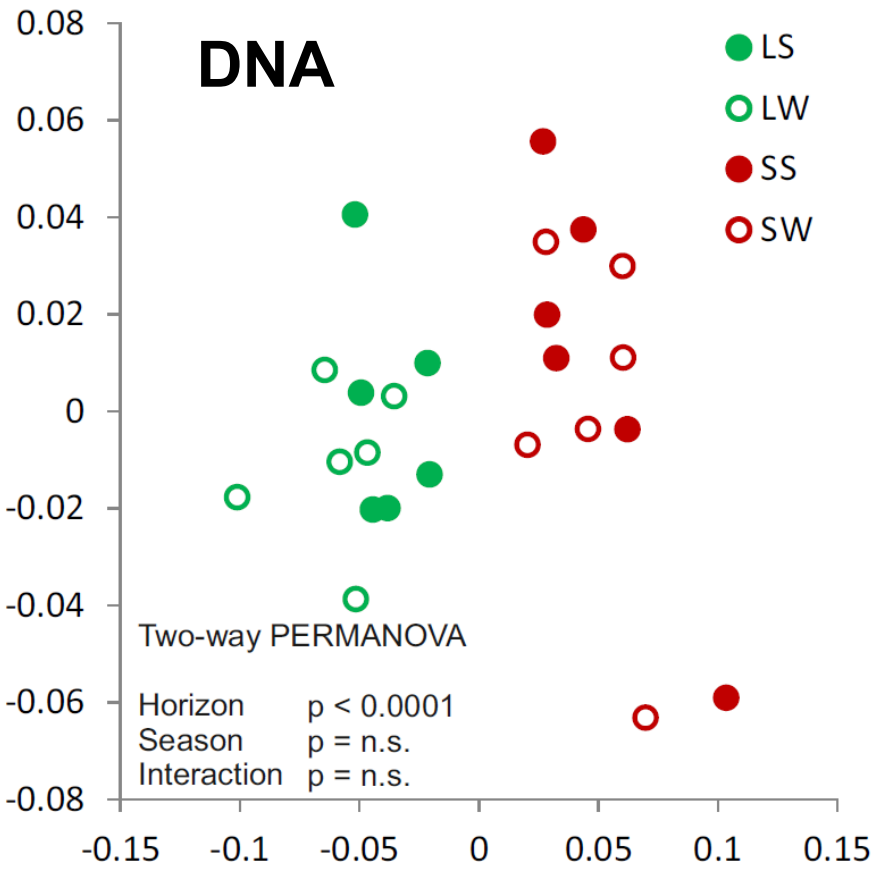
Number of soil cores: *8 cores per site*

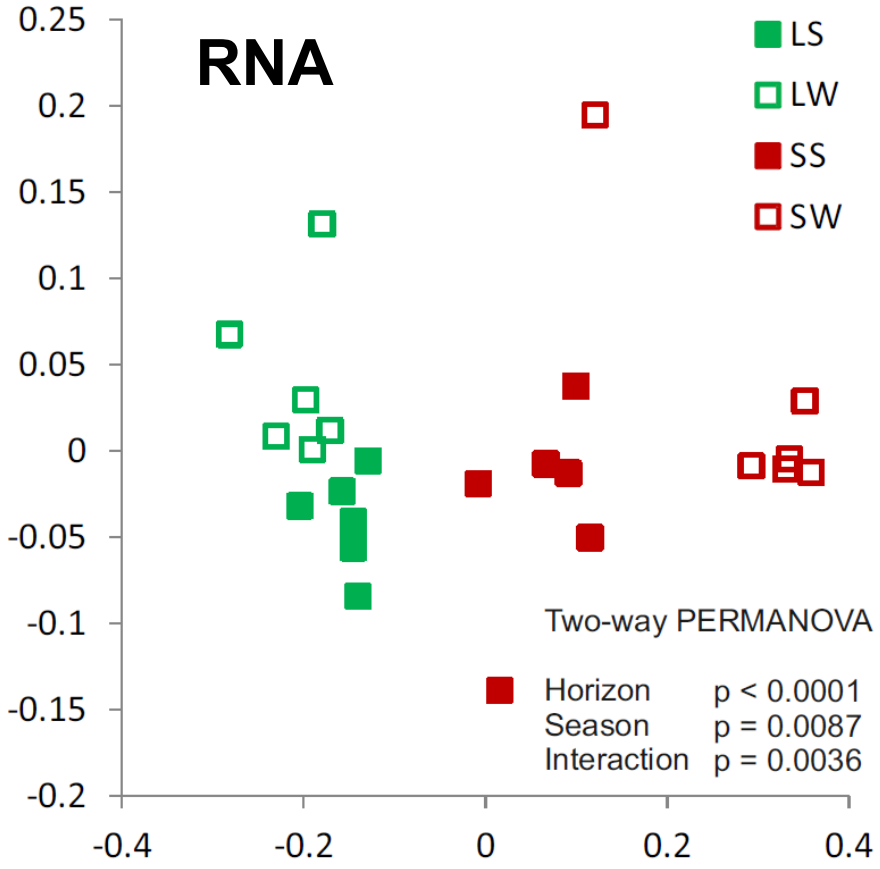
RNA extractions: *Minimum 3 high-quality replicates, min. 250 mg/sample*
RNA preservation in -80°C for up to 1 year

DNA extractions: *Minimum 2 high-quality replicates, min. 250 mg/sample*
DNA preservation in -24°C for up to 2 years

PCR replication: *At least three PCR replicates of each sample*

Seasonality of microbial carbohydrate use





Soil horizons differ in gene pool (community composition) as well as expression

Seasonality is apparent in the transcript pool only

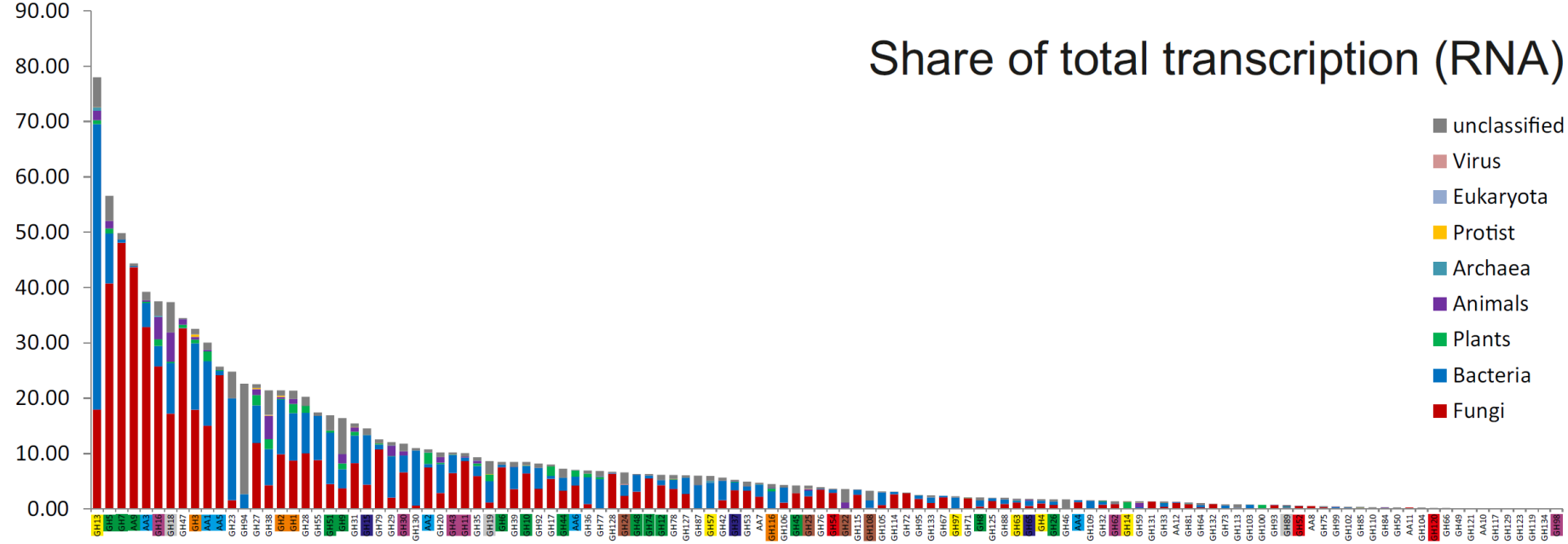
Carbohydrate utilization in forest soil - high functional redundancy

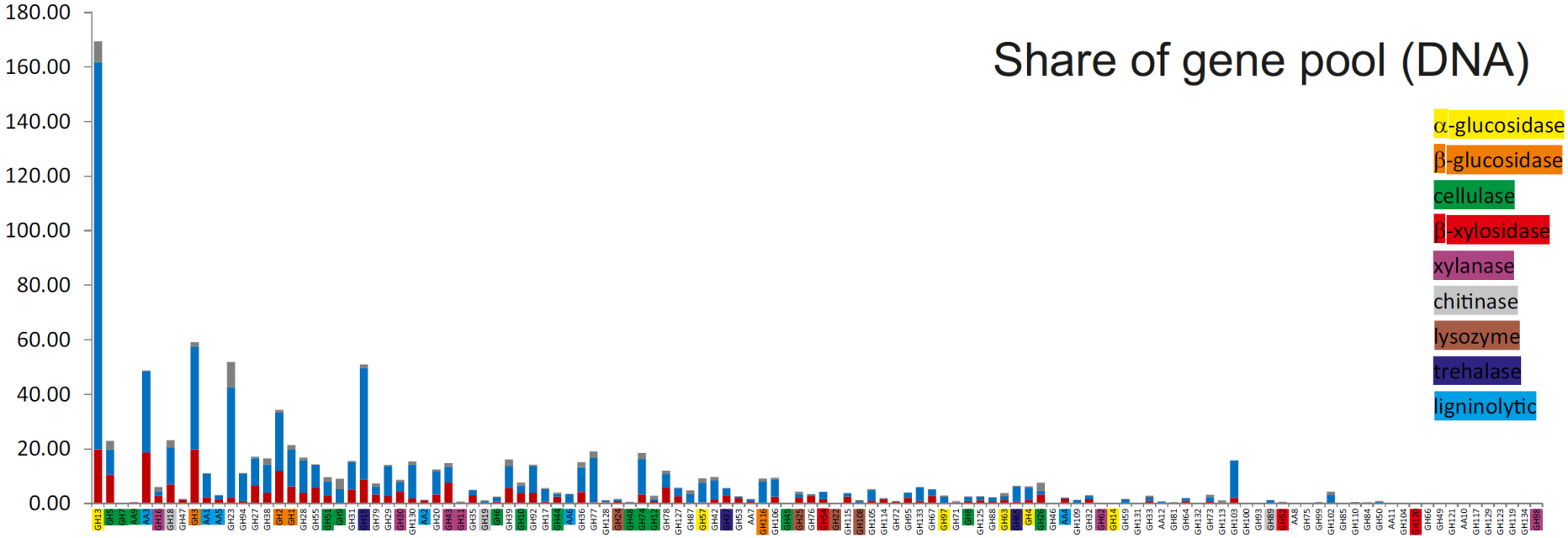
Family	Activity	DNA	RNA
GH1	beta-glucosidase/beta-xylosidase	2060	939
GH2	beta-glucosidase	3359	1145
GH3	beta-glucosidase/beta-xylosidase/endoxylanase	5571	1707
GH4	alpha-glucosidase	526	79
GH5	beta-glucosidase/exocellulase/endocellulase/beta-xylosidase/endoxylanase	2427	2011
GH6	exocellulase/endocellulase	241	256
GH7	exocellulase/endocellulase	29	576
GH8	endocellulase/endoxylanase	345	112
GH9	beta-glucosidase/exocellulase/endocellulase/enodxylanase	889	426
GH10	endocellulase/endoxylanase	812	386
GH11	endoxylanase	58	178
GH12	endocellulase/endoxylanase	242	170
GH13	amylase/alpha-glucosidase/trehalase	16412	4574
GH14	amylase	1	34
GH15	trehalase	4542	803
AA1	laccase	897	1262
AA2	ligninolytic peroxidase(s)	70	366
AA9	lytic polysaccharide monooxygenase(s)	60	522

Typically hundreds to thousands genes of a single CAZY family are present in the metagenome

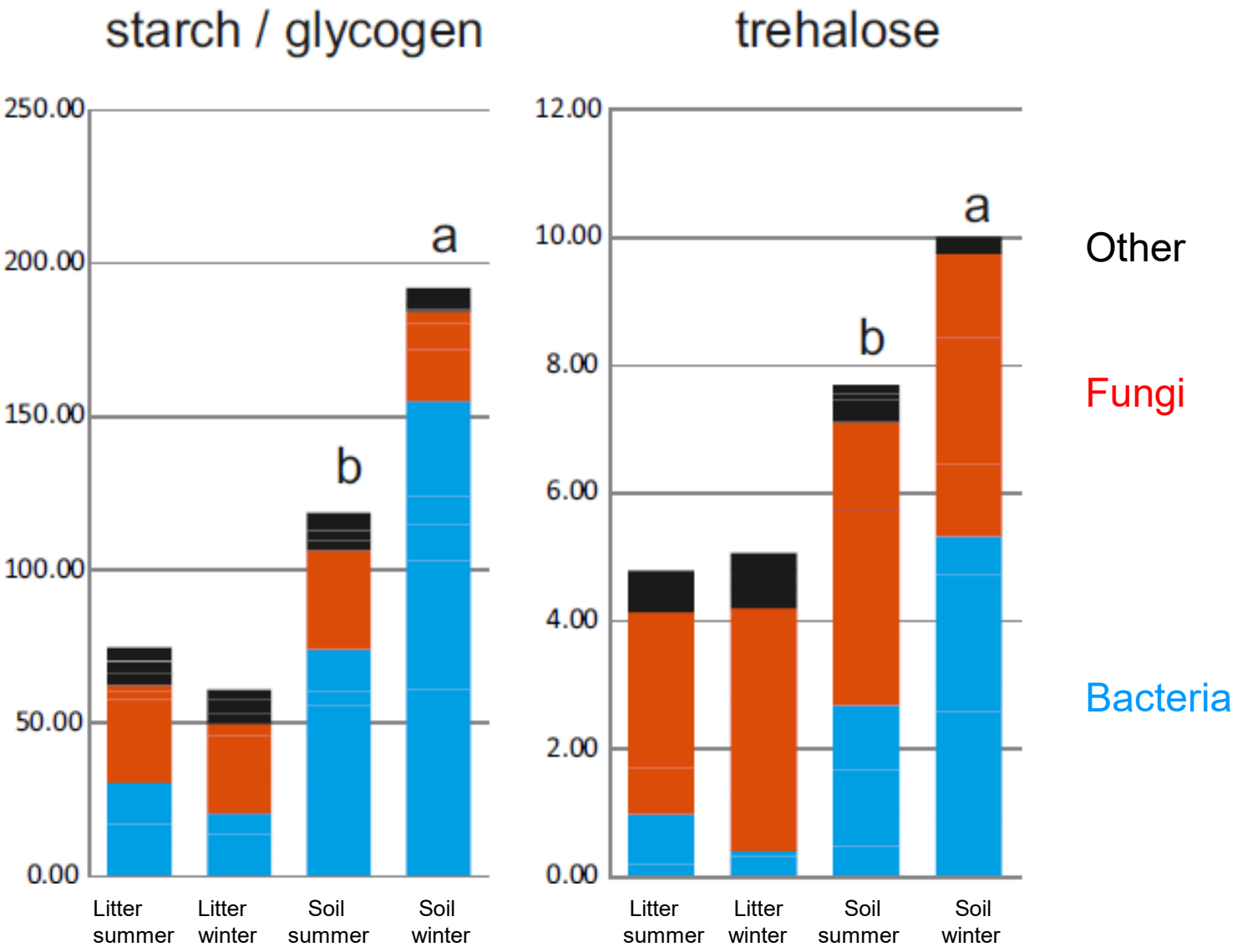
The number of expressed genes also ranges in hundreds to thousands

Seasonality of microbial carbohydrate use: most trascribed gene families





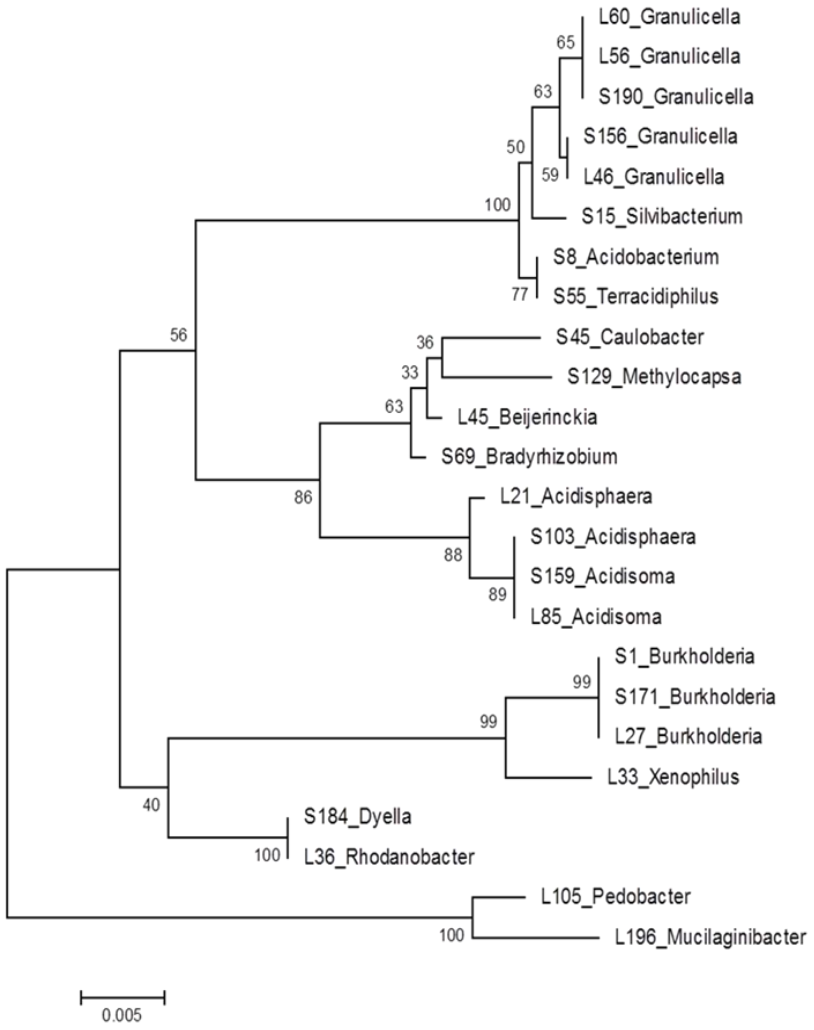
Highly expressed in winter soil – use of storage compounds

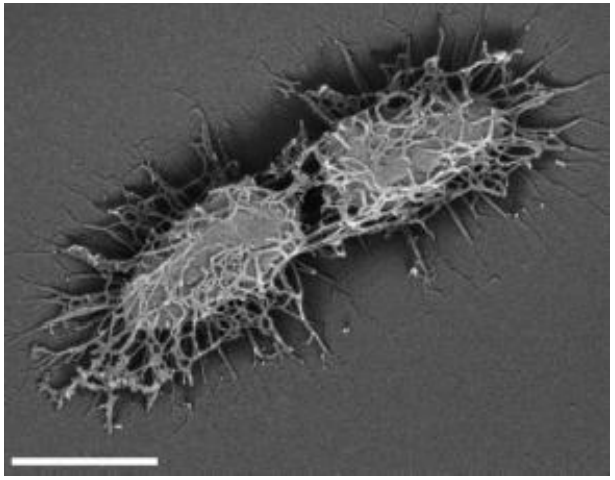


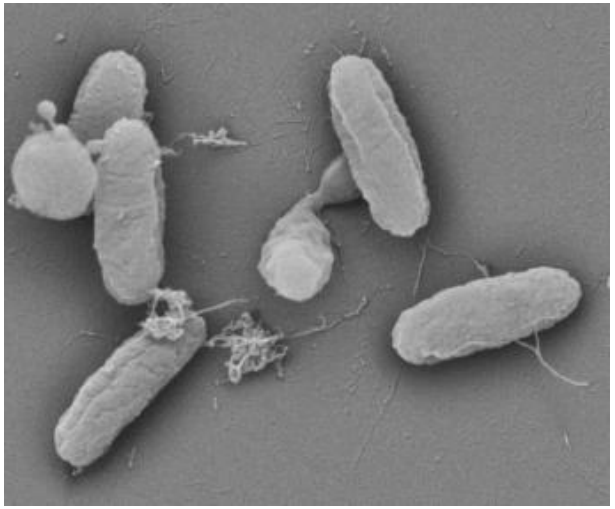
Utilization of starch and trehalose increases in C-starved winter soil

Žifčáková et al. Microbiome 2017

Genomics of microbial isolates to aid the analysis of ecosystem processes







García-Fraile et al. Appl. Environ. Microbiol. 2016
Lladó et al. Syst. Appl.. Microbiol. 2016

Dominant bacteria: Expression in forest soil and litter across seasons

Isolate	Putative ID	Phylogenetic group	Genes	Abundance DNA (%)	Abundance RNA (%)	Transcription rate (ppm)	% Genes expressed	Horizon specific	Seasonality litter	Seasonality soil
L105	Pedobacter sp.	Bacteroidetes	5752	0.02	0.06	27.1	5.7	**		
L196	Mucilaginibacter sp.	Bacteroidetes	4762	0.21	0.43	165.6	23.1	**		*
L21	Acidisphaera sp.	Alphaproteobacteria	6036	0.22	0.80	0.7	2.1			
L27	Burkholderia sp.	Betaproteobacteria	6012	0.12	0.40	25.7	25.0	**	**	**
L33	Xenophilus sp.	Betaproteobacteria	6648	0.07	0.24	7.5	15.2	**		*
L36	Rhodanobacter sp.	Gammaproteobacteria	3641	0.26	0.19	4.6	6.3	.		*
L45	Beijerinckia sp.	Alphaproteobacteria	5702	0.73	0.20	1.5	4.4	**		
L46	Granulicella sp.	Acidobacteria	4784	0.90	0.39	188.1	55.2	**	*	*
L56	Granulicella sp.	Acidobacteria	4114	1.23	0.46	7.3	11.4	*		**
L60	Granulicella sp.	Acidobacteria	5207	2.46	1.06	29.6	36.5	*		*
L85	Acidisoma sp.	Alphaproteobacteria	6275	0.24	0.27	28.4	48.1	**		
S1	Burkholderia sp.	Betaproteobacteria	9312	0.24	5.24	57.1	36.5	**	**	*
S103	Acidisphaera sp.	Alphaproteobacteria	7760	0.48	1.01	15.6	24.4	**		*
S129	Methylocapsa sp.	Alphaproteobacteria	7007	6.87	2.55	7.2	13.0	**		
S15	Silvibacterium sp.	Acidobacteria	5703	3.96	1.45	23.3	38.1	**	.	*
S156	Granulicella sp.	Acidobacteria	5387	1.54	0.67	34.9	50.5	**	.	*
S159	Acidisoma sp.	Alphaproteobacteria	6228	0.24	0.27	87.6	48.0	**		*
S171	Burkholderia sp.	Betaproteobacteria	8091	0.54	2.01	105.5	44.6	*	**	**
S184	Dyella sp.	Gammaproteobacteria	3711	0.44	0.26	21.8	38.5		*	**
S190	Granulicella sp.	Acidobacteria	4490	1.68	0.67	26.3	39.4	**		.
S45	Caulobacter sp.	Alphaproteobacteria	4412	0.25	0.26	1.6	6.2			
S55	Terracidophilus sp.	Acidobacteria	4617	3.46	1.24	41.1	31.8	**		
S69	Bradyrhizobium sp.	Alphaproteobacteria	7066	3.66	1.77	33.5	34.3	.	*	*
S8	Acidobacterium sp.	Acidobacteria	4961	6.47	2.05	13.4	29.3	**		.

Bacterial transcription differs between litter and soil.

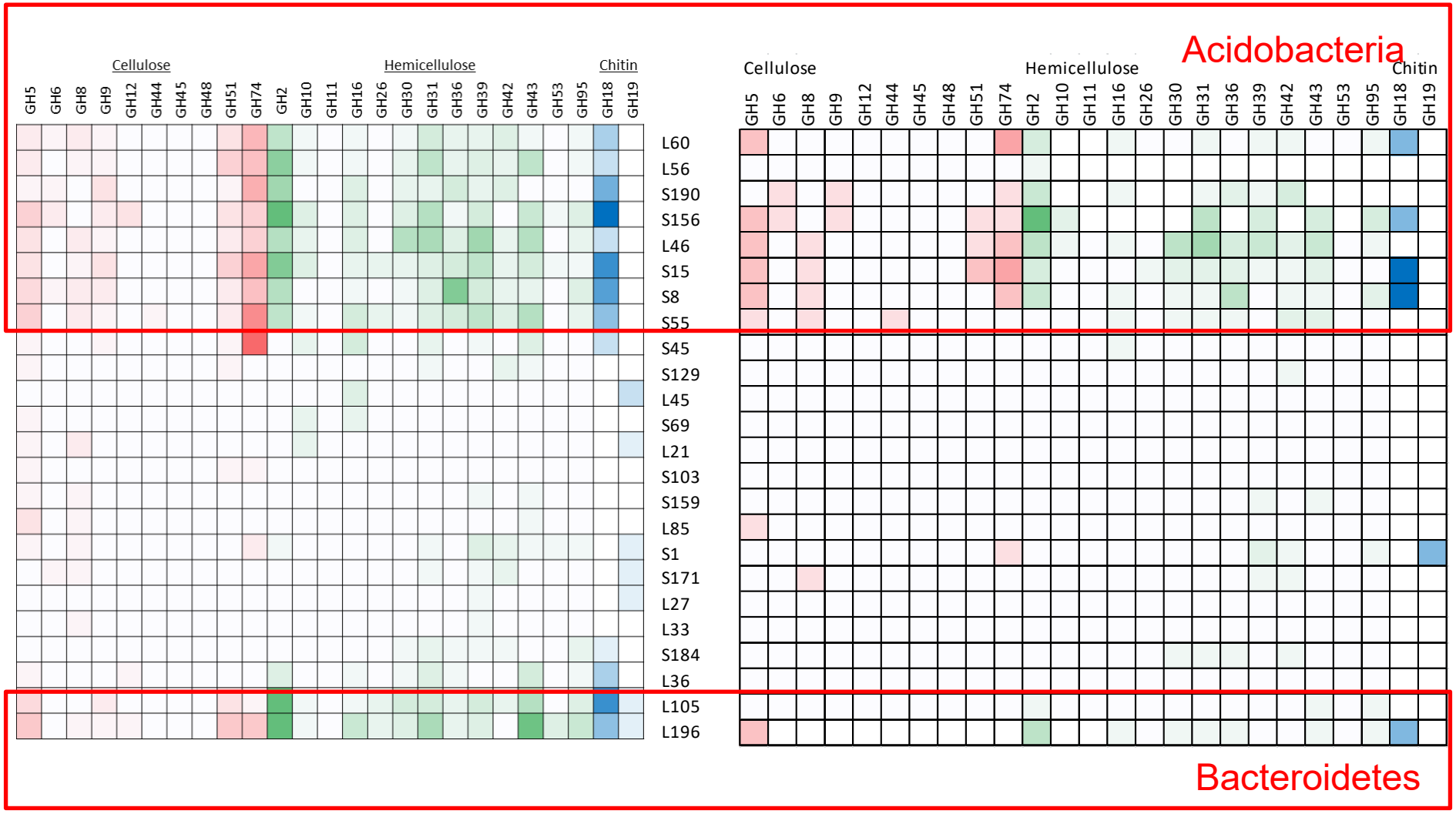
In soil, most taxa transcribe different genes in different seasons.

Lladó et al. Soil Biol. Biochem. 2019

Genomic potential – carbohydrate utilization

Genomic potential (gene counts)

Transcription in soil



Some Acidobacteria and Bacteroidetes with high genomic potential also highly express decomposition-related enzymes

Lladó et al. Soil Biol. Biochem. 2019

Data visualization and statistics: Examples from five selected papers

The ISME Journal (2018) 12:692–703
https://doi.org/10.1038/s41396-017-0027-3

ARTICLE

Clearcutting alters decomposition processes and initiates complex restructuring of fungal communities in soil and tree roots

Petr Kohout[1,2,3] · Markéta Charvátová[1] · Martina Štursová[1] · Tereza Mašínová[1] · Michal Tomšovský[4] · Petr Baldrian[1]

FEMS Microbiology Ecology, 93, 2017, fix157

doi: 10.1093/femsec/fix157
Advance Access Publication Date: 8 November 2017
Research Article

FEMS Microbiology Ecology, 92, 2016, fiw185

doi: 10.1093/femsec/fiw185
Advance Access Publication Date: 7 September 2016
Research Article

RESEARCH ARTICLE

Bacteria associated with decomposing dead wood in a natural temperate forest

Vojtěch Tláskal[1,*,†], Petra Zrůstová[1], Tomáš Vrška[2] and Petr Baldrian[1]

RESEARCH ARTICLE

Small-scale spatial heterogeneity of ecosystem properties, microbial community composition and microbial activities in a temperate mountain forest soil

Martina Štursová[1,*], Jiří Bárta[2], Hana Šantrůčková[2] and Petr Baldrian[1]

Received: 19 April 2017 | Revised: 4 September 2017 | Accepted: 13 September 2017
DOI: 10.1002/ldr.2817

WILEY

RESEARCH ARTICLE

Development of microbial community during primary succession in areas degraded by mining activities

Lenka Harantová[1] | Ondřej Mudrák[2] | Petr Kohout[1,5,6] | Dana Elhottová[3] | Jan Frouz[4] | Petr Baldrian[1]

RESEARCH ARTICLE
Ecological and Evolutionary Science

Complementary Roles of Wood-Inhabiting Fungi and Bacteria Facilitate Deadwood Decomposition

Vojtěch Tláskal,[a,b] Vendula Brabcová,[a] Tomáš Větrovský,[a] Mayuko Jomura,[c] Rubén López-Mondéjar,[a] Lummy Maria Oliveira Monteiro,[d] João Pedro Saraiva,[d] Zander Rainier Human,[a] Tomáš Cajthaml,[a] Ulisses Nunes da Rocha,[d] Petr Baldrian[a]

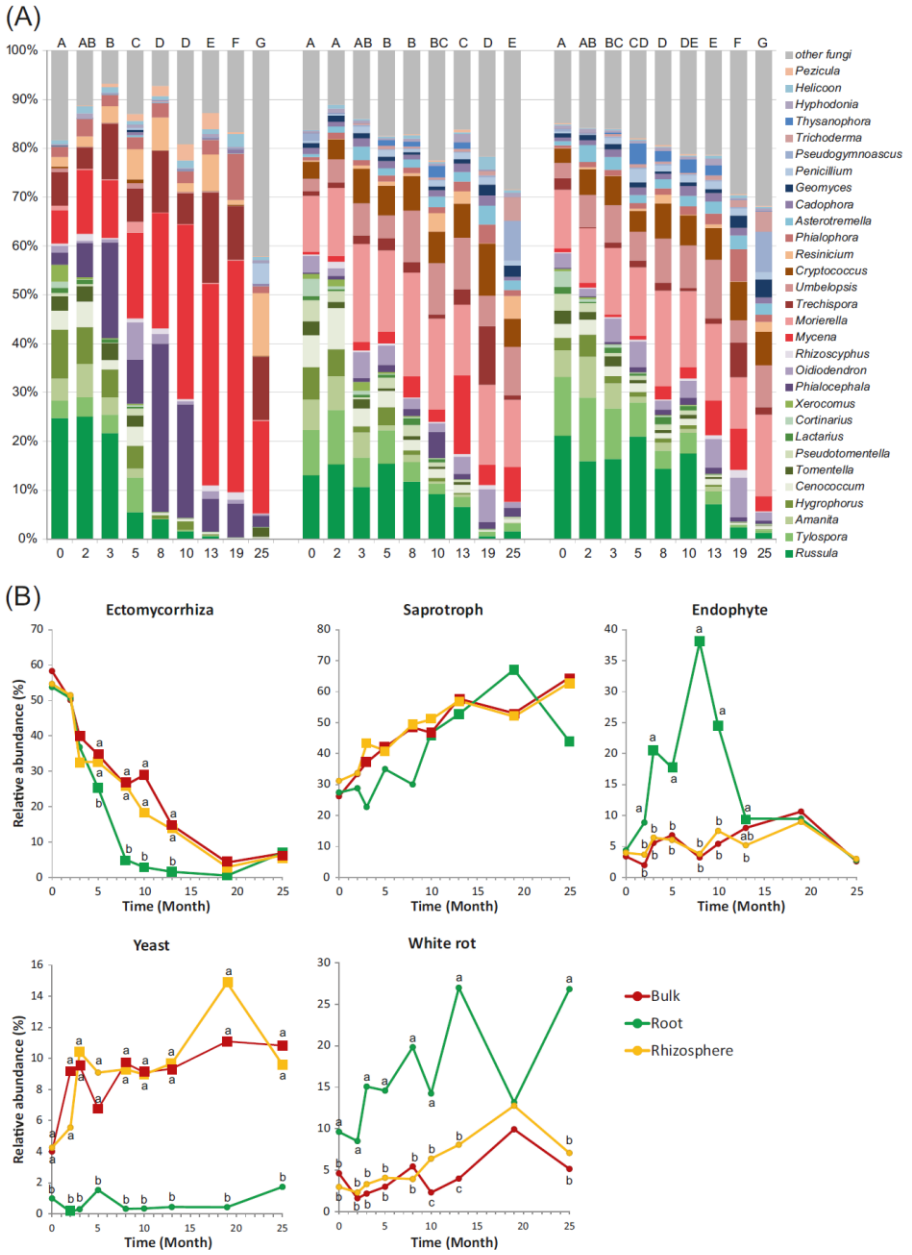


Fig. 4 Abundances of major fungal genera and functional groups in the soil, rhizosphere and roots of a *Picea abies* stand before (month 0) and after clearcutting. Values represent means ($n = 10$) of the relative abundances of ITS2 amplicons expressed as percentages. **a** Abundance of major fungal genera. Different letters indicate significant differences between timepoints. **b** Abundance of functional groups. Different letters indicate significant differences in abundance between soil, rhizosphere and root, and square symbols indicate significantly different abundance compared to the *P. abies* stand on month 0

Barplot shows mean community composition across timepoints (months of sampling).

Letters above the columns indicate, what timepoints differ significantly.

Line graph is another option for the same: here the abundance of fungal ecological guilds.

Here, letters indicate which treatments are different at certain point in time.

Kohout et al.

Table 1 Factors affecting the composition of fungal communities in three different compartments as revealed by PERMANOVA

	Root			Rhizosphere soil			Bulk soil		
	Df	*P* value	Adj. R^2	Df	*P* value	Adj. R^2	Df	*P* value	Adj. R^2
Fungal OTUs									
Spatial distribution	4	**0.001**	**0.029**	5	**0.001**	**0.089**	5	**0.001**	**0.107**
Time after clearcutting	1	**0.001**	**0.128**	1	**0.001**	**0.1**	1	**0.001**	**0.103**
Water content	NS	NS	NS	1	**0.005**	**0.005**	1	**0.004**	**0.008**
Ecological guild									
Spatial distribution	NS	NS	NS	NS	NS	NS	1	**0.006**	**0.021**
Time after clearcutting	1	**0.001**	**0.232**	1	**0.001**	**0.298**	1	**0.001**	**0.399**
Water content	NS	NS	NS	NS	NS	NS	NS	NS	NS

Factors significantly affecting composition of the fungal communities on the taxon level (relative sequence abundance of OTUs) and on the ecological guild level (relative sequence abundance of ecological guilds) are in bold. Adjusted R^2 was determined only for statistically significant factors

PERMANOVA permutational multivariate analysis of variance, *Df* degree of freedom, *NS* not significant, *OTU* Operational Taxonomic Unit

PERMANOVA was used to explore which effects were significant. Was it spatial location of samples, time, or water content in samples?

Kohout et al.

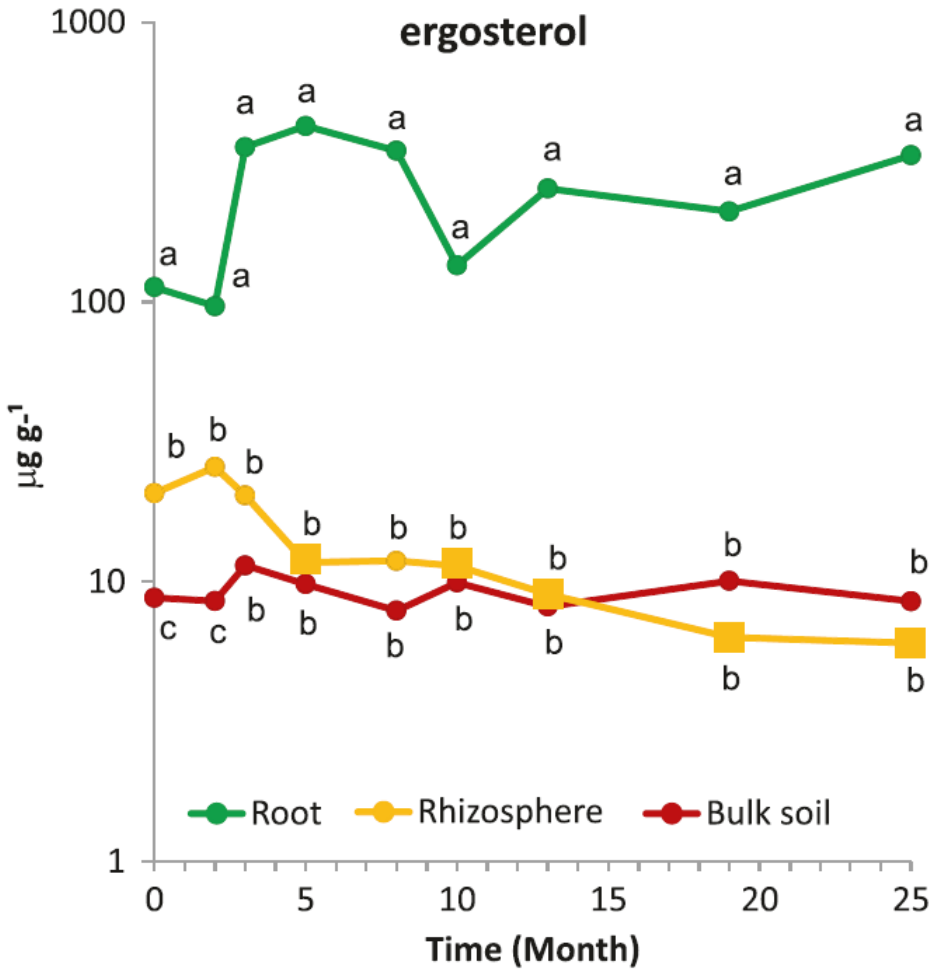


Fig. 2 Fungal biomass in the soil, rhizosphere and roots of a *Picea abies* stand before (month 0) and after clearcutting expressed in µg ergosterol g^{-1} of dry soil or roots. Values represent means ($n = 10$) of the ergosterol content. Different letters indicate significant differences in ergosterol content between soil, rhizosphere and roots, and squares indicate significantly different ergosterol content compared to the *P. abies* stand in month 0

Some data (e.g. environmental variables such as pH or biomass content) show normal distribution and testing of treatment differences can be performed using ANOVA which is more sensitive.

Kohout et al.

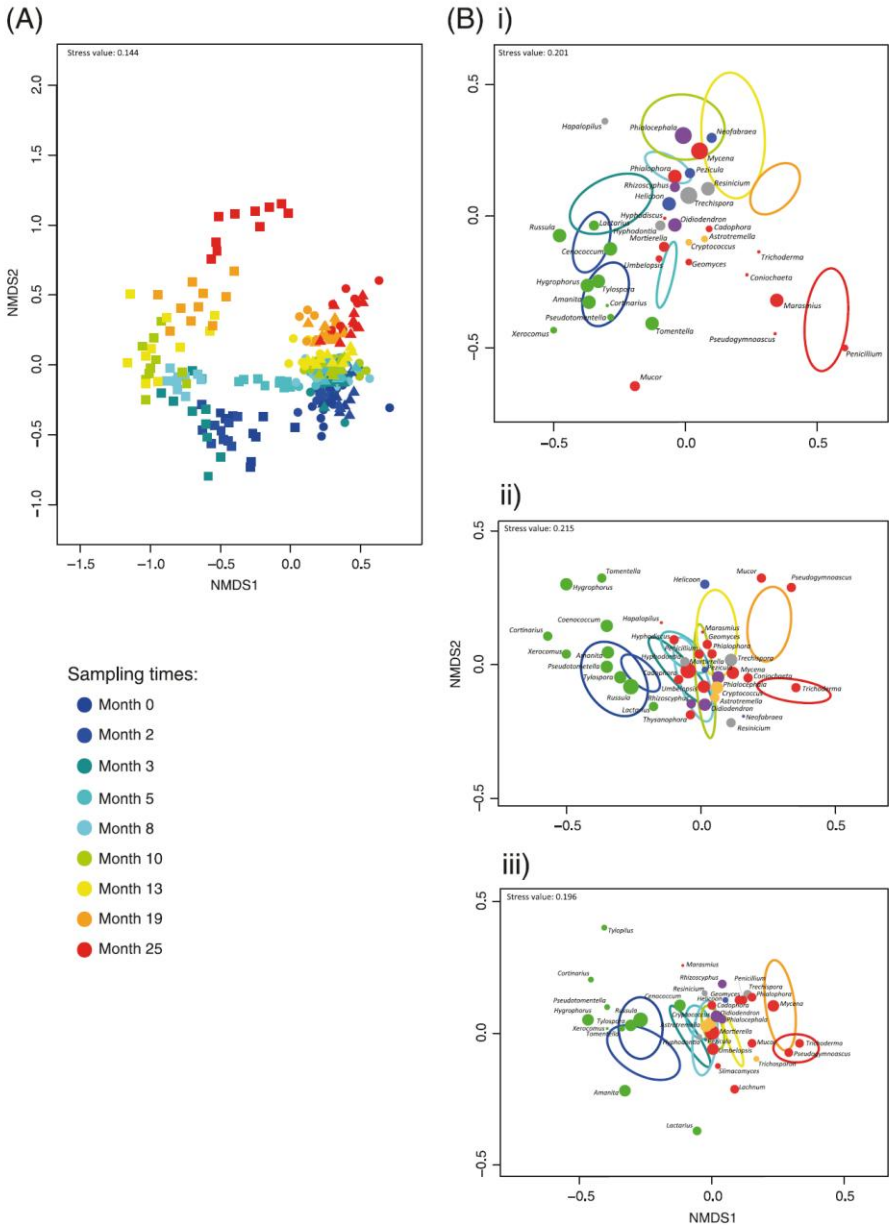


Fig. 3 **a** Non-metric multidimensional scaling (NMDS) plot of the community composition of fungal genera. Each symbol represents one root (square), rhizosphere (triangle) or soil (circle) sample. Time after clearcutting is indicated by colour. **b** NMDS plots of the composition of fungal genera communities for three studied compartments: (i) root, (ii) rhizosphere and (iii) soil. Ellipses represent ordination confidence intervals (95%) for different sampling times after clearcutting. The relative abundances of fungal genera are expressed by the size of symbols, and putative fungal ecophysiology is indicated by colours: EcMF—green, endophytic fungi (ENF)—violet, saprotrophs—red, plant pathogens—blue, white-rot—grey and yeast—orange

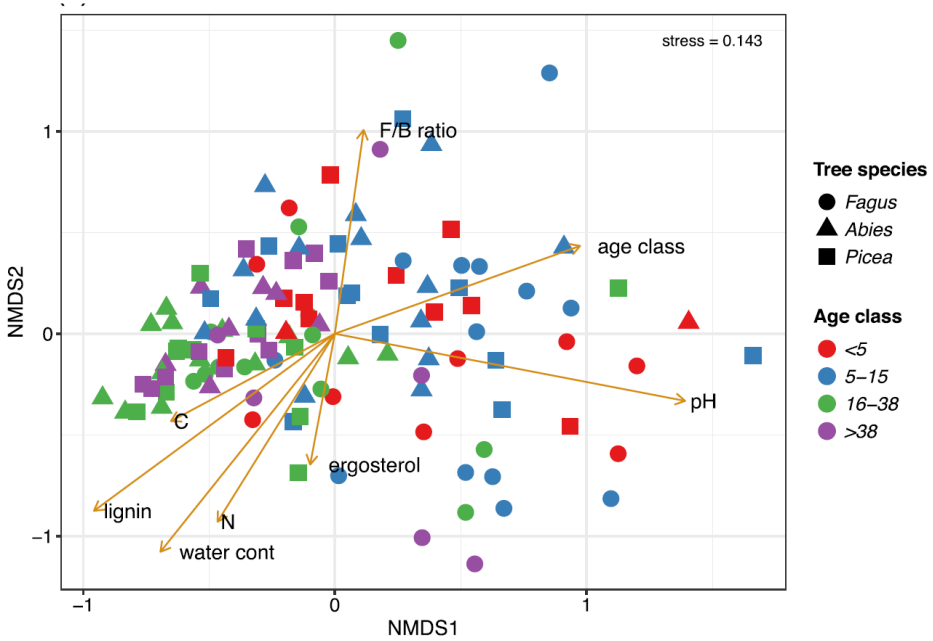
Non-metric multidimensional scaling (NMDS) can visualize effects of multiple treatments on sample similarity.

NMDS does not require normality of data, but can not quantify the share of explained variability.

a) *Samples are labelled by treatment (time = colour, symbol = habitat)*

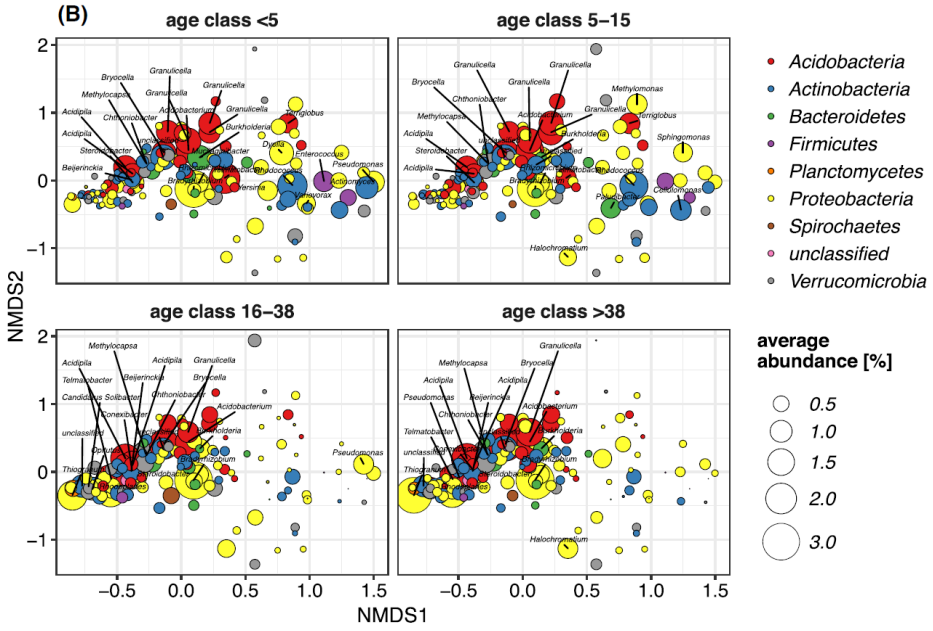
b) *Since habitat was found to affect samples most, subsequent NMDS was performed separately for each habitat to show time effect (colours and coloured elipses).*

Kohout et al.



NMDS can map vectors into its plot and test if they significantly associate with the x and y axes. This is some indication that they affect sample composition

Upper panel shows vectors of environmental variables that are significantly linked with axes



While NMDS typically shows sample distribution, individual species can be also mapped showing their association with axes. Each circle is one fungal taxon, size corresponds to relative abundance at certain treatment (here age class of deadwood). Colours show higher bacterial taxa (phyla).

Figure 4. Non-metric multidimensional scaling of bacterial communities in the CWD of *Fagus sylvatica*, *Abies alba* and *Picea abies* in the Žofín natural forest. (A) Plot of individual CWD samples. Vectors of environmental variables are included. (B) Plot of dominant bacterial taxa and their average abundances (circle size) across CWD of various age classes.

Tláskal et al. 2017

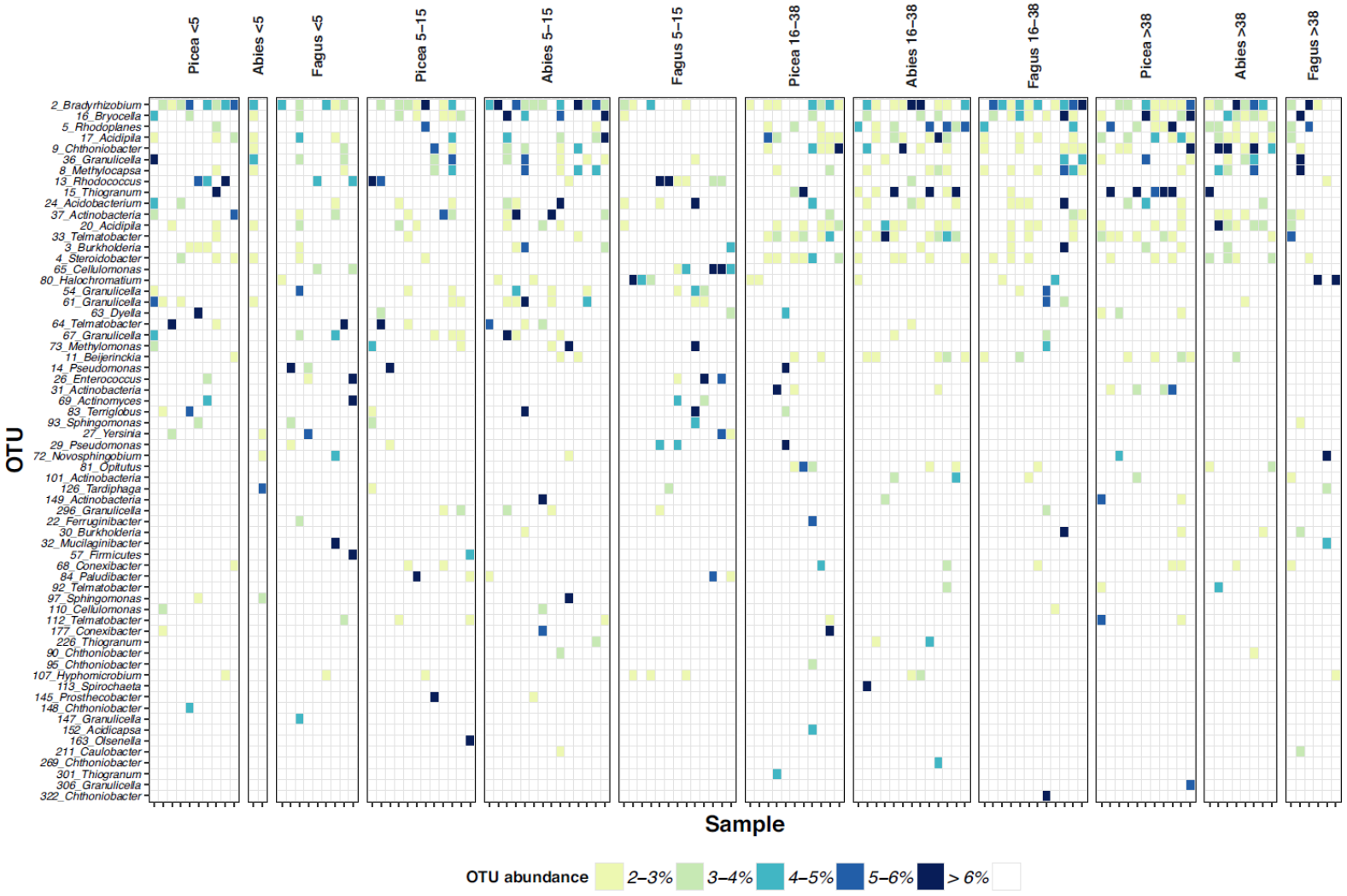


Figure 3. Distribution of abundant OTUs in individual CWD of *Fagus sylvatica*, *Abies alba* and *Picea abies* in the Žofín natural forest. Colours of points indicate the relative abundances of each OTU in each CWD sample (x axis). OTUs are specified by their number and the genus of the best hit on the y axis.

Heatmaps can display more values in space (here taxon relative abundances across samples), but values may be more difficult to read than in a bar plot.

Tláskal et al. 2017

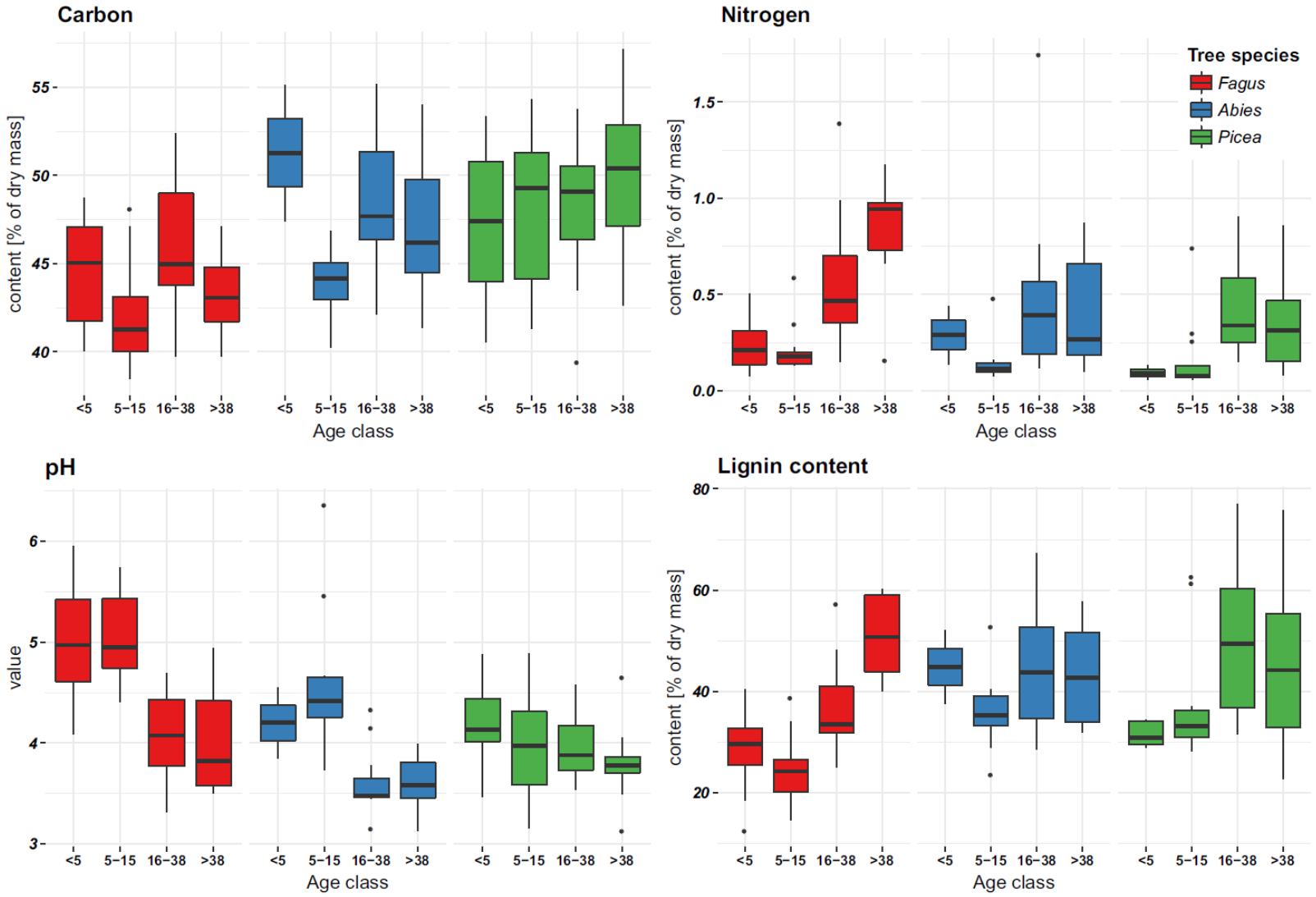


Figure 1. Chemistry of the CWD of *Fagus sylvatica*, *Abies alba* and *Picea abies* in the Žofín natural forest. Data on C and N content and pH are from (Baldrian *et al.* 2016).

Box-and-whisker plots show not only the averages, but the variation within treatment. The wider the plot, the higher variability.

Tláskal et al. 2017

Table 1. Bacterial indicator OTUs associated with CWD age classes in a natural beech-dominated forest. Abundance values are means for particular age class.

Age class	OTU	Best hit	Specificity	Fidelity	*P* value	Abundance (%)			
						<5	5–15	16–38	>38
<5	27	*Yersinia (Gammaproteobacteria)*	0.62	0.86	0.0013**	0.79	0.43	0.01	0.05
	63	*Dyella (Gammaproteobacteria)*	0.54	0.90	0.0155*	1.21	0.35	0.17	0.47
	64	*Telmatobacter (Acidobacteria)*	0.64	0.76	0.0108*	1.98	0.86	0.20	0.06
	52	*Variovorax (Betaproteobacteria)*	0.46	1.00	0.0006***	0.67	0.48	0.14	0.17
	13	*Rhodococcus (Actinobacteria)*	0.42	1.00	0.0306*	1.84	1.67	0.46	0.42
	333	*Beijerinckia (Alphaproteobacteria)*	0.50	0.81	0.0022**	0.18	0.16	0.01	0.01
	93	*Sphingomonas (Alphaproteobacteria)*	0.45	0.90	0.0416*	0.48	0.39	0.02	0.17
	167	*Mycobacterium (Actinobacteria)*	0.45	0.90	0.0203*	0.53	0.39	0.09	0.17
	126	*Tardiphaga (Alphaproteobacteria)*	0.44	0.90	0.0371*	0.49	0.38	0.05	0.19
	205	*Sphingomonas (Alphaproteobacteria)*	0.46	0.81	0.0171*	0.27	0.23	0.02	0.05
	419	*Candidatus Xiphinematobacter (Verrucomicrobia)*	0.60	0.57	0.0043**	0.29	0.15	0.03	0.01
	188	*Opitutus (Verrucomicrobia)*	0.45	0.62	0.0456*	0.44	0.26	0.13	0.13
	57	*unclassified (Firmicutes)*	0.68	0.38	0.0213*	0.53	0.25	0.00	0.00

Indicator species analysis can identify which microbial taxa are associated with certain treatment (here deadwood younger than 5 years). The lower the P-value, the higher the specificity of the taxon for the treatment.

Tláskal et al. 2017

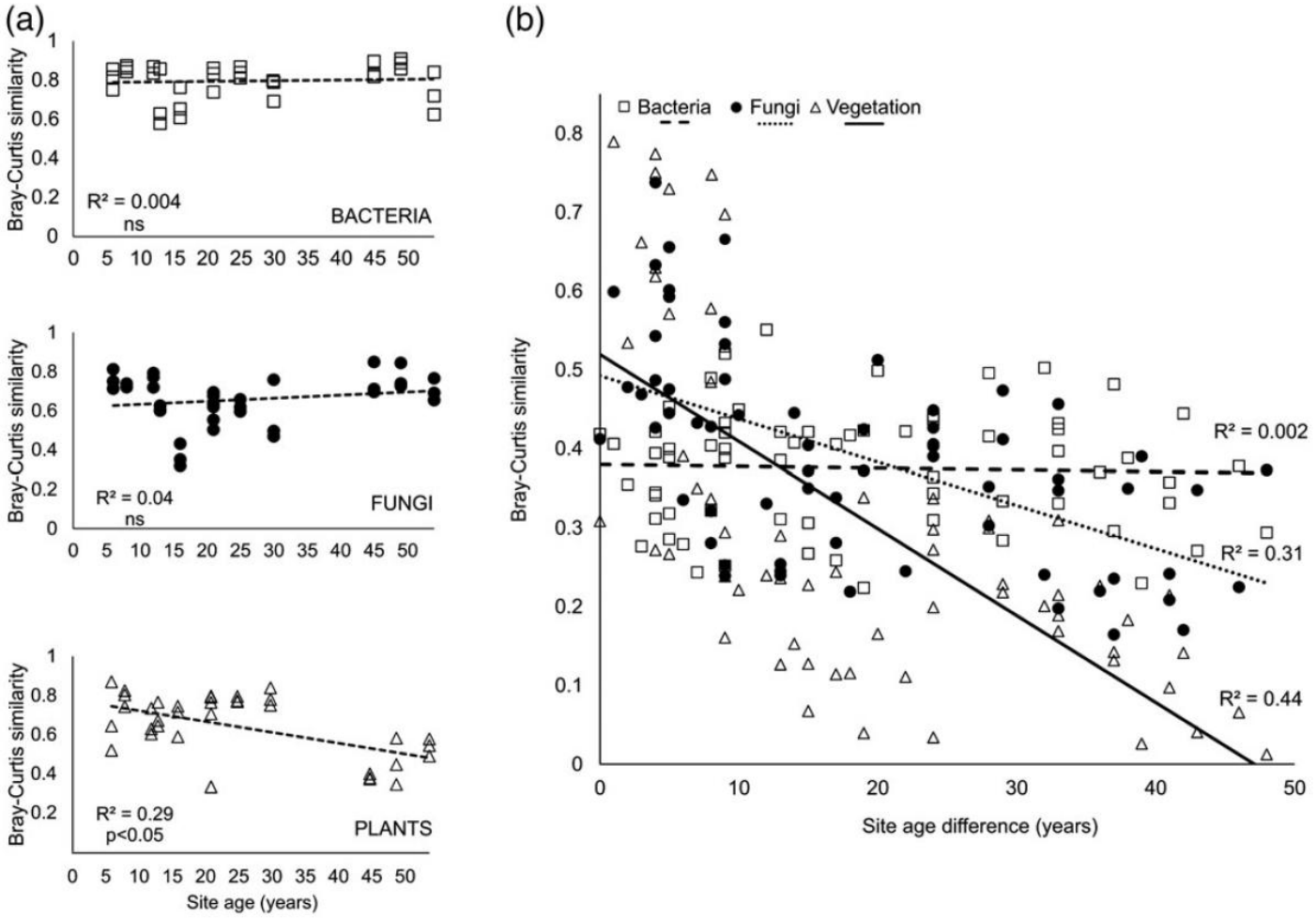


FIGURE 4 Time-dependent variations of Bray–Curtis similarity of bacterial, fungal, and plant communities (a) within sampling sites of the same plot and age and (b) between different plots representing the chronosequence

Linear correlations indicate the dependence between two numerical variables.

Correlations may be highly affected by outlier points.

Harantová et al.

TABLE 3 Correlations between environmental variables and the composition of bacterial, fungal, and plant communities at postmining sites of different successional ages (Mantel test on Bray–Curtis similarities among pairs of samples, n = 9,999, significant values in bold ($p<0.05$))

	Age		pH		N_{tot}		C_{org}		Vegetation	
	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p
Mantel test										
Bacteria	**.37**	**<.001**	**.24**	**<.001**	.12	.07	.04	.24	**.47**	**<.001**
Fungi	**.54**	**<.001**	**.28**	**<.001**	**.22**	**<.001**	**.29**	**<.001**	**.57**	**<.001**
Vegetation	**.67**	**<.001**	**.28**	**<.001**	**.30**	**<.001**	**.32**	**<.001**		
Partial Mantel test corrected for site age										
Bacteria			**.13**	**<.05**	.04	.26	.00	.43	**.32**	**<.001**
Fungi			**.13**	**<.01**	**.12**	**<.05**	**.27**	**<.001**	**.34**	**<.001**
Vegetation			.06	.08	**.21**	**<.01**	**.33**	**<.01**		

Mantel tests are used to make correlations between matrices, such as, in this case, between the matrix of community fungal or bacterial composition (OTU table) and matrix of plant community composition (list of observed plant species).

Mantel tests can be also used to correlate matrix and variable (e.g. pH).

The question asked is, for example, whether community composition of fungi is more similar in samples with more similar plant communities.

Harantová et al.

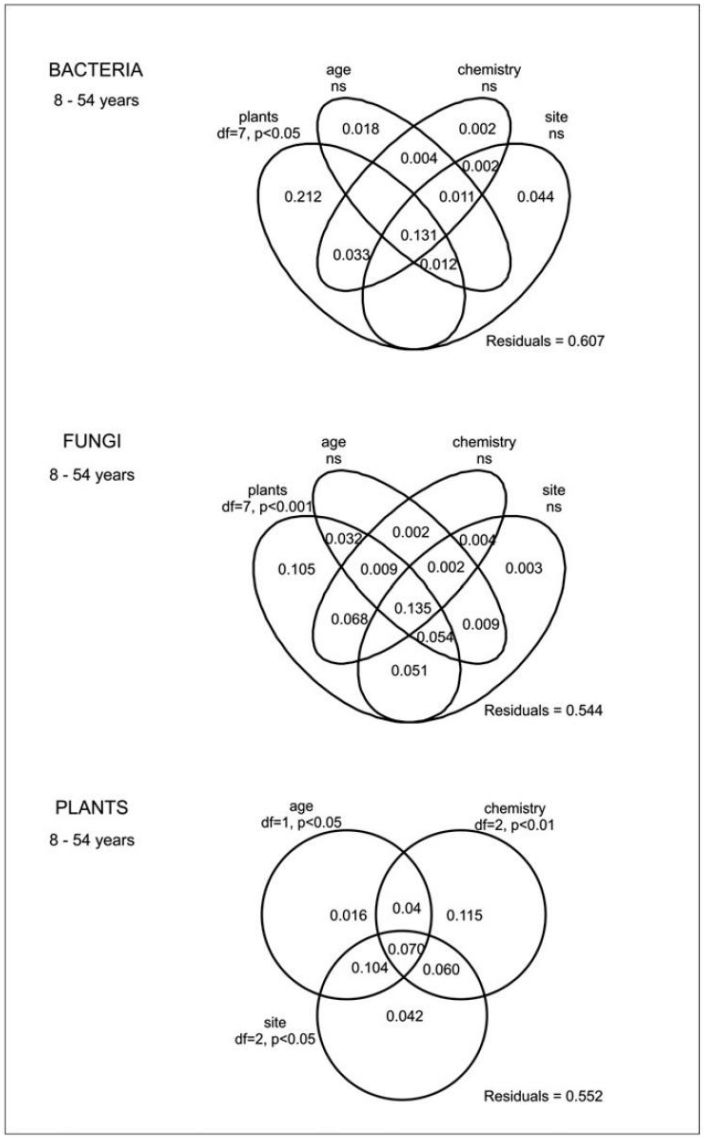


FIGURE 5 Drivers of the vegetation and microbial community composition according to the variation partitioning analysis showing the effects of site age, soil chemistry, site identity, and plants (only for bacteria and fungi). Numbers within individual compartments indicate the percentage of explained variation. *df* = degrees of freedom

Variation partitioning (VP) is able to divide total variability between factors that determine it.

Here, effects of plant community composition, site age, site chemistry and site location were analysed and variation partitioning quantified pure effects of individual components and their significance.

VP also identifies effects shared between variables, but the significance of this shared variability can not be tested.

Harantová et al.

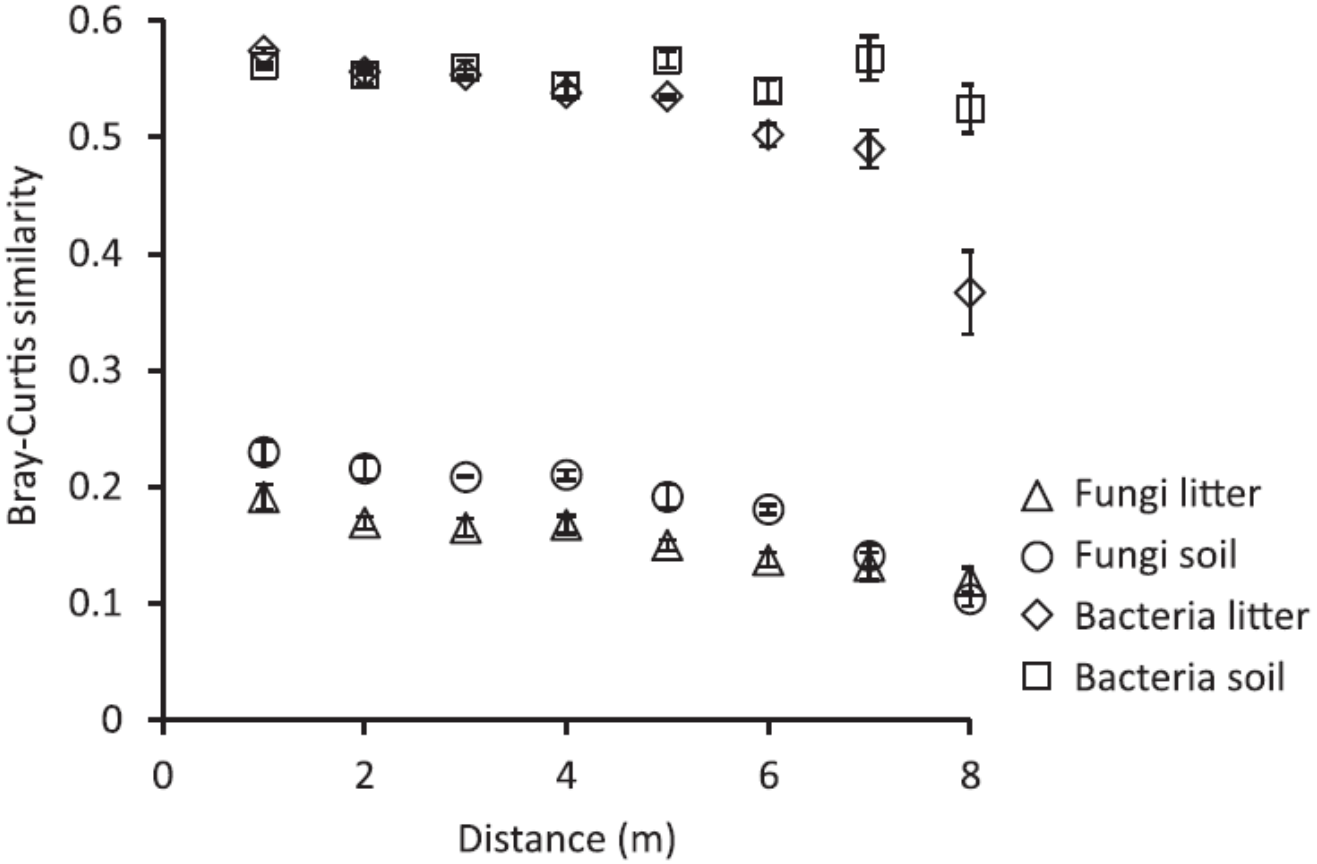


Figure 4. Relationship between the pairwise sample distance and the similarity of microbial communities in the regenerating mountainous forest.

Simple application of spatial statistics is the correlation between sample distance and similarity.
Especially in soil fungi, communities in closely located samples were more similar.

Šťursová et al.

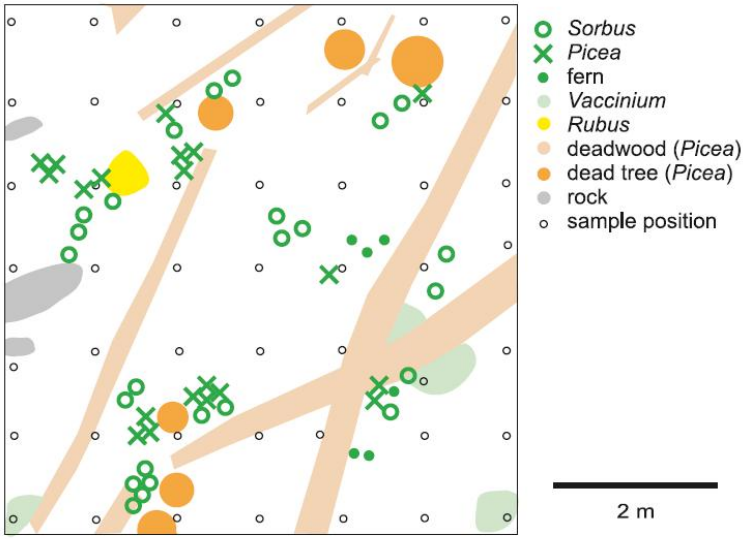


Figure 1. Study area within a regenerating mountainous forest, soil surface features and locations of the sampling sites in the studied plot.

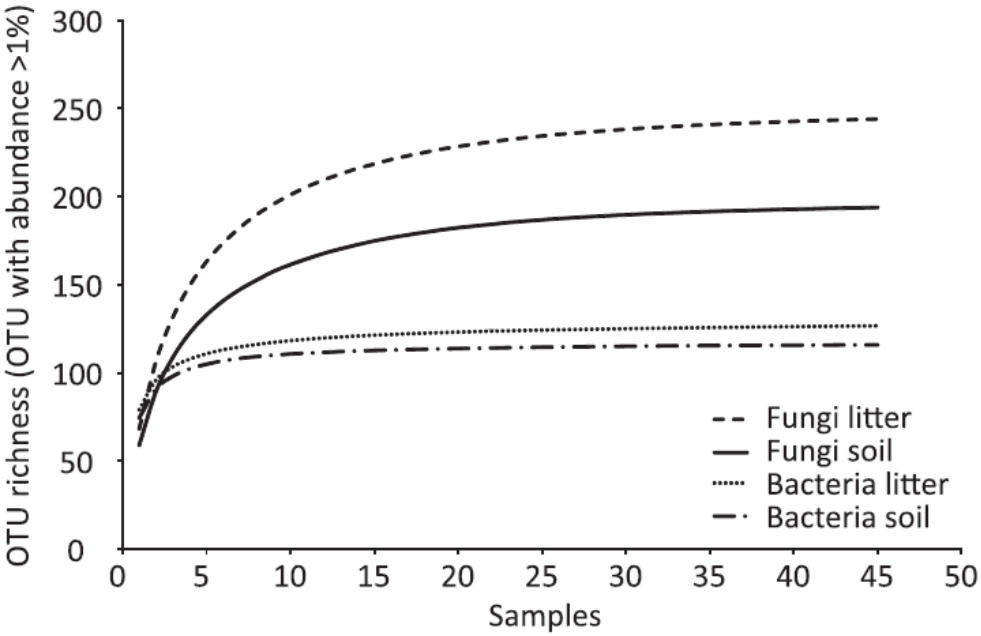


Figure 3. Beta-diversity of the abundant members of the bacterial and fungal communities in the regenerating mountainous forest. The data represent the relationship between the sampling effort (number of sites) and the estimates of total OTU richness of the taxa, for OTUs whose abundance was >1% in at least

While alpha-diversity expresses the number of species in sample, beta diversity indicates how much diversity increases with adding additional samples.
Here, individual soil cores were analysed that were collected at 1-m distances.

Štursová et al.

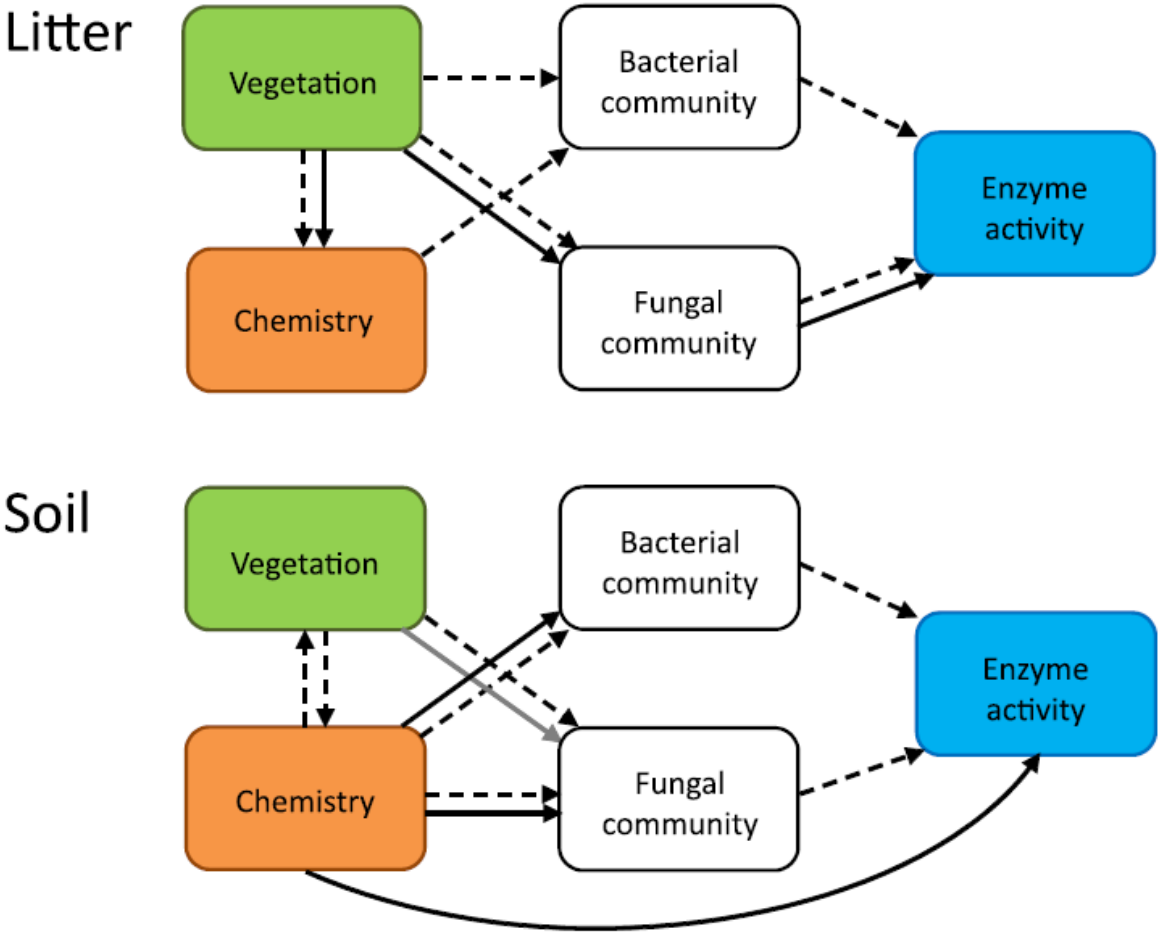


Figure 2. Hypothetical and observed effects of soil and vegetation properties on the bacterial and fungal communities in a regenerating mountainous forest and their relationship with enzymatic activity. Potential effects are indicated with dashed lines, and observed effects are indicated with continuous lines; a grey line indicates a marginally significant effect.

In statistical models, significant interactions can indicate what may be the potential drivers of observed processes.

To analyse the causes and consequences, one can use Structural Equation Modelling (SEM).

However, SEM needs much higher replication (n=60) than used in this study.

Štursová et al.

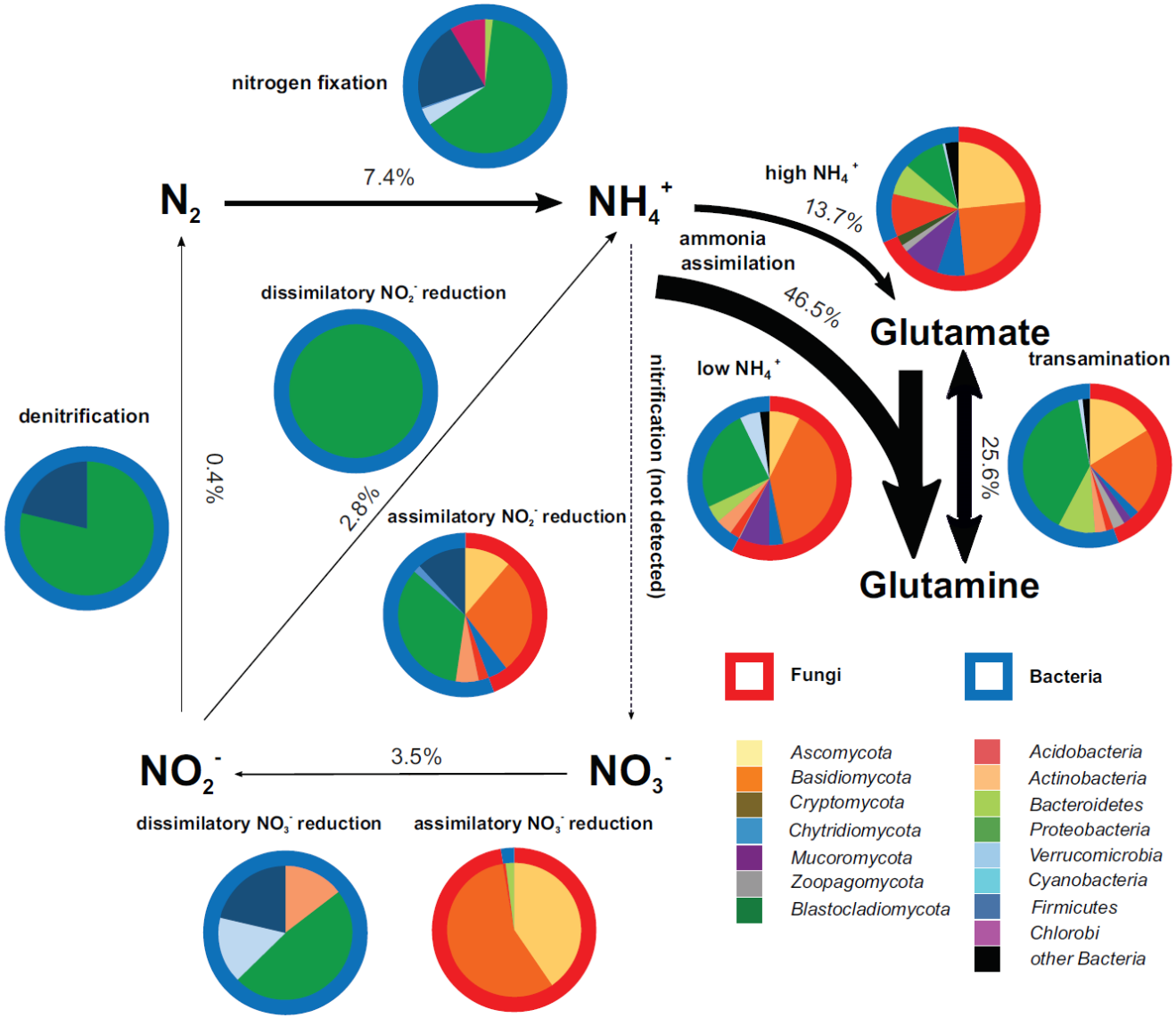


FIG 4 Nitrogen cycle and ammonia assimilation pathways occurring in deadwood. N-cycling intensity is expressed as the relative share of expression of each function across all N-cycling genes. Pie charts indicate the relative shares of bacterial and fungal transcription in each process. Nitrification was not detected.

Complex plots are sometimes required to represent proces rates and their microbial drivers alongside.

Function: Nitrogen cycling rates as thickness of arrows

Phylogeny: Involvement of Microbiome members in transcriptions in pie-charts

Tláskal et al. 2021

Environmental microbiology, basic methodical approaches