Bioinformatics for Microbiomes:

Methods in Microbial Ecology
Experimental Design
Data Presentation

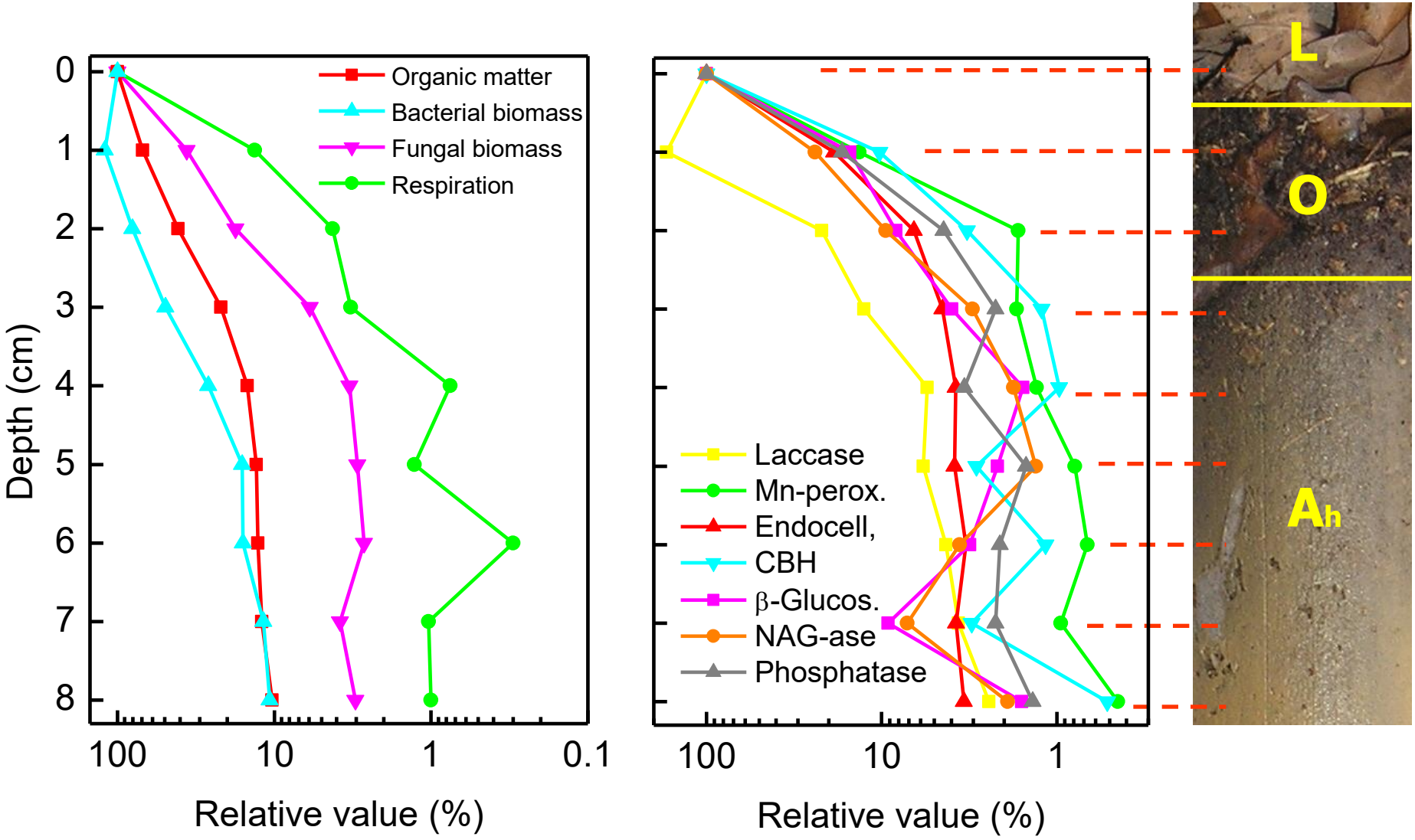
Petr Baldrian

Institute of Microbiology of the Czech Academy of Sciences
baldrian@biomed.cas.cz

Environment and microorganisms

Microorganisms sense the environment at scales different from those of the macroorganisms.
Natural environment is highly complex and heterogeneous at such scales.

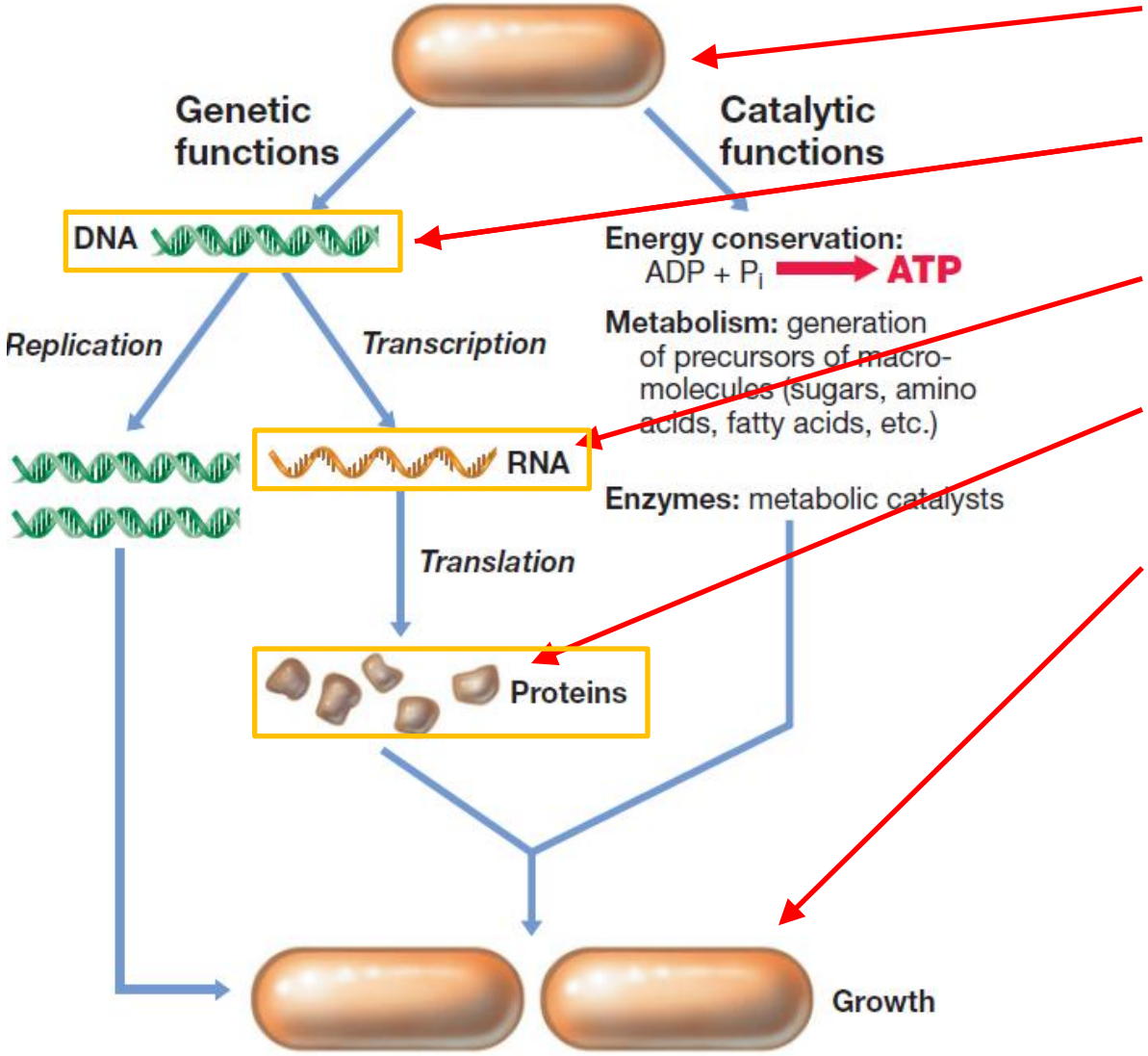
Spatial heterogeneity



Soil properties along a vertical profile of forest topsoil.

Šnajdr et al. Soil Biol Biochem 2008

Methodical approaches of microbiology



Brock Biology of Microorganisms (2012)

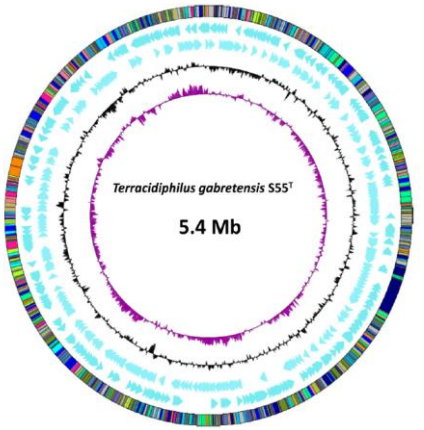
Classical microbiology

Genomics

Transcriptomics

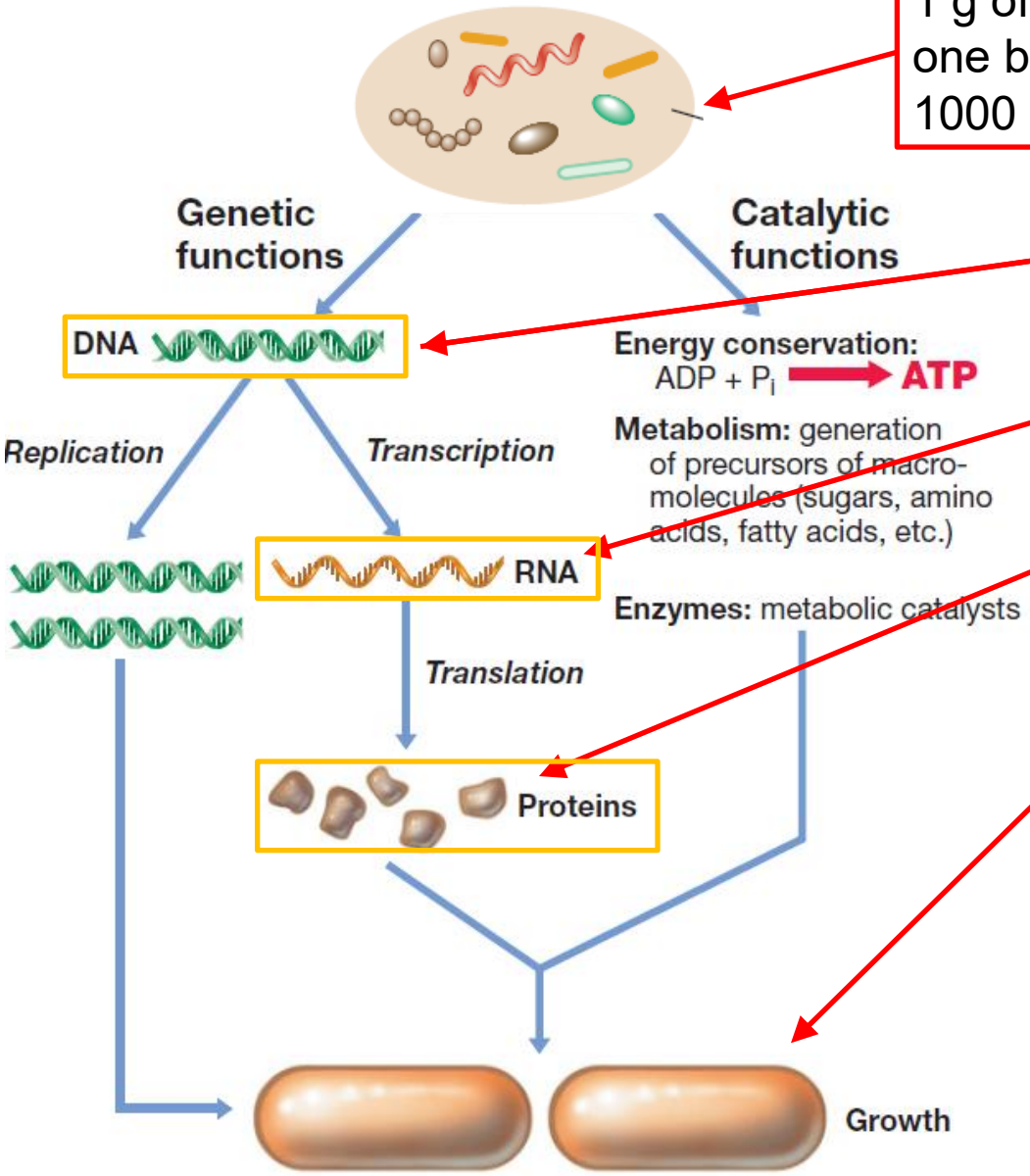
Proteomics

Metabolomics
Strain characterization



Methodical approaches of microbial ecology

1 g of soil:
one billion bacterial cells
1000 bacterial / 200 fungal species



Brock Biology of Microorganisms (2012)

Metagenomics

Metatranscriptomics

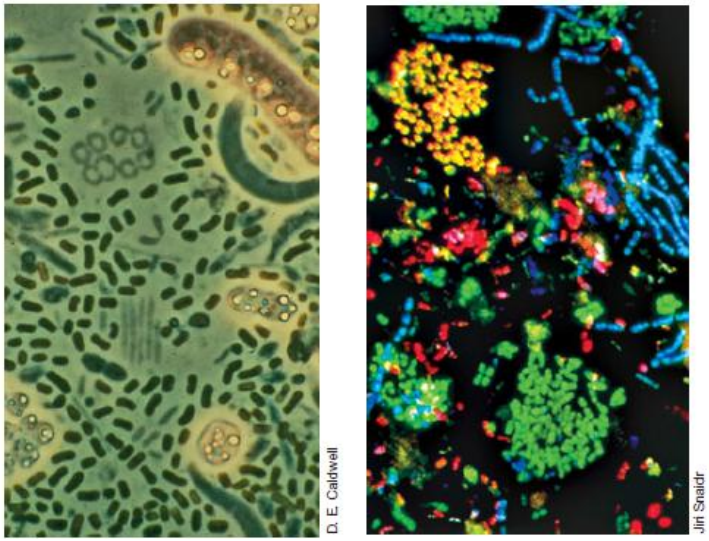
Metaproteomics

(Meta)metabolomics

>IEH00WJ01BK2Y9 xy=533_195
ATTAGCAGTAAGATCCGTAGGTGAACCTGCGGAAGGATCATTATCGAAACGAATAGGAATGGGGGAGCAAGACGGCAAAGC
AAAAGACTGTCGCTGGCTAGATTCATTTCTGGCATGTGCACGTCCTTGCTTTTTTCGTCGACCTTTCTCTTTCTTTCTTTCACA
CCTGTGCACCCGTTGTAGGTCCTCGAAAGAGGATC
>IEH00WJ01DVNL6 xy=1473_2492
TAGTGTAGATAGATCCGTAGGTGAACCTGCGGAAGGATCATTATCGAATCGTAGGCGAGGGTTGTCGCTGGCCTCTCGGGG
CATGTGCACGCCCGAGCCCTTGAATCCCAAACACCACATGTGAACCCACCGTAGGCCTTCGGGCCTATGTCTTATCATATAA
TCTGAATGTCTAATAGAATGTAAACCCATTTCGTTGC
>IEH00WJ01CDN03 xy=858_2645
CGTATGGACAGATCCGTAGGTGAACCTGCGGAGGAGCATATCAGTAAGCGGAGGAGCATATCAATAAGCGGAGGAGCATAT
CAATAAGCGGAGGAATCCGTAGTGAACCTGCGGAGGAGCATATCAATAAGCGGAGGAGCATATCAATAAGCGGAGGAGCAT
ATCAATAAGCGGAGGAGCATATCAATAAGCGGAGGAGT
>IEH00WJ01BNPAZ xy=562_3657
TAGTCGCATAAGATCCGTAGGTGAACCTGCGGAGGGATCATTACAAGAACGCCCGGGCTTCGGCCTGGTTATTCATAACCCT
TTGTTGTCCGACTCTCTTGCCTCCGGGGCGACCCTGCCTTCGGGCGGGGGCTCCGGGTGGACACTTCAAACTCTTGCGTAA
CTTTGCAGTCTGAGTAAACTTAATTAATAAATTAAAA
>IEH00WJ01B2AI9 xy=729_323
ATACGACGTAAGATCCGTAGGTGAACCTGCGGAAGGATCATTATTGAATACGTTTGGTTCTGATGCTGGCTCGTCACTGAGC
ATGTGCTCGATCCATAACTATTTATCTTCTCTTGTGCACCTTTTGTAGTCTTTCAGAGCAAGTGATAACTCTCGCAGCAATGCG
GTTTGGGGGACTTGGGCGTGAGCCCTTCCCCTTC
>IEH00WJ01DAD4P xy=1231_1655
TATCACTCAGAGATCCGTAGGTGAACCTGCGGAAGGATCATTACTGAAGTAGAGGGCCCCTGGGGTCCAACCTACCCACCC
GGTGTTTAATTGTAACCTTGTTGCTTCGGTGCGCCGCCTCACGGTCCGCCGGGGGGCTTCTGCCCCGGGTCCGCGCGCACC
GGTAGACACCATTGAACTTCTTGTTCTGAACGATTGCAC
>IEH00WJ01AEBMK xy=45_4042
CTATGTACAGAGATCCGTAGGTGAACCTGCGGAAGGATCATTACAGTGTTCTCTGCCCTCACGGGTAGAAACGCTCACCCTT
GTATATTATATCTTTGTTGCTTTGGCAGGCCGCCCTCGGGCACCGGCTCCGGCTGGATCGCGCCTGCCAGAGGAAACCCAA
ACTCTGAATGTTAGTGTCGTCTGAGTACTATCTAATA
>IEH00WJ01DFSEK xy=1292_3578
TAGTGTAGATAGATCCGTAGGTGAACCTGCGGAGGGATCATTACAGAGTTCATGCCCTTTAGGGTAGATCTCCCATCCTTTG
TCTATATACCTCTGTTGCTTTGGCAGGCCGAACTATTAGTCTACCGGCTCTGCTGGTAAGCGCCTGCCAGAGGACCCCCACT
CTGAGAGTTAGTGTCGTCTGAGTACTATATAATAGT
>IEH00WJ01CMQ2E xy=962_500
CAGTACTGCGAGATCCGTAGGTGAACCTGCGGAAGGATCATTATCGTAAACCGAGGTGCGAGGGCTGTCGCTGACCTTTTT
TGGTCGTGCACGCCCGAGCGCTCTCACACAATCCATCTCACCCCTTGTGCACCACCGCGTGGGTTCCCTTTCTGGCTTGTC
CGAAGGGGGCTCGCGTTTTCACACAAACTTGAATTGGT

Methodical approaches of microbial ecology

Cultivation-dependent methods



Brock Biology of Microorganisms (2012)

Direct visualization

```
>IEH00WJ01BK2Y9 xy=533_195
ATTAGCAGTAAGATCCGTAGGTGAACCTGCGGAAGGATCATTATCGAAACGAATAGGAATGGGGGAGCAAGACG
GCAAAGCAAAAGACTGTCGCTGGCTAGATTCATTTCTGGCATGTGCACGTCCTTGCTTTTTTCGTCGACCTTTCTC
TTTCTTTCTTTCACACCTGTGCACCCGTTGTAGGTCCTCGAAAGAGGATC
>IEH00WJ01DVNL6 xy=1473_2492
TAGTGTAGATAGATCCGTAGGTGAACCTGCGGAAGGATCATTATCGAATCGTAGGCGAGGGTTGTCGCTGGCCT
CTCGGGGCATGTGCACGCCCGAGCCCTTGAATCCCAAACACCACATGTGAACCCACCGTAGGCCTTCGGGCCTA
TGTCTTATCATATAATCTGAATGTCTAATAGAATGTAAACCCATTTCGTTGC
>IEH00WJ01CDN03 xy=858_2645
CGTATGGACAGATCCGTAGGTGAACCTGCGGAGGAGCATATCAGTAAGCGGAGGAGCATATCAATAAGCGGAGG
AGCATATCAATAAGCGGAGGAATCCGTAGTGAACCTGCGGAGGAGCATATCAATAAGCGGAGGAGCATATCAATA
AGCGGAGGAGCATATCAATAAGCGGAGGAGCATATCAATAAGCGGAGGAGT
>IEH00WJ01BNPAZ xy=562_3657
TAGTCGCATAAGATCCGTAGGTGAACCTGCGGAGGGATCATTACAAGAACGCCCGGGCTTCGGCCTGGTTATTC
ATAACCCTTTGTTGTCCGACTCTCTTGCCTCCGGGGCGACCCTGCCTTCGGGCGGGGGCTCCGGGTGGACACTT
CAAACTCTTGCGTAACTTTGCAGTCTGAGTAAACTTAATTAATAAATTAAAA
>IEH00WJ01B2AI9 xy=729_323
ATACGACGTAAGATCCGTAGGTGAACCTGCGGAAGGATCATTATTGAATACGTTTGGTTCTGATGCTGGCTCGTC
ACTGAGCATGTGCTCGATCCATAACTATTTATCTTCTCTTGTGCACCTTTTGTAGTCTTTCAGAGCAAGTGATAACT
CTCGCAGCAATGCGGTTTGGGGGACTTGGGCGTGAGCCCTTCCCCTTC
>IEH00WJ01DAD4P xy=1231_1655
TATCACTCAGAGATCCGTAGGTGAACCTGCGGAAGGATCATTACTGAAGTAGAGGGCCCCTGGGGTCCAACCTA
CCCACCCGGTGTTTAATTGTAACCTTGTTGCTTCGGTGCGCCGCCTCACGGTCCGCCGGGGGCTTCTGCCCCGG
GTCCGCGCGCACCGGTAGACACCATTGAACTTCTTGTTCTGAACGATTGCAC
>IEH00WJ01AEBMK xy=45_4042
CTATGTACAGAGATCCGTAGGTGAACCTGCGGAAGGATCATTACAGTGTTCTCTGCCCTCACGGGTAGAAACGCT
CACCCTTGTATATTATATCTTTGTTGCTTTGGCAGGCCGCCCTCGGGCACCGGCTCCGGCTGGATCGCGCCTGC
CAGAGGAAACCCAAACTCTGAATGTTAGTGTCGTCTGAGTACTATCTAATA
>IEH00WJ01DFSEK xy=1292_3578
TAGTGTAGATAGATCCGTAGGTGAACCTGCGGAGGGATCATTACAGAGTTCATGCCCTTTAGGGTAGATCTCCCA
TCCTTTGTCTATATACCTCTGTTGCTTTGGCAGGCCGAACTATTAGTCTACCGGCTCTGCTGGTAAGCGCCTGCCA
GAGGACCCCCACTCTGAGAGTTAGTGTCGTCTGAGTACTATATAATAGT
>IEH00WJ01CMQ2E xy=962_500
CAGTACTGCGAGATCCGTAGGTGAACCTGCGGAAGGATCATTATCGTAAACCGAGGTGCGAGGGCTGTCGCTGA
CCTTTTTTGGTCGTGCACGCCCGAGCGCTCTCACACAATCCATCTCACCCCTTGTGCACCACCGCGTGGGTTCCC
TTTCTGGCTTGTCCGAAGGGGGCTCGCGTTTTCACACAAACTTGAATTGGT
```

Analysis of nucleic acids

- comprises all microorganisms including those unculturable and unknown
- „species“ are formed artificially, based on mathematical similarity of sequences
- sequence identity (the gene and its producer) is identified indirectly, based on sequence similarity (higher or lower) to a known sequence

Cultured and uncultured microorganisms

There are some 10 000 culturable microbial species.
Some microbial phyla do not contain any culturable species

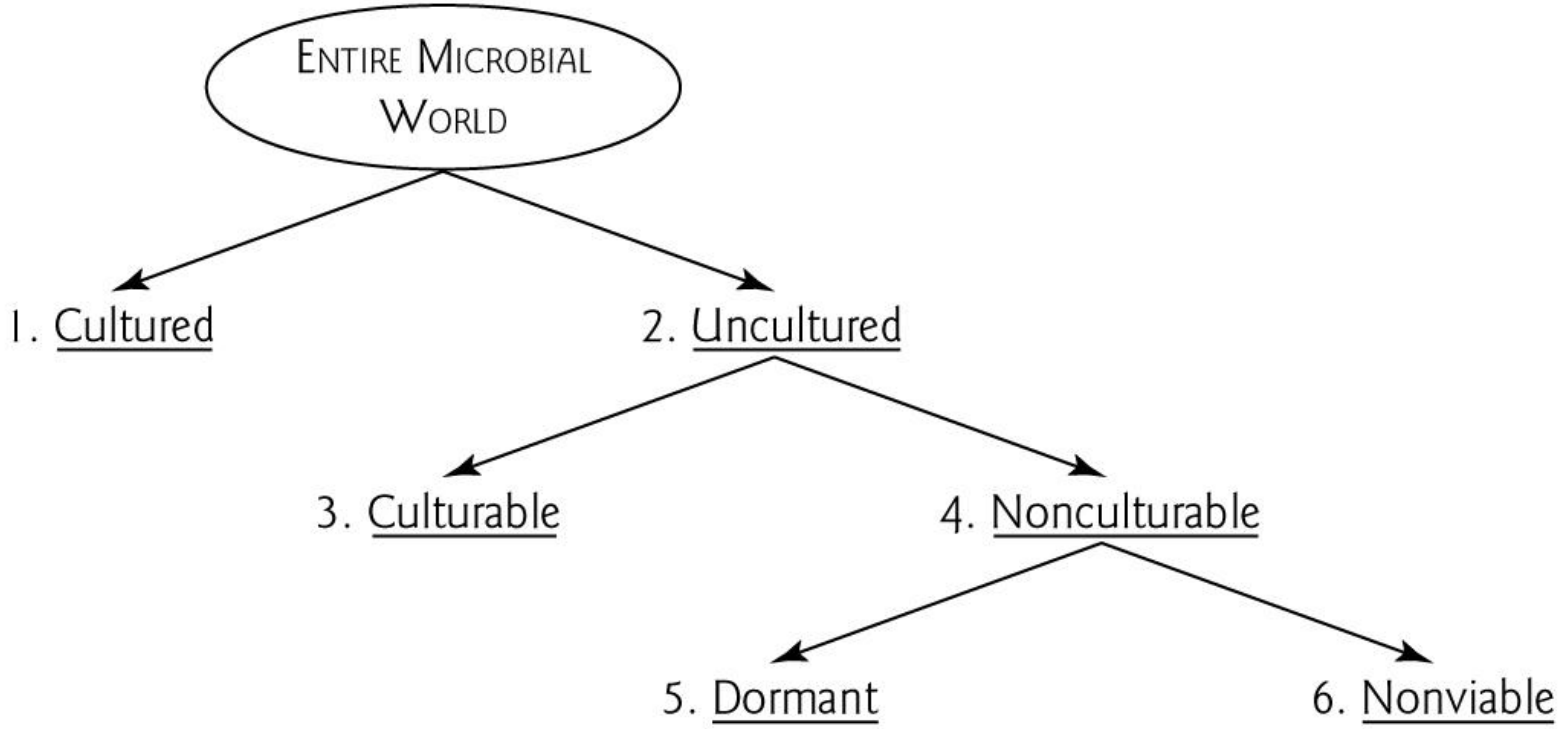


Figure 5.3 Six categories of culturability for microorganisms in nature. The categories are operationally defined – based on techniques of detection that include microscopy and both traditional and novel procedures for growth and isolation. See text for details.
Science and the citizen

Fundamental methodical approaches in microbial ecology

Box 5.1

Introducing the four fundamental methodological approaches in environmental microbiology

There are four fundamental methodological approaches that generate information addressing environmental microbiological questions and issues. These approaches are:

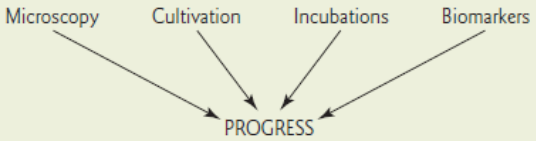
1 *Microscopy*: direct imaging at the microscale to verify the presence of microbial cells. Microscopy provides direct, irrefutable data on the total abundance of microorganisms in environmental samples. Recent technological advancements in the resolution and types of information gathered during imaging aim to identify individual cells and probe their biogeochemical activities.

2 *Cultivation*: environmental matrixes (e.g., soil, sediment, water) are diluted and suspended microorganisms are transferred to liquid or solid media where nutrients are provided – with hopes that the cells present will grow. Assays for assessing growth include colony formation (on agar plates), increased cell numbers in liquid media, and chemical endpoints indicative of a physiological process. By counting individual colonies growing from known dilutions of environmental samples, different types of microorganisms can be enumerated.

3 *Physiological incubations*: microbial populations occur as complex communities in environmental samples. Environmental samples can be brought into the laboratory, sealed in vessels, and subjected to assays that demonstrate the physiological potential of the microorganisms, such as production of CO_2 or CH_4, nitrogen fixation, sulfate reduction, or metabolism of pollutant compounds.

4 *Biomarker extraction and analysis*: prokaryote-specific molecular structures (e.g., nucleic acids, proteins, lipids) can be directly extracted from environmental samples. After analysis, these provide insightful clues about the presence and activity of microorganisms (see also Section 2.2).

No single approach leads to a thorough understanding or answer to a given question. Information from all four approaches can complement and confirm one another. When this confluence occurs, the discipline of environmental microbiology is advanced.



Sequencing

Planning of experiments in microbial ecology

In contrast to laboratory experiments, methodical approaches, experimental design and data analysis in microbial ecology are complicated for several reasons: (1) the environment (as well as the collected sample) is highly complex (2) its composition is typically only partly known (3) sampling of environment is challenging due to its heterogeneity and variability in space and time.

*When **planning sampling or experiments** it is essential to consider*

- suitable methodology of sample collection and analysis
- representativity of results at the level of sample (size of sample)
- representativity of results at the (eco)system level (number of replicates)

*The **results** have to be statistically tested and following questions need to be answered:*

What are the samples like?

Are differences among samples statistically significant? At what level?

Is it possible to explain why are samples different?

Experimental work in microbial ecology

(1) Analysis of the present state-of-the-art

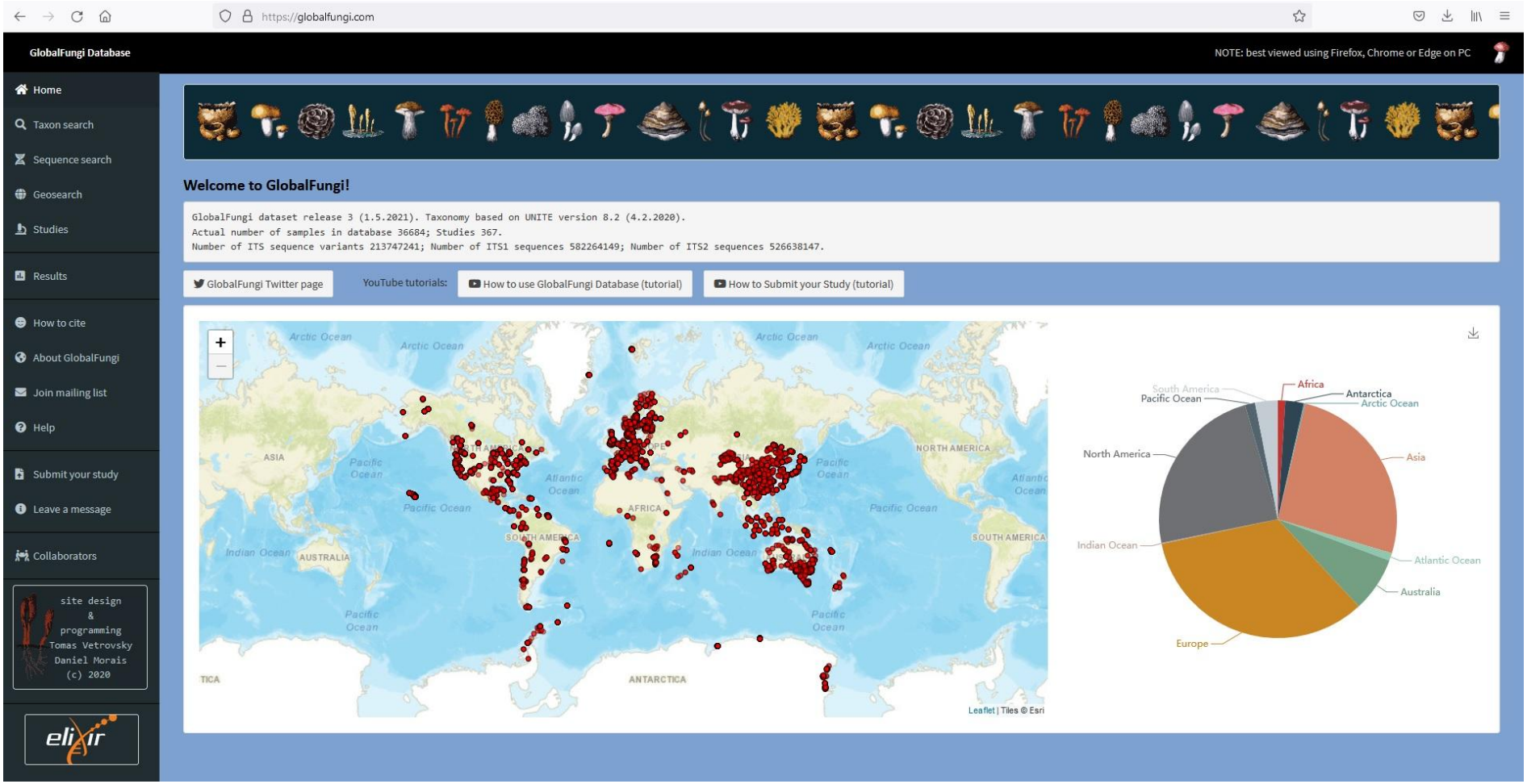
(2) Exploration and description of the environment

(3) Formulation of hypotheses

(4) Design of the experiment to test hypotheses

(5) Experimental verification / falsification of hypotheses

Meta-analysis of existing data



Information encoded in microbial genomes

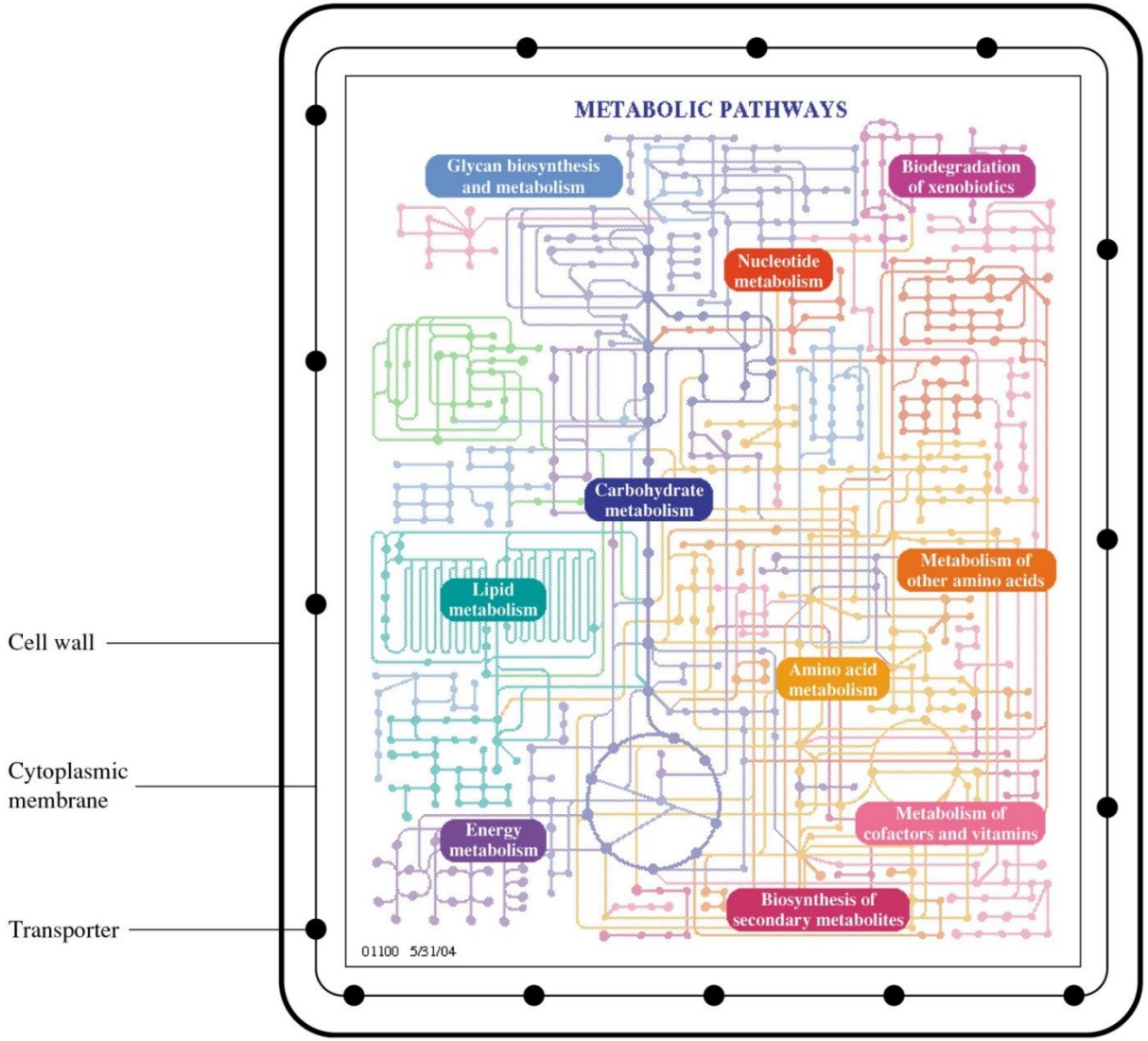


Figure 3.2 Conceptual cell model with bioinformatic template for organizing recognized metabolic genes from each microbial genome-sequencing project into a model prokaryotic cell. (Modified from M. Kanehisa, Kyoto University and KEGG, with permission.)

Information encoded in microbial genomes

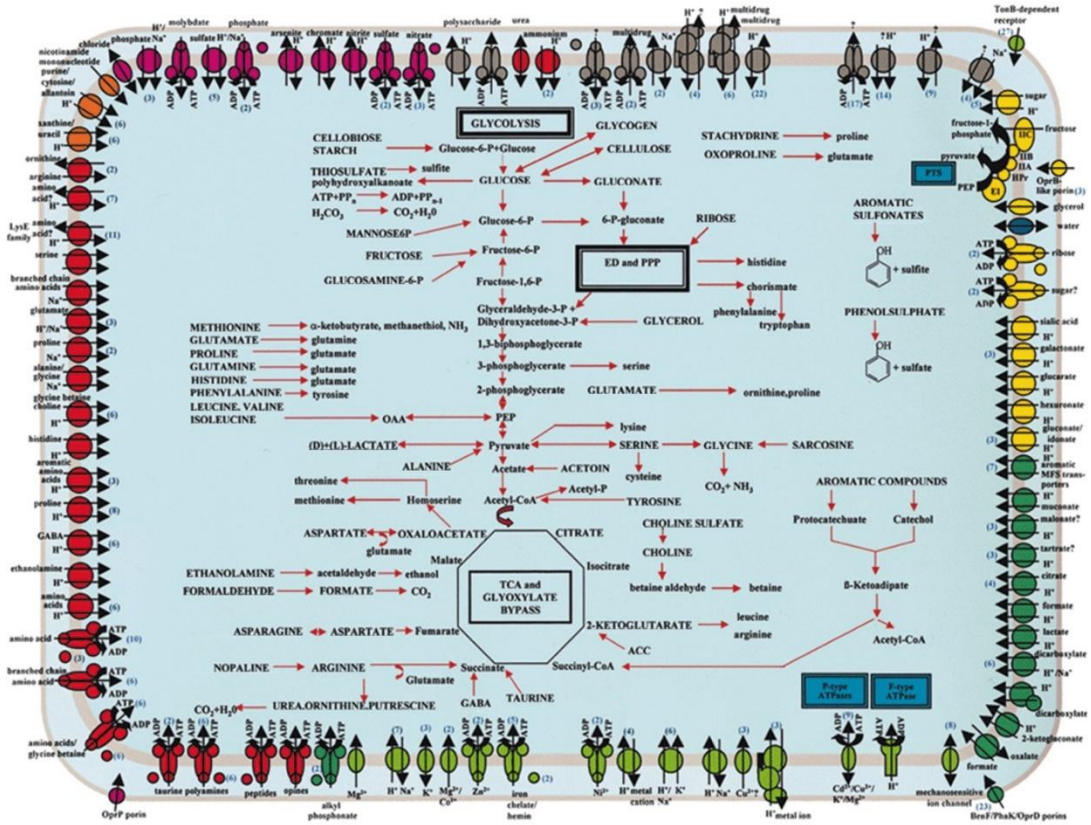


Figure 3.3 (*opposite*) A modeled genome for *Pseudomonas putida* KT2440. Shown is an overview of metabolism and transport based on the 6.18 Mb completed genome. Predicted pathways for energy production and metabolism of organic compounds are shown. Predicted transporters are grouped by substrate specificity: inorganic cations (light green), inorganic anions (pink), carbohydrates (yellow), amino acids, peptides, amines, purines, pyrimidines, and other nitrogenous compounds (red), carbohydrates, aromatic compounds, and other carbon sources (dark green), water (blue), drug efflux and other (dark gray). Question marks indicate uncertainty about the substrate transported. Export or import of solutes is designated by the direction of the arrow through the transporter. The energy-coupling mechanisms of the transporters are also shown: solutes transported by channel proteins are shown with a double-headed arrow; secondary transporters are shown with two arrowed lines indicating both the solute and the coupling ion; ATP-driven transporters are indicated by the ATP hydrolysis reaction; transporters with an unknown energy-coupling mechanism are shown with only a single arrow. The P-type ATPases are shown with a double-headed arrow to indicate that they include both uptake and efflux systems. Where multiple homologous transporters with similar predicted substrate exist, the number of that type of transporter is indicated in parentheses. The outer and inner membrane are sketched in gray, the periplasmic space is indicated in light turquoise, and the cytosol in turquoise. (From Nelson, K.E., C. Weinel, I.T. Paulsen, et al., 2002. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ. Microbiol.* **4**:799–808. With permission from Blackwell Publishing, Oxford, UK.)

Genomes of micro- and macroorganisms

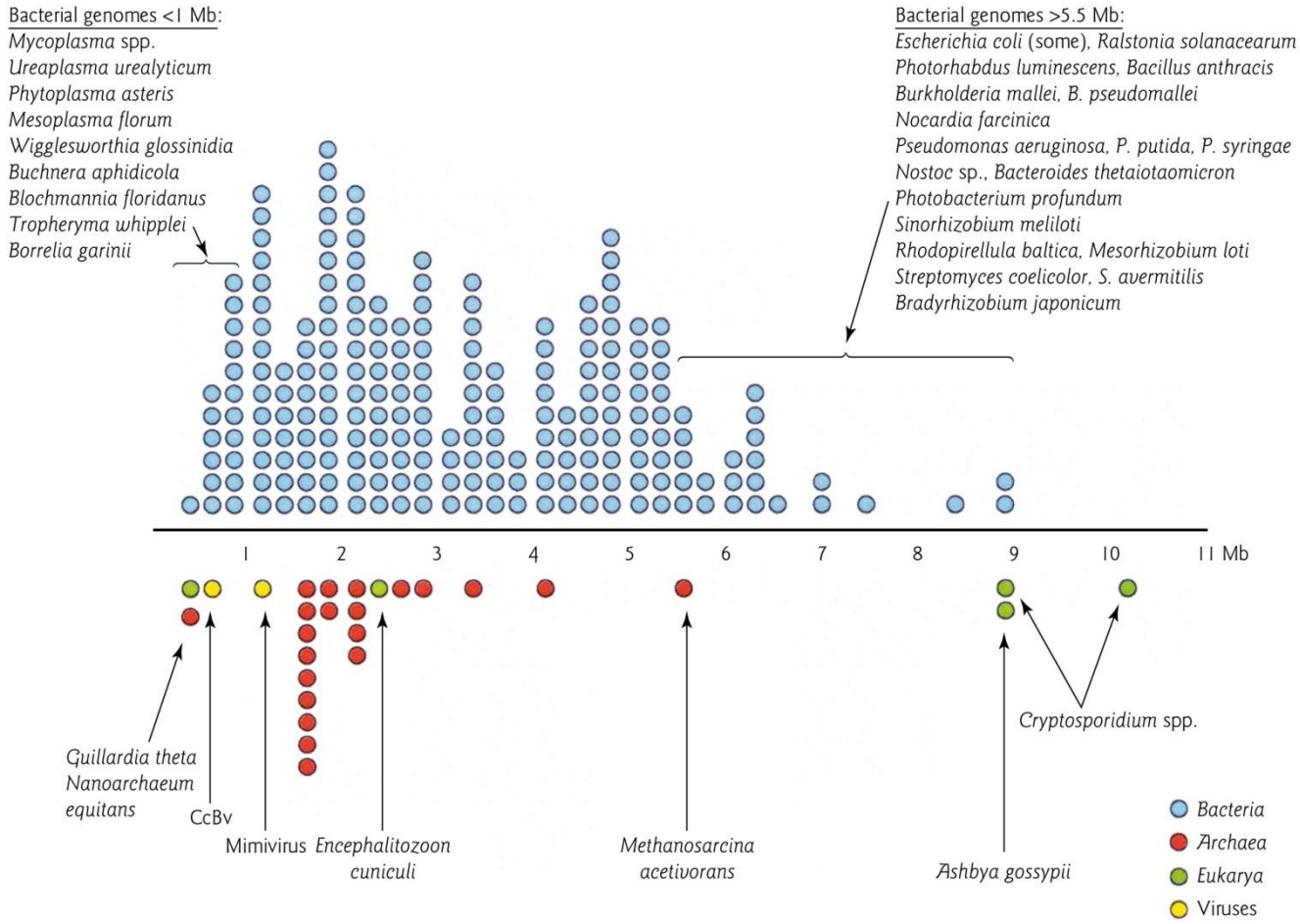


Figure 5.21 Depiction of overlapping genome size in members of the *Bacteria* (blue), *Archaea* (red), *Eukarya* (green), and viruses (yellow), in the size range (approximately 0.5–10.5 Mb) in which this overlap has been found to occur. The number of circles at a given point on the scale indicate the number of completed genomes that possess a specific size. Circles that represent unusually small (<1 Mb) or large (>5.5 Mb) bacterial genomes are labeled with the species name. (Reprinted from Ward, N. and C.N. Fraser. 2005. How genomics has affected the concept of microbiology. *Curr. Opin. Microbiol.* **8**:564–571. Copyright 2005, with permission from Elsevier.)

Tree of life or a network: horizontal gene transfer

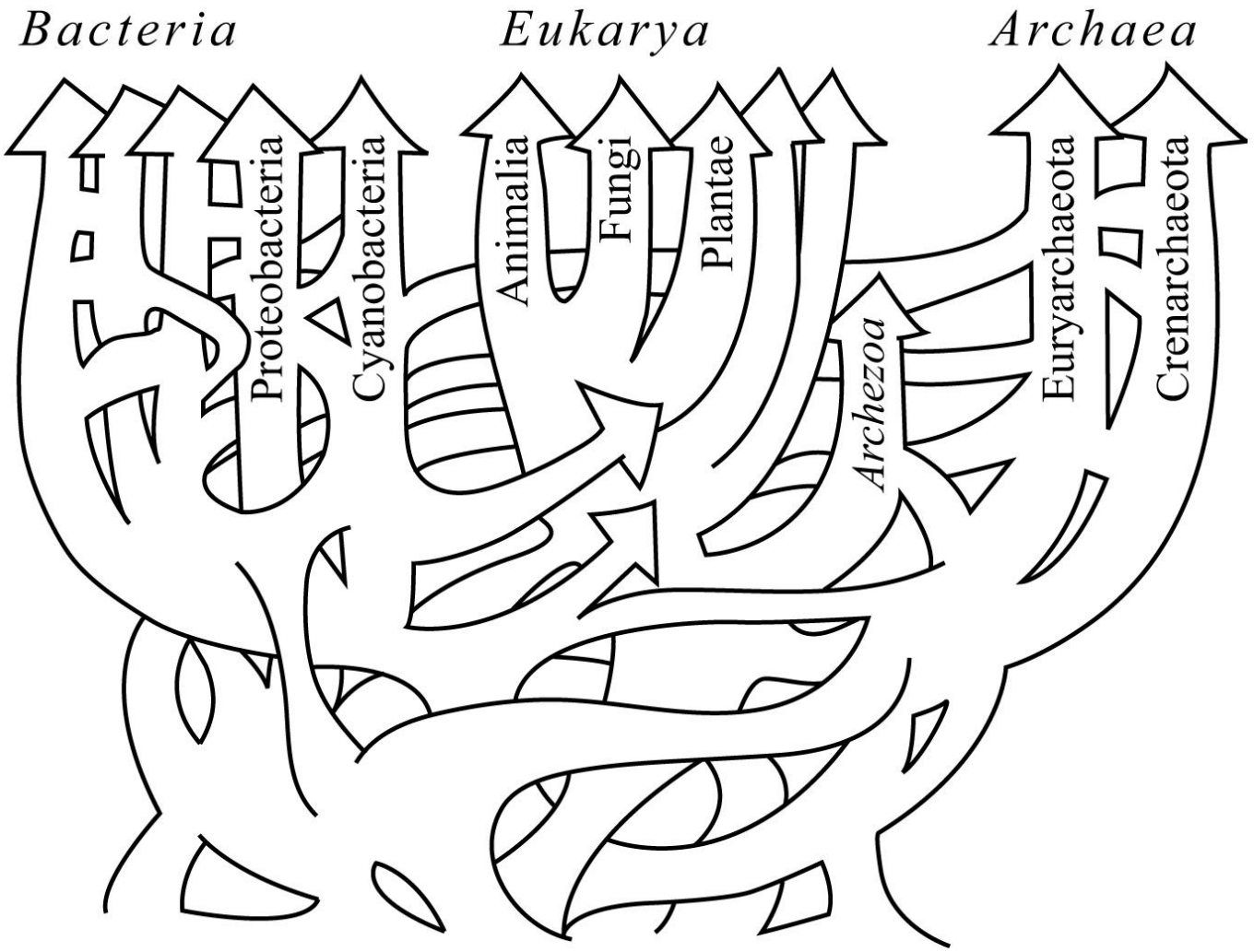
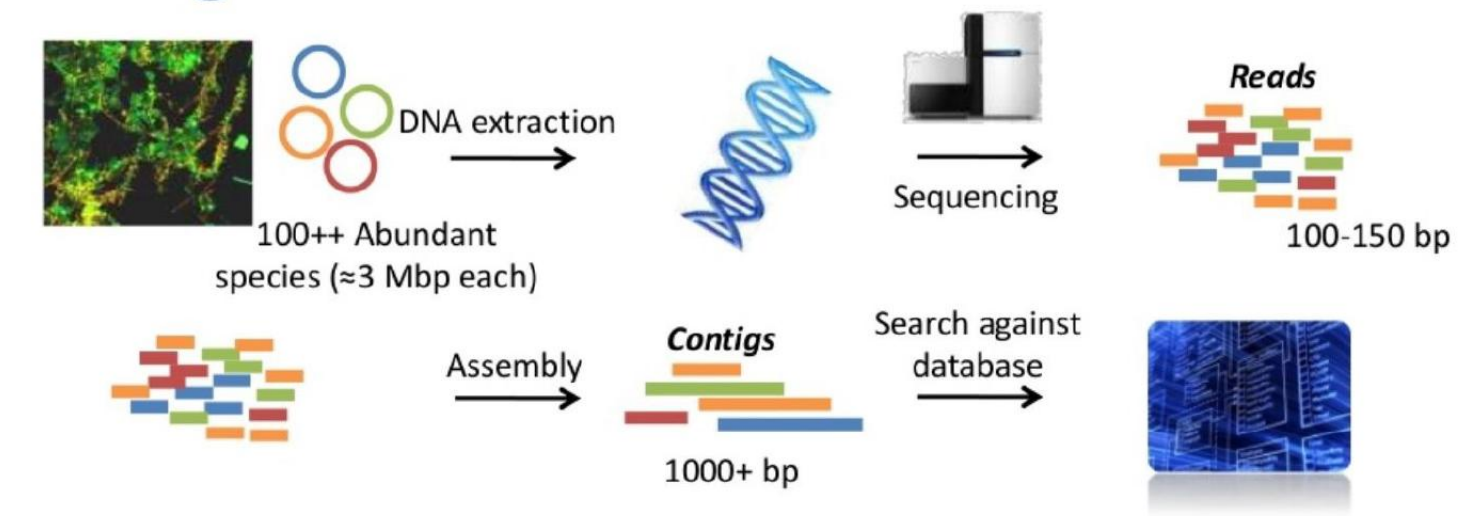


Figure 5.22 A version of the tree of life (based on small subunit rRNA sequences) that incorporates horizontal gene transfer processes in shaping the genetic composition of the three domains, *Eukarya*, *Archaea*, and *Bacteria*. (From Doolittle, R.F. 1999. Phylogenetic classification and the universal tree. *Science* **284**:2124–2129. Reprinted with permission from AAAS.)

Analysis of Environmental DNA/RNA: complex and computing-intensive

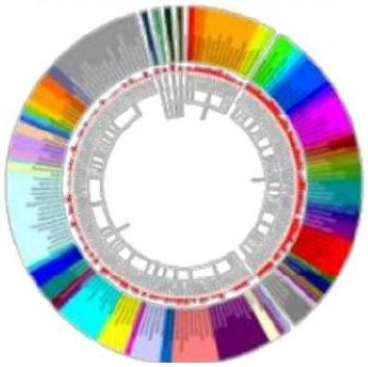
Metagenomes represent mixture of genome fragments: in 1 g of soil, there are up to several thousands of bacterial species.



Phylogenetic classification

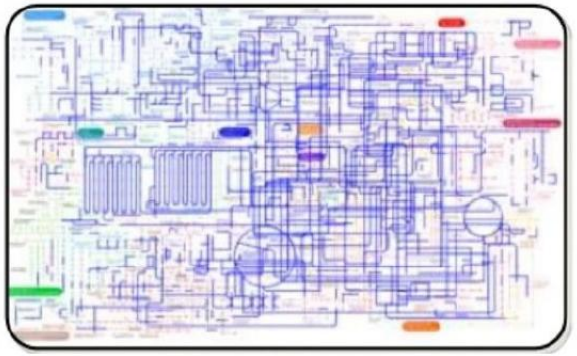
Who is there?

Bacterium A
Bacterium B
...
Bacterium X



Functional classification

What can they do?



Gene A
Gene B
...
Gene X

Mads Albertsen

Analysis of microbial communities using amplicon sequencing

Making species census

Ribosomal RNA:

- *structural molecules that form a backbone binding ribosomal proteins together*
- *conserved regions suitable for primer design*
- *variable regions with sufficient taxonomic resolution*

- *bacteria: 16S rRNA*
- *eukaryota: 18S rRNA, ITS*

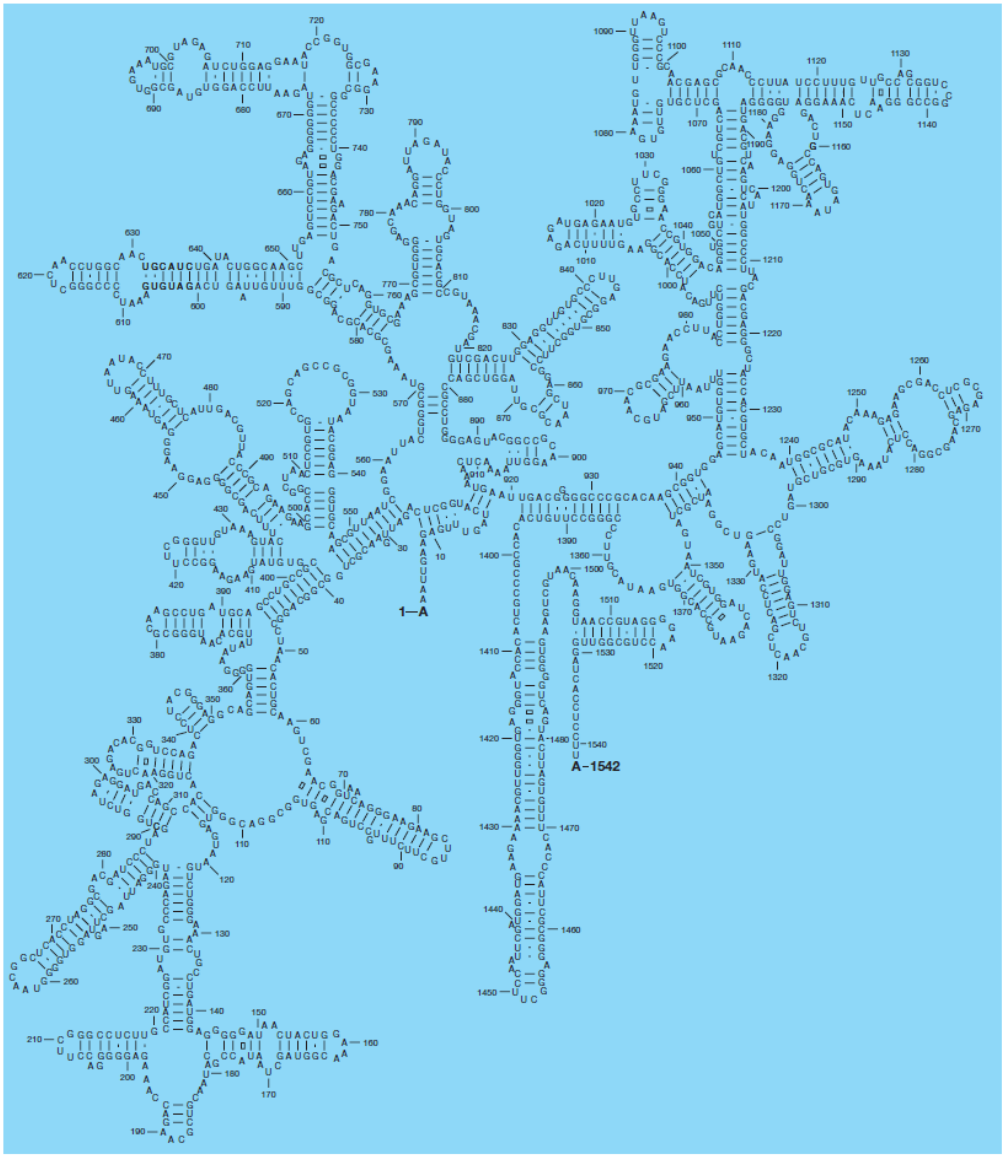
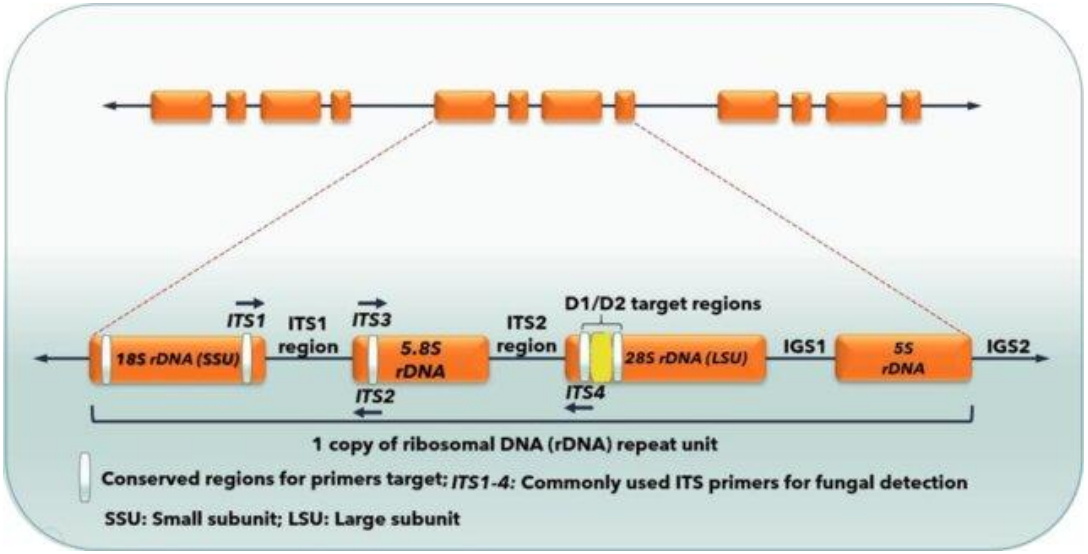


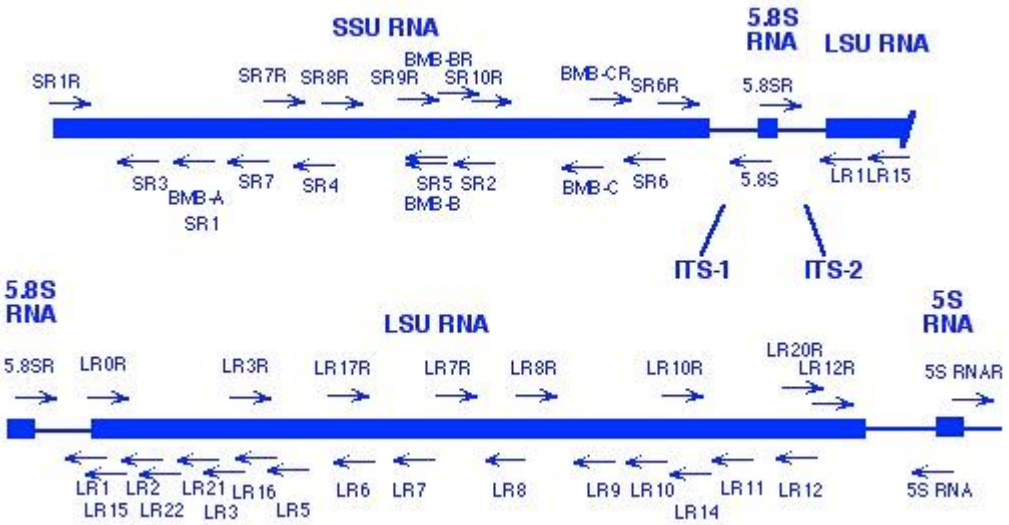
Figure 16.11 Ribosomal RNA (rRNA). Primary and secondary structure of 16S rRNA from *Escherichia coli* (*Bacteria*). The 16S rRNA from *Archaea* is similar in secondary structure (folding) but has numerous differences in primary structure (sequence).

The ribosomal casette (eukaryotes) – 5 kB of information



More information in the rDNA – diversity and phylogeny

Yahaya 2023 Academia Biology



Many primers: nothing is perfect

https://sites.duke.edu/vilgalyslab/rdna_primers_for_fungi/

Species definition in microorganisms

Standard criteria based on sexual crossing can not be used even in many cultured microorganisms

Box 5.2

Criteria to establish if microorganisms belong to the same species

There are three main taxonomic and molecular criteria used for deciding that two cultivated microorganisms belong to the same species (Rosselló-Mora and Amann, 2001; Stackebrandt et al., 2002; Gevers et al., 2005):

1. *DNA–DNA hybridization*: (i) 70% or greater DNA–DNA relatedness in whole genome reannealing tests; and (ii) <5°C difference in the temperature of DNA helix dissociation (ΔTm) for the two strains.
2. *16S rRNA gene*: >97% identity in sequence. This criterion is not absolute, as many phenotypically distinctive species have been discovered to have >97% 16S rRNA gene sequence identity.
3. *Whole genome sequence comparison*: here the criteria are not yet clear. Many strains of the same species can differ substantially in genome size and content. Identifying the critical phenotypic and genotypic traits that define "species" is an ongoing challenge, largely because subjective judgments must be made by taxonomists. A multigene strategy, known as multilocus sequence analysis (MLSA) has immense promise for developing new definitive criteria.

Biological species and molecular species

- Species are ecological entities. We can consider all their members functionally equal (this need not be true always).

- Marker gene sequences are conserved within microbial species (ideally, their species specific and show little variation within species)

- In a sequencing data, species are represented by sequences of their molecular marker – group of highly similar sequences

- Sometimes, biological species can be recognised, because its marker gene sequence is known. However, some microbial species were not yet observed (or assigned name).

- Molecular species are groups of sequences sharing some level of similarity (e.g., 97% similarity of 16S rRNA fragment) that likely represent the same species. They are often termed Operational Taxonomic Unit (OTU). Ideally, one OTU corresponds to one biological species.

- Species are created mathematically based on sequencing data.

Community analysis - OTU Table

samples

Tree ID	16376_2	16292_1	16302_1	16403_3	16405_2	16926_1	16919_1	16322_1	16259_1	16161_1	16124_1	16002_1	16026_1
Group name:	VOJ001	VOJ002	VOJ003	VOJ004	VOJ005	VOJ006	VOJ007	VOJ008	VOJ009	VOJ010	VOJ011	VOJ012	VOJ013
|CL00000000002	94	58	103	102	80	116	68	10	41	63	162	62	86
|CL00000000003	102	12	113	74	125	38	4	13	42	8	136	24	27
|CL00000000004	80	6	89	84	155	33	4	19	102	54	62	121	58
|CL00000000005	46	11	32	27	0	21	18	18	70	17	15	75	6
|CL00000000008	59	2	25	10	7	25	1	4	35	2	9	27	15
|CL00000000009	21	0	83	52	59	27	8	2	74	9	32	80	60
|CL00000000011	32	5	140	64	35	19	2	15	109	16	48	52	48
|CL00000000013	31	30	35	44	24	74	248	5	1	19	48	3	36
|CL00000000014	3	3	3	0	2	22	36	317	3	6	34	2	2
|CL00000000015	15	0	3	13	0	0	2	12	48	0	3	11	0
|CL00000000016	28	2	69	44	9	10	8	7	51	30	23	18	62
|CL00000000017	38	1	47	32	23	7	0	6	52	18	55	91	26
|CL00000000020	10	0	29	67	5	4	0	7	128	34	28	29	72
|CL00000000022	30	4	16	114	16	4	1	0	28	17	26	63	54
|CL00000000023	33	1	16	29	5	16	1	6	24	6	9	57	17
|CL00000000024	11	2	40	37	67	13	3	7	21	24	62	15	51
|CL00000000026	0	3	0	0	0	0	6	198	0	0	1	0	3

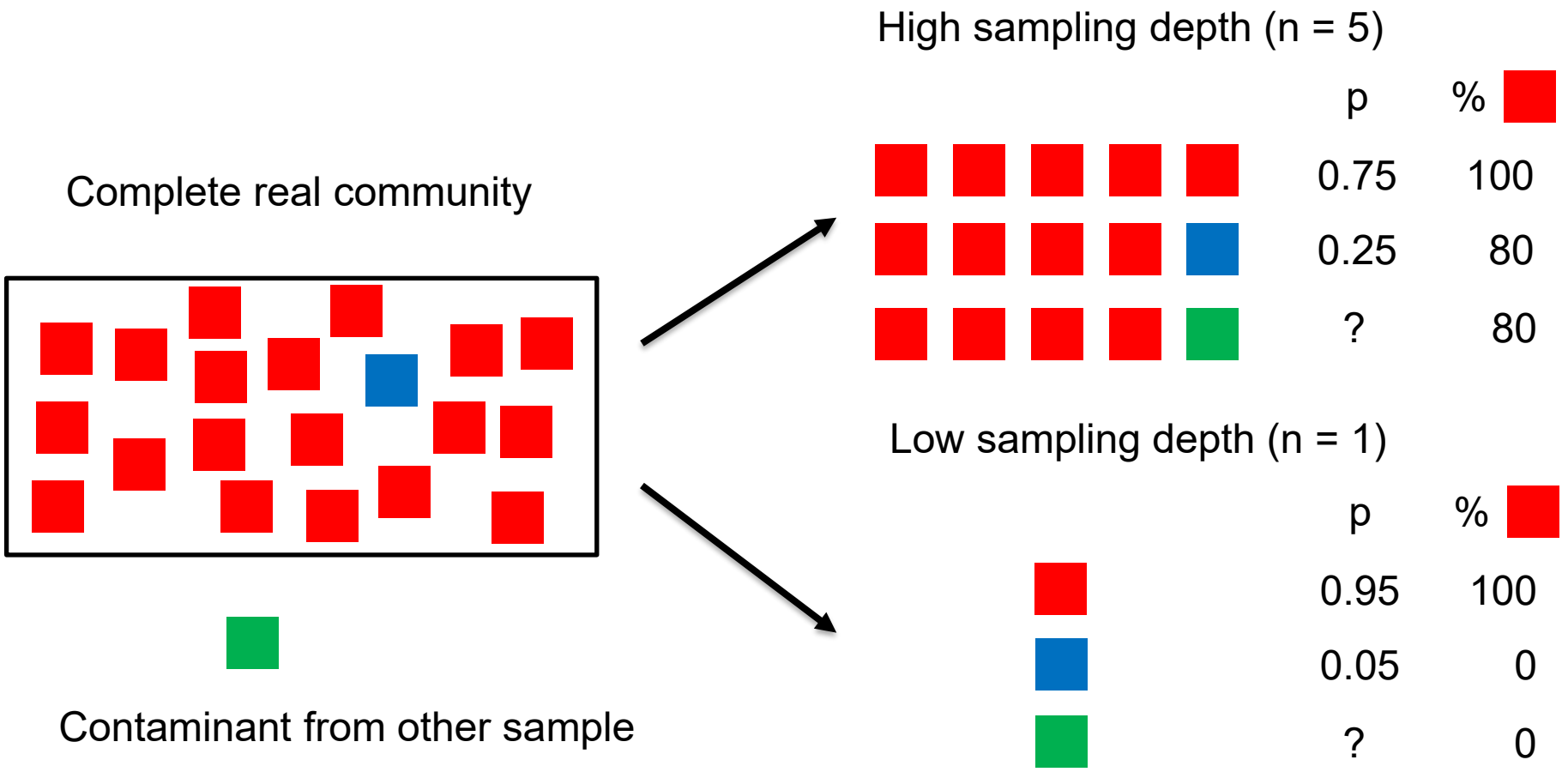
OTU (species)

sequence count of given species in given sample

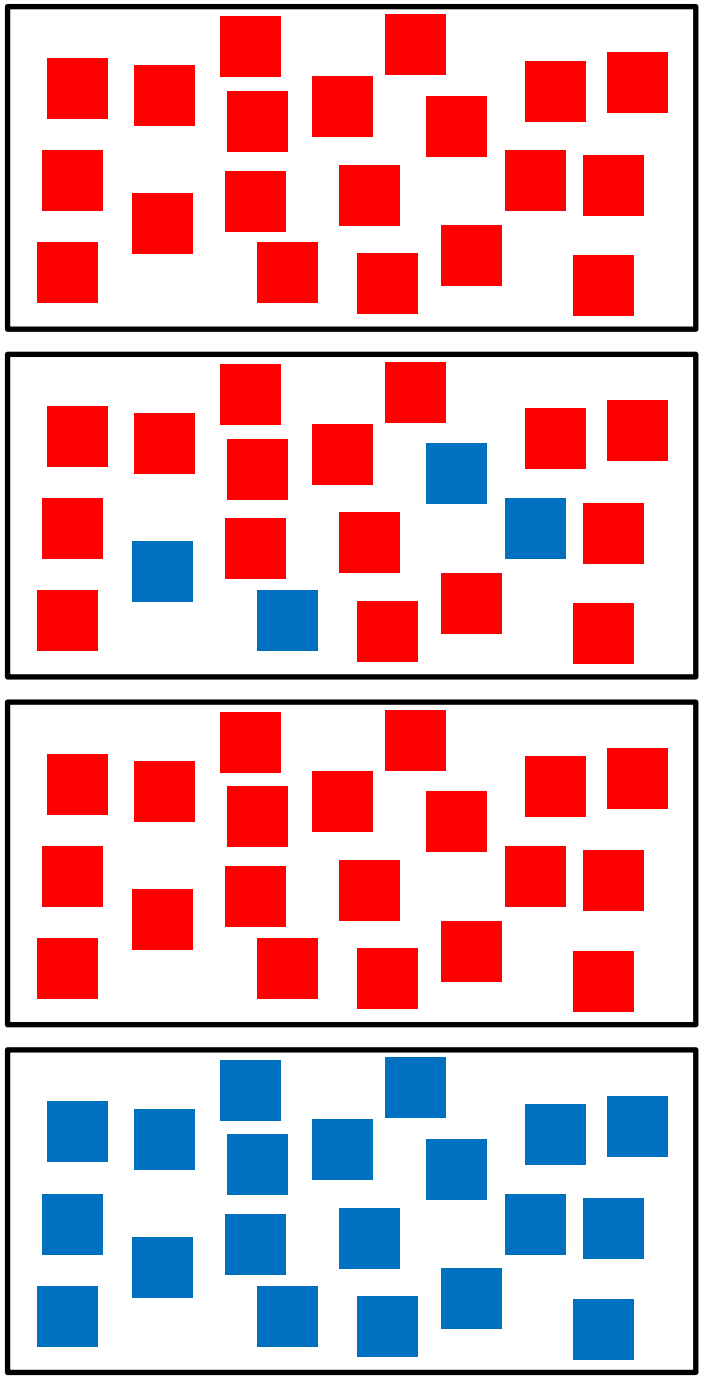
local singleton (sometimes removed)

Sampling of microbial communities is incomplete: this brings probability effects

Microbial community is represented by sequences sorted into molecular species (OTU)

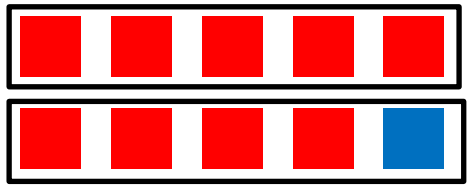


While plant community surveys can be „complete“, microbial can not: there are too many microbes in each sample



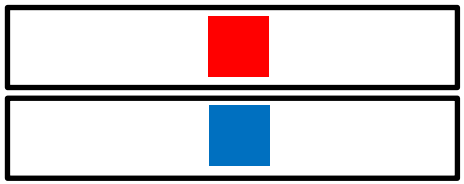
What we predict from sequencing results?

High sampling depth (n = 5)



p	%
0.75	100
0.25	80

Low sampling depth (n = 1)



p	%
0.95	100
0.05	0

Relative abundance estimates are imperfect due to probability: especially for low-rank taxa.

Sources of bias in amplicon sequencing

- Incomplete extraction of DNA (soft cells / resistant cells). Fragmentation of DNA by extended sample homogenization.
- PCR primer bias: primers bind some sequences more or less often than others
- PCR amplification bias: if markers differ in length, shorter amplicons are produced more often
- PCR is partly random (can be reduced by multiplication of PCR reactions)
- PCR may create chimeric molecules
- Sample contamination / cross-contamination
- Imperfect representation of biological species by molecular markers

Sources of bias in amplicon sequencing

OPEN ACCESS Freely available online PLOS ONE

The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses

Tomáš Větrovský, Petr Baldrian*
Laboratory of Environmental Microbiology, Institute of Microbiology of the Academy of Sciences of the Czech Republic, Praha, Czech Republic

Received: 10 July 2018 | Revised: 22 November 2018 | Accepted: 27 November 2018
DOI: 10.1111/mec.14995

Variability in bacterial 16S genes

ORIGINAL ARTICLE WILEY MOLECULAR ECOLOGY

Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles

Lotus A. Lofgren[1] | Jessie K. Uehling[2] | Sara Branco[3] | Thomas D. Bruns[2] | Francis Martin[4] | Peter G. Kennedy[1,5]

Variability in fungal ITS

FUNGAL ECOLOGY 6 (2013) 1–11

From gene copies to DNA and biomass content (in fungi)

Estimation of fungal biomass in forest litter and soil

Petr BALDRIAN, Tomáš VĚTROVSKÝ, Tomáš CAJTHAML, Petra DOBIÁŠOVÁ, Mirka PETRÁNKOVÁ, Jaroslav ŠNAJDR, Ivana EICHLEROVÁ*
Institute of Microbiology of the ASCR, Vídeňská 1083, 14220 Praha 4, Czech Republic

MOLECULAR ECOLOGY RESOURCES
Molecular Ecology Resources (2016) **16**, 388–401 doi: 10.1111/1755-0998.12456

The *rpb2* gene represents a viable alternative molecular marker for the analysis of environmental fungal communities

TOMÁŠ VĚTROVSKÝ,[1] MIROSLAV KOLAŘÍK,[1] LUCIA ŽIFČÁKOVÁ, TOMÁŠ ZELENKA and PETR BALDRIAN
Institute of Microbiology of the ASCR, v.v.i., Vídeňská 1083, 14220, Praha 4, Czech Republic

Quality assessment of primers for amplicon-sequencing of fungi

Sources of bias in amplicon sequencing: 16S of bacteria

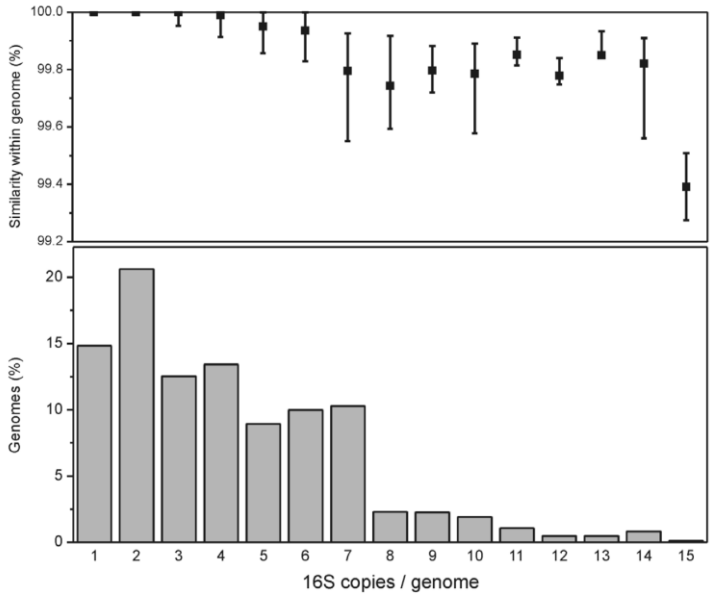


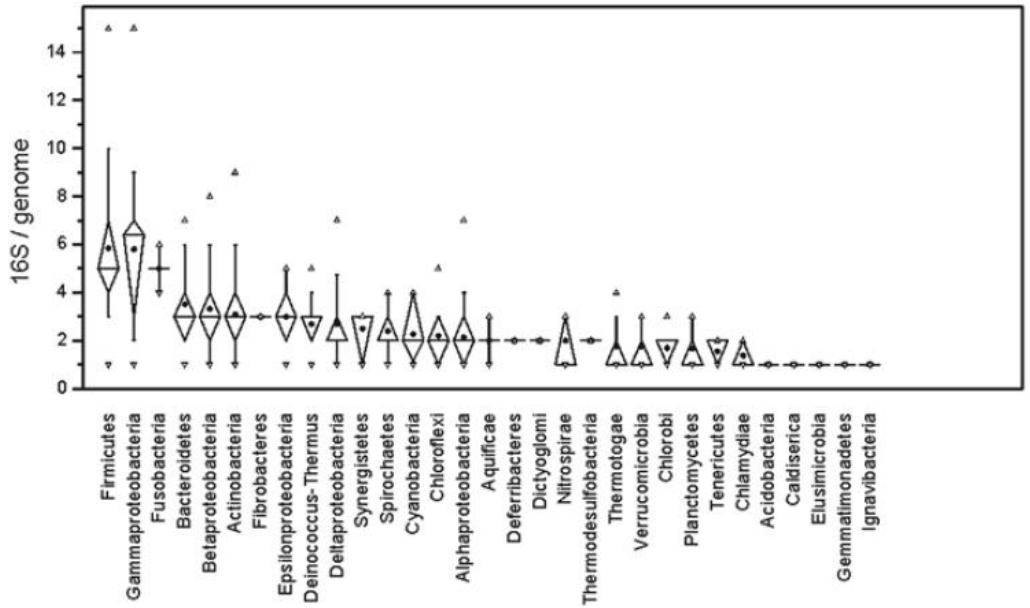
Figure 1. 16S rRNA within-genome similarity and copy numbers in bacterial genomes. Upper panel: the similarity of genomes with various copy numbers: the values indicated represent the first, the second and the third quartile. Lower panel: distribution of 16S rRNA copy numbers per genome in 1,690 sequenced bacterial genomes.
doi:10.1371/journal.pone.0057923.g001

Bacteria contain variable numbers of rRNA copies

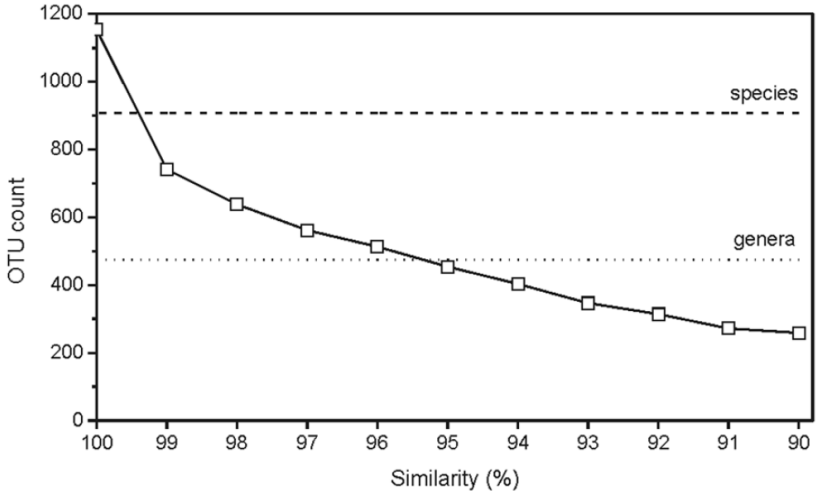
Sequencing gives counts of these copies, not of genomes or cells

16S copies in one genome show sequence differences

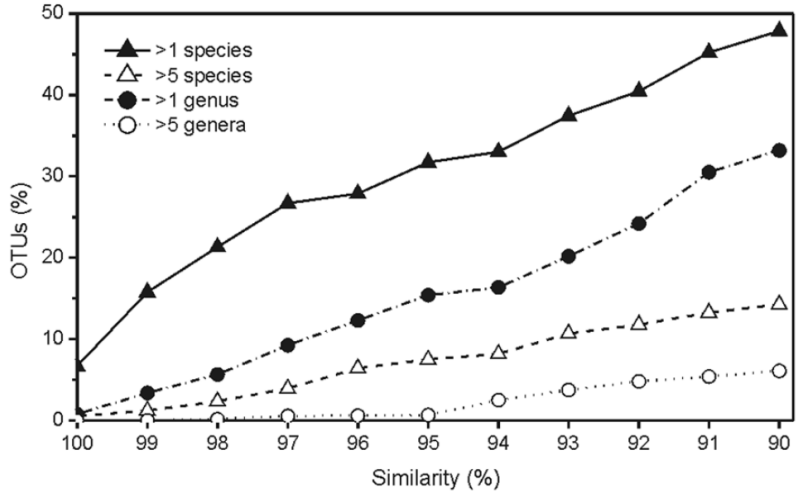
To some degree, copy counts of bacteria in genome are phylogenetically defined.

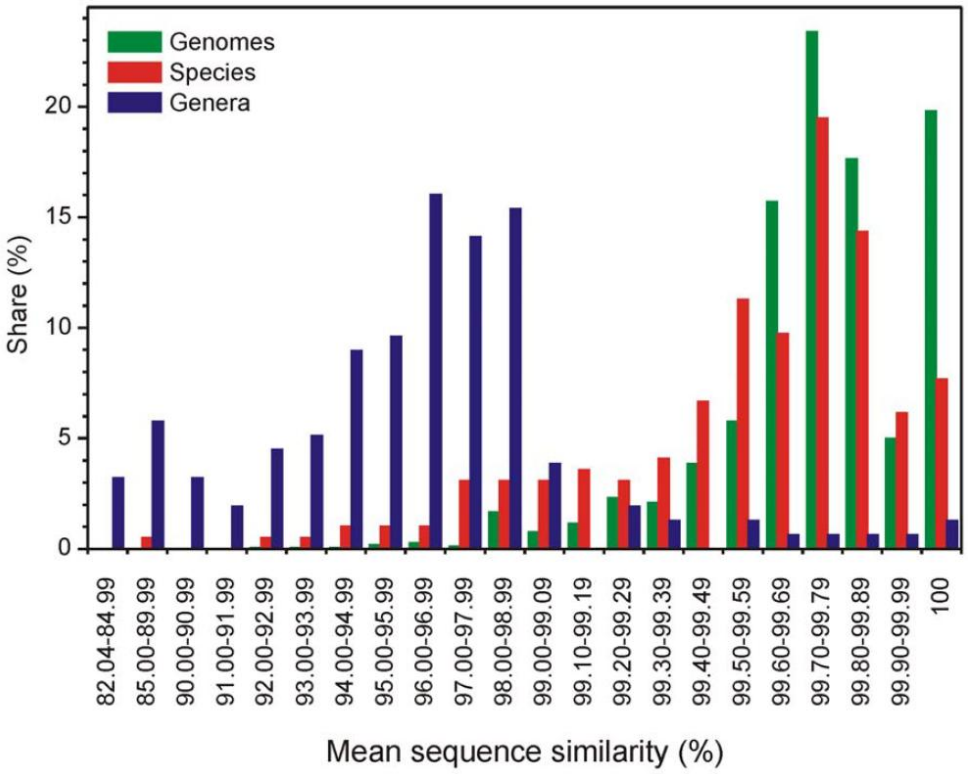


Sources of bias in amplicon sequencing: 16S of bacteria



The choice of sequencing similarit for OTU definition defines the number of OTU and how well they represent bacterial species.





How to correct bias in 16S amplicon sequencing?

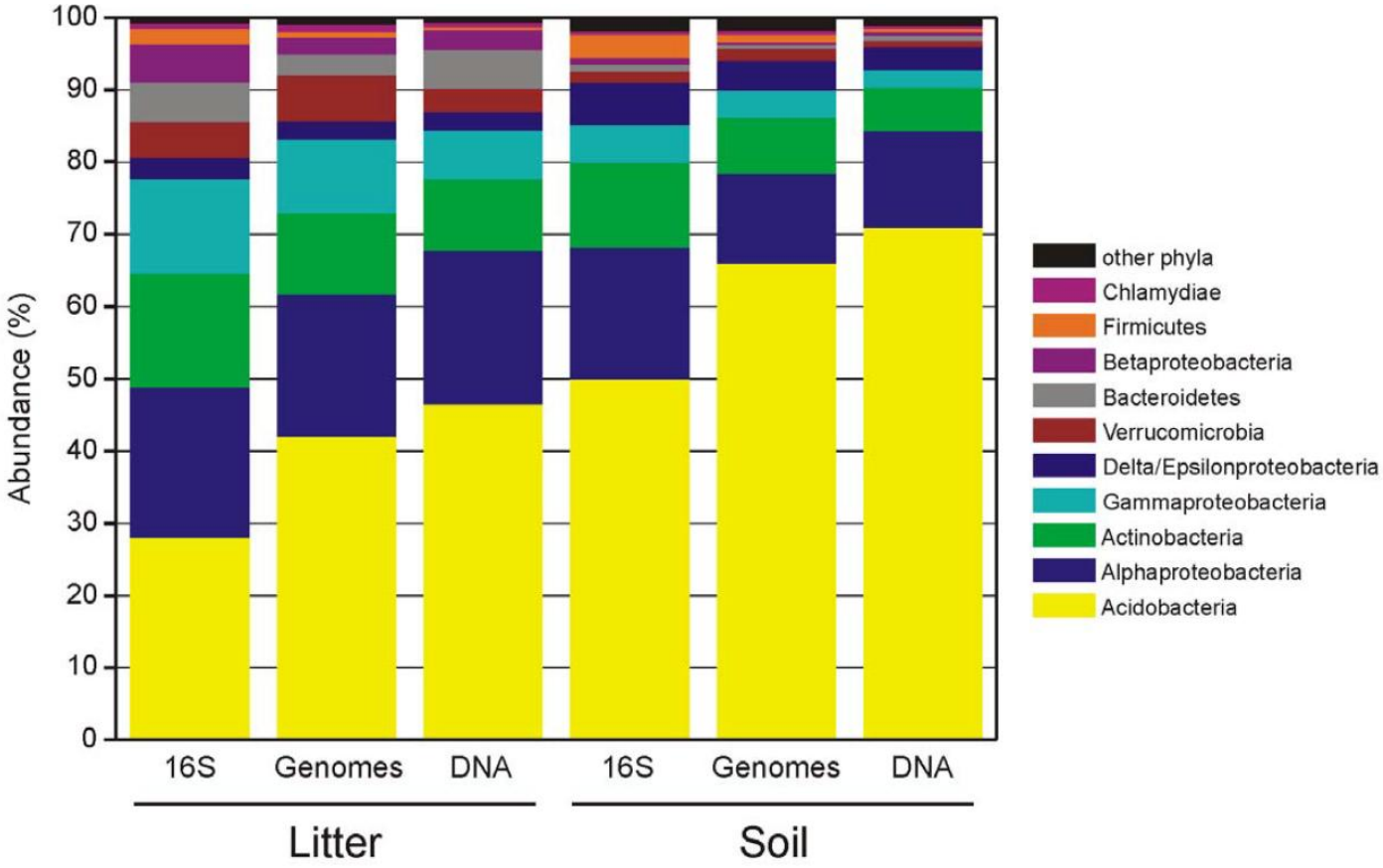
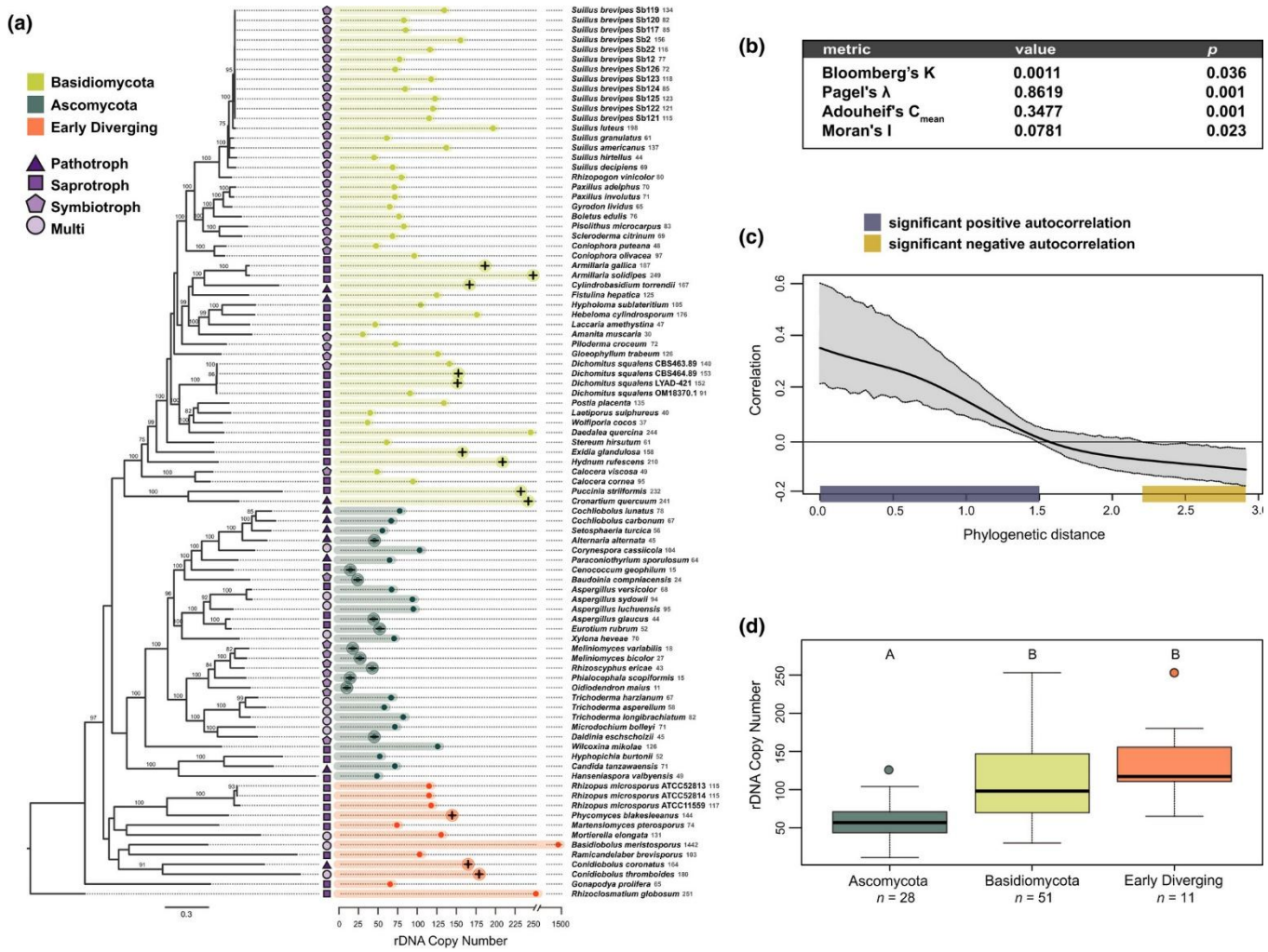


Figure 6. Abundance of bacterial 16S rRNA sequences, genomes and DNA in forest litter and soil. Relative abundance of bacterial 16S rRNA sequences in the amplicon pool from *Picea abies* litter and soil (Baldrian et al., 2012), and estimates of the relative abundance of bacteria genomes and DNA. The estimates were calculated using the values of 16S rRNA copy numbers and genome sizes of the closest hits to each bacteria OTU.

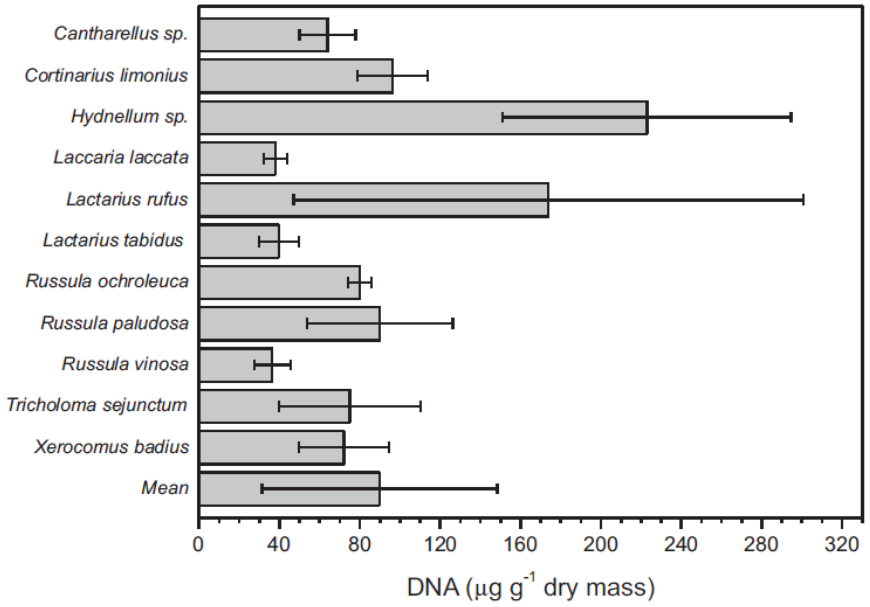
Bacterial genome counts in samples can be calculated when correcting 16S copy numbers by dividing sequence abundance by the 16S count in the most similar sequenced genome. This value probably best represents bacteria (cells) in sample.

Sources of bias in amplicon sequencing: rDNA of fungi



Ribosomal DNA copies per fungal genome: 20-1500, but typically 40-150
Taxonomic bias: Ascomycota have less copies than Basidiomycota

Sources of bias in amplicon sequencing: rDNA of fungi



Fungi show variation in the amount of DNA per g of biomass and copy numbers of rDNA (ITS) per genome or per ng DNA.

The counts of copies are estimated in tens to hundreds per fungal genome, but are unknown for vast majority of fungal species.

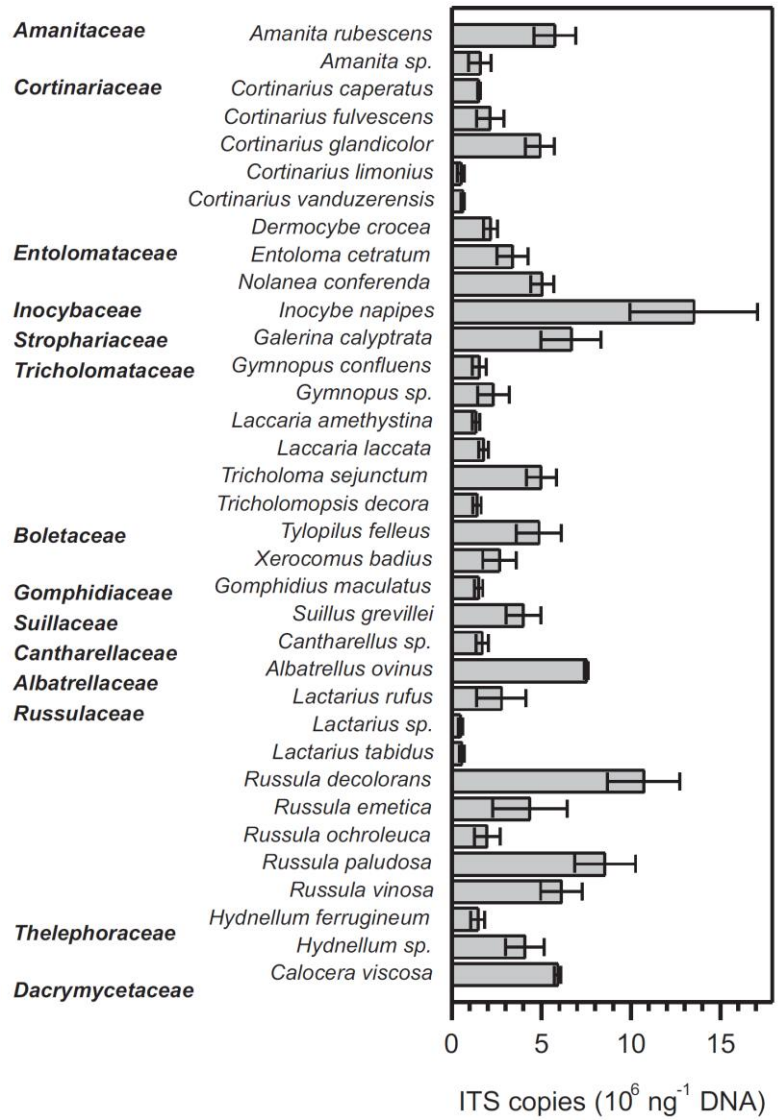


Fig 2 – Content of ITS in the DNA of fungi from the *Picea abies* forest. The values represent means and standard deviations from three replicates.

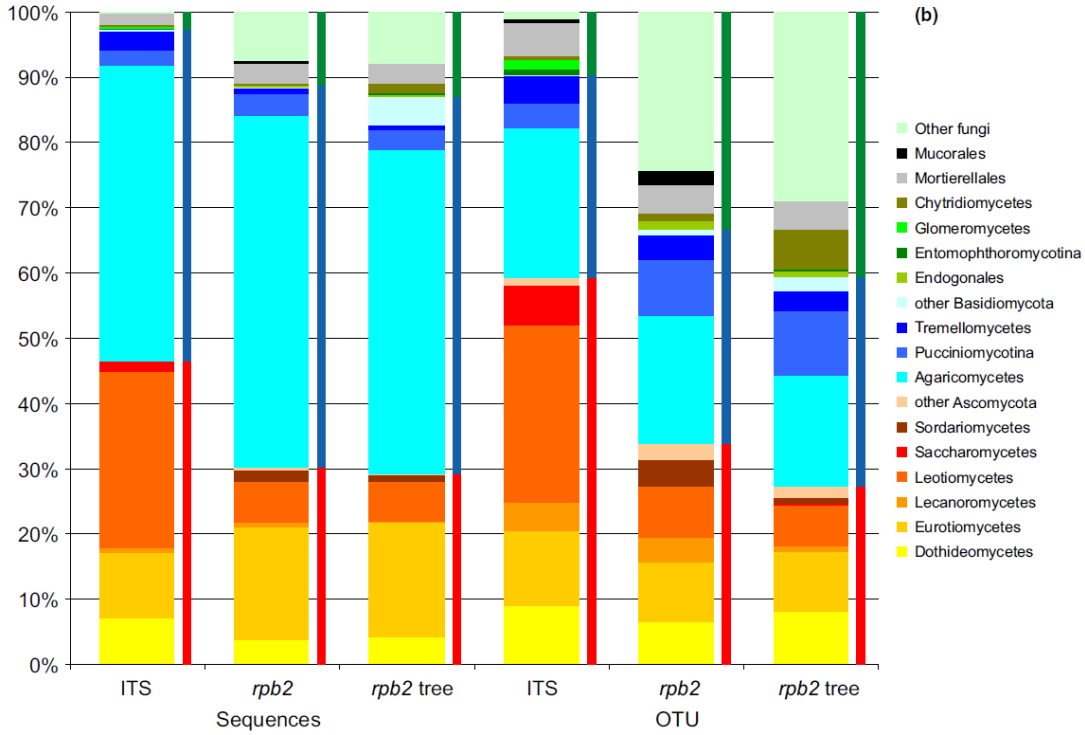
Sources of bias in amplicon sequencing: markers of fungi

Table 3 Comparison of the representation of major fungal groups in the mock community and in the sequence data sets obtained with ITS and *rpb2* as molecular markers

		ITS			*rpb2*		
	Mock community	Species detected	OTUs	Reads (%)	Species detected	OTUs	Reads (%)
Basidiomycota	83 (63.8%)	64 (63.4%)	205 (74.5%)	90.0	68 (70.1%)	133 (75.1%)	62.9
Ascomycota	42 (32.3%)	32 (31.7%)	62 (22.5%)	8.6	25 (25.8%)	38 (21.5%)	35.3
Mucoromycotina and Mortierellomycotina	5 (3.9%)	5 (4.9%)	8 (3.0%)	1.4	4 (4.1%)	6 (3.4%)	1.8

Copies of rDNA or other markers may differ within fungal genomes.

Depending on markers used, estimates of community composition may differ.



Sources of bias in amplicon sequencing: amplicon length

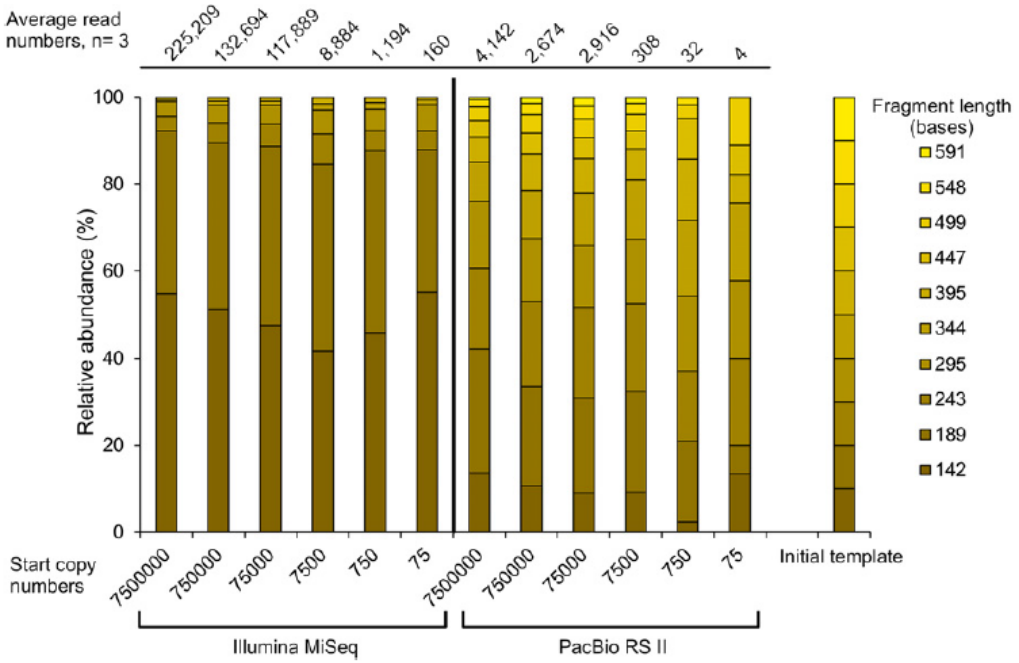


Fig. 2 Results from Expt 2, 'even communities'. Relative abundances of 10 mock community members varying in fragment size (142–591 bp) based on community sequencing with Illumina MiSeq or PacBio RS II. The ITS2 mock community consisted of 10 ITS2-like fragments, added in equal proportions to the initial PCR (initial template), but with different total starting quantities in the PCR (75–7500 000 copies). Above the graph, the averaged observed sequencing output is given for each sample ($n = 3$).

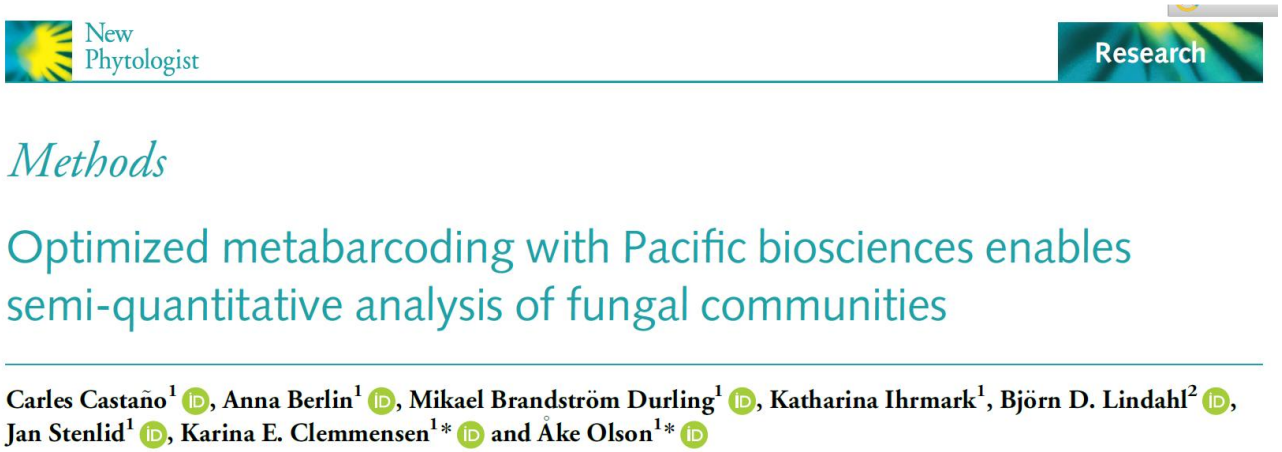
Castano et al. 2020 New Phytologist

Some amplified marker genes differ in length – typically the ITS1 markers of fungi with amplicon lengths between 142 and 591.

The shorter the amplicon, the more likely it is generated in PCR.

Bacterial 16S markers typically have equivalent length.

Barcoding of long amplicons – Is this the future?



Long read sequencing (up to 5000 bases) on recent Pacific Biosciences instruments slowly becomes established as a barcoding alternative for fungi.

Advantages:
- *longer reads can cover higher stretches of taxonomically informative regions*
- *more uniform read lengths are better represented after PCR*
- *better suitable PCR primer sites can be found across long distances*

Disadvantages:
- *higher costs of sequencing (per sequence approximately 4x)*
- *slightly lower quality*

<u>Oxford Nanopore Promethion Solo 2 as an alternative</u> – cheaper, poor quality

Short reads – platforms comparison

Illumina MiSeq

12 years on Market

Recent standard

Illumina i100

New in 2024

Illumina NextSeq 1000

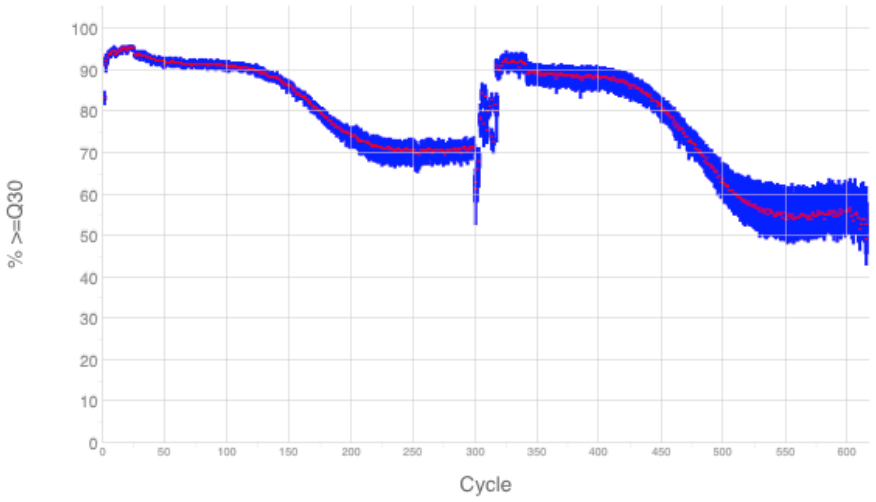
New in 2022

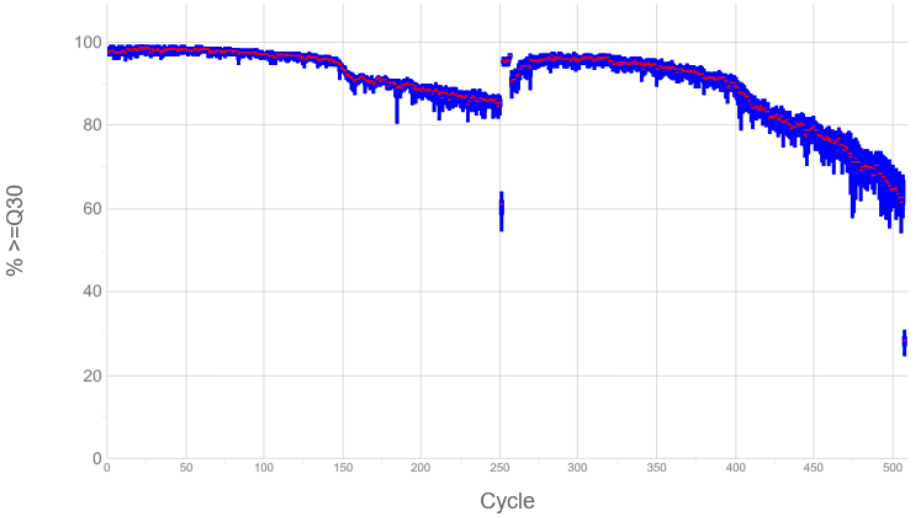
AVITI

New in 2024

… and others

Illumina NextSeq 1000 / Illumina MiSeq quality performance





Illumina NextSeq 1000

>Q30 per base

Illumina MiSeq

>Q30 per base

Platforms price comparison

PLATFORM	AVITI	Illumina NextSeq 1000	Illumina MiSeq	Illumina i100
Kit	AVITI 2x300 Sequencing Kit	NextSeq 1000 2x300 small	MiSeq 2x250	i100 2x300
Kit price without VAT	55275	47286.77686	43250	37000
Kit price with VAT	66882.75	57217	52332.5	44770
Kit throughput (pair end reads)	100000000	100000000	15000000	25000000
Read length	300	300	250	300
Price per 10 000 pair-end reads	6.688275	5.7217	34.88833333	17.908
Price per 1000000 pair-end bases	2229.425	1907.233333	13955.33333	5969.333333
Samples per run	2000	2000	600	900
Price per sample	33.441375	28.6085	87.22083333	49.74444444
Reads per sample (75% recovery)	37500	37500	18750	20833.33333
Knihovny	20	20	6	9
Library prep / library with DPH	2000	1800	1800	1800
Library prep	40000	36000	10800	16200
Library prep / sample	20	18	18	18
Total price per sample (CZK)	53.44	46.61	105.22	67.74
Total price per 10 000 sequences (CZK)	1.43	1.24	5.61	3.25
Quality	High	Lowest	Low	Highest
Q30 at 250 base	99%	68%	76%	99%
Q30 at 300 base	75%	66%	45%	99%

Alternatives to amplicon-basedbarcoding to represent the microbiome

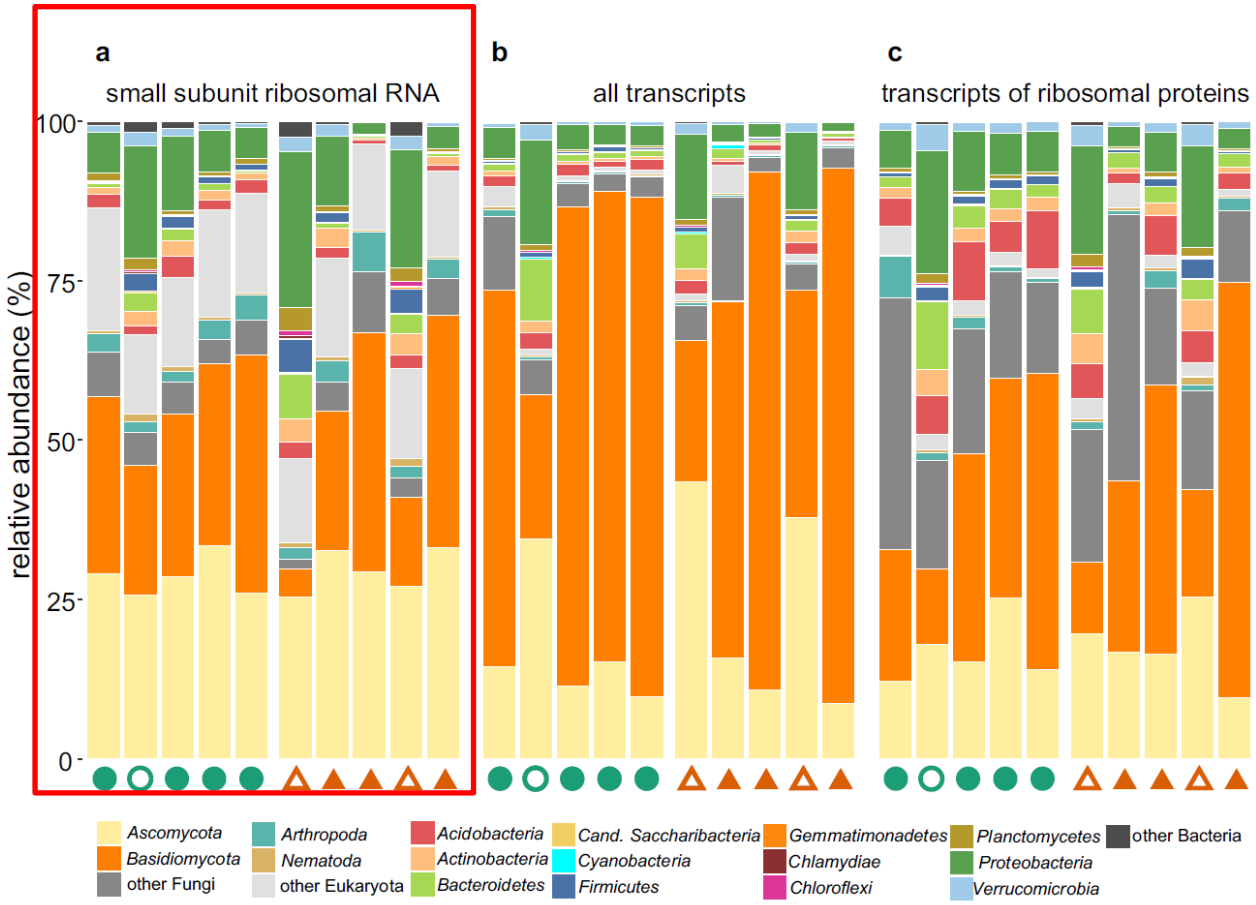


FIG 2 Taxonomic composition and activity of deadwood-associated organisms. Composition of the community of deadwood-associated organisms based on the relative abundance of small ribosomal subunit RNA, corresponding to ribosome counts (a), their activity based on the relative abundance of all mRNA transcripts (b), and their growth based on the relative abundance of mRNA of genes encoding ribosomal proteins (c). Filled symbols indicate samples dominated by Basidiomycota, and open symbols indicate those rich in Ascomycota and bacteria.

a) *Total RNA sequencing and analysis of the SSU sequences represents biomass content. Good coverage of all taxa, limited taxonomic resolution compared to marker gene amplicons*

Tláskal et al. 2021 mSystems

Alternatives to amplicon-basedbarcoding to represent the microbiome

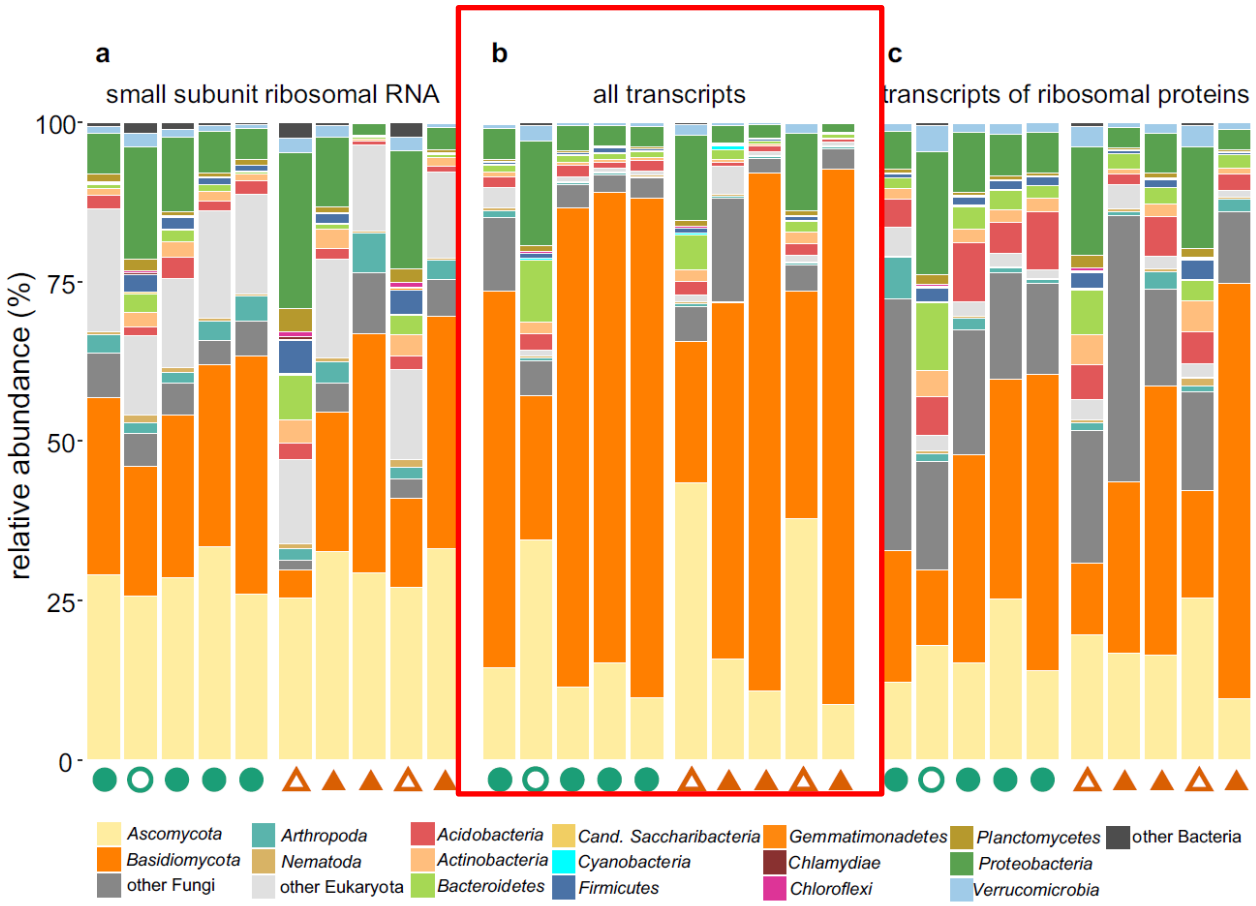


FIG 2 Taxonomic composition and activity of deadwood-associated organisms. Composition of the community of deadwood-associated organisms based on the relative abundance of small ribosomal subunit RNA, corresponding to ribosome counts (a), their activity based on the relative abundance of all mRNA transcripts (b), and their growth based on the relative abundance of mRNA of genes encoding ribosomal proteins (c). Filled symbols indicate samples dominated by Basidiomycota, and open symbols indicate those rich in Ascomycota and bacteria.

b) Assignment of all transcripts to taxa – active community members. Limited taxonomic information (phylum or class level), varying quality of assignment across taxa

Tláskal et al. 2021 mSystems

Alternatives to amplicon-basedbarcoding to represent the microbiome

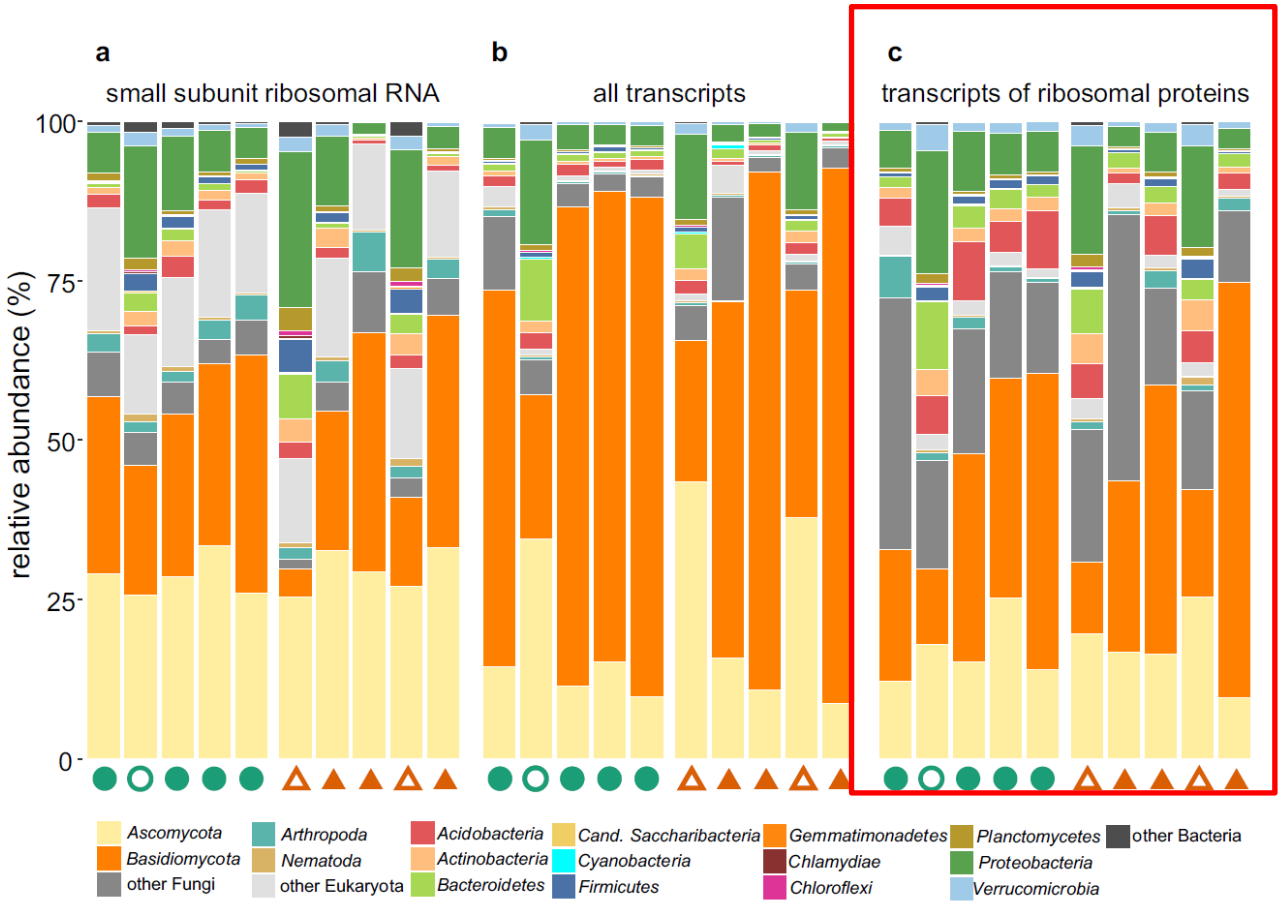


FIG 2 Taxonomic composition and activity of deadwood-associated organisms. Composition of the community of deadwood-associated organisms based on the relative abundance of small ribosomal subunit RNA, corresponding to ribosome counts (a), their activity based on the relative abundance of all mRNA transcripts (b), and their growth based on the relative abundance of mRNA of genes encoding ribosomal proteins (c). Filled symbols indicate samples dominated by Basidiomycota, and open symbols indicate those rich in Ascomycota and bacteria.

c) Assignment of transcripts of ribosomal protein genes – those community members that actively grow. Slightly more reliable taxonomic information (phylum or class level), comparable quality of assignment across taxa

Tláskal et al. 2021 mSystems

Alternatives to amplicon-basedbarcoding to represent the microbiome

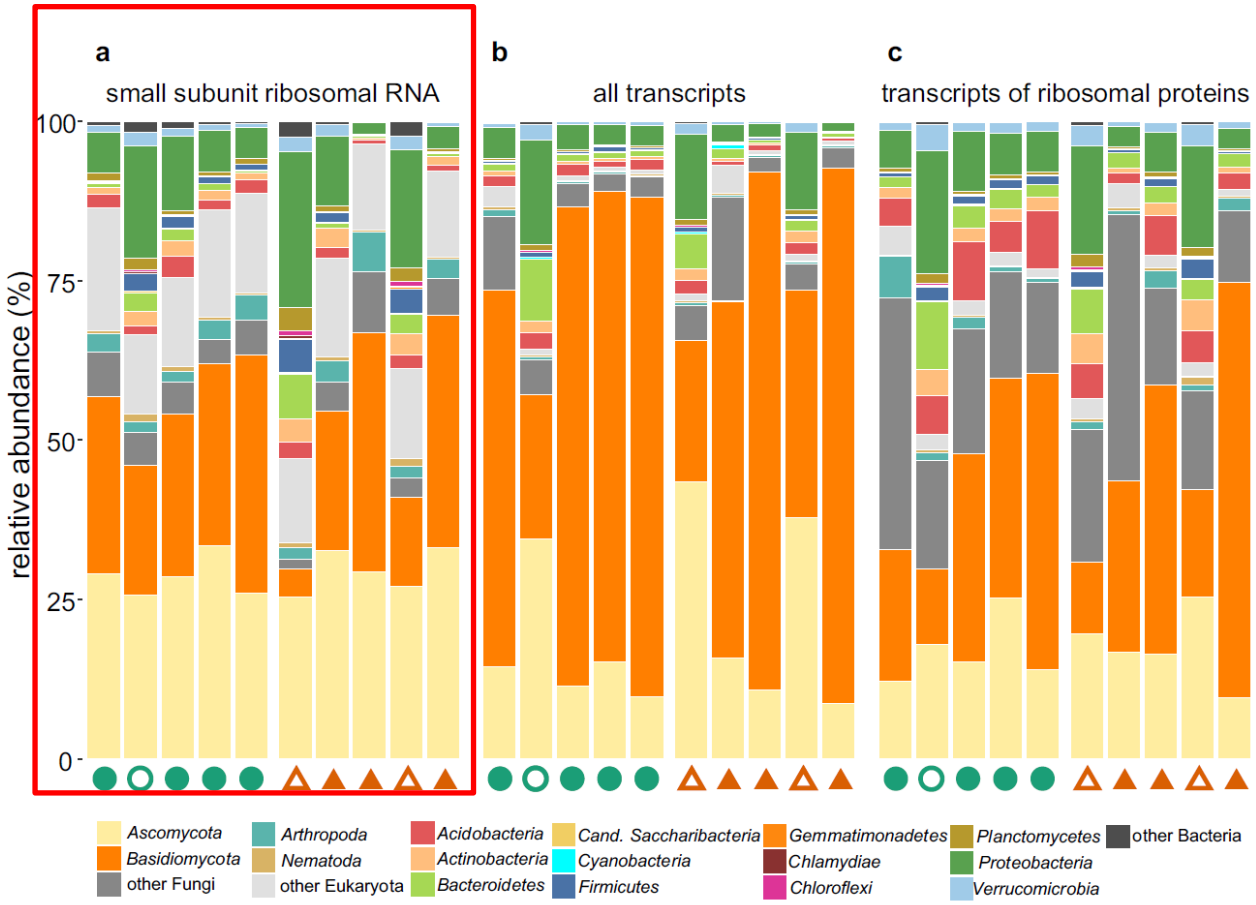


FIG 2 Taxonomic composition and activity of deadwood-associated organisms. Composition of the community of deadwood-associated organisms based on the relative abundance of small ribosomal subunit RNA, corresponding to ribosome counts (a), their activity based on the relative abundance of all mRNA transcripts (b), and their growth based on the relative abundance of mRNA of genes encoding ribosomal proteins (c). Filled symbols indicate samples dominated by Basidiomycota, and open symbols indicate those rich in Ascomycota and bacteria.

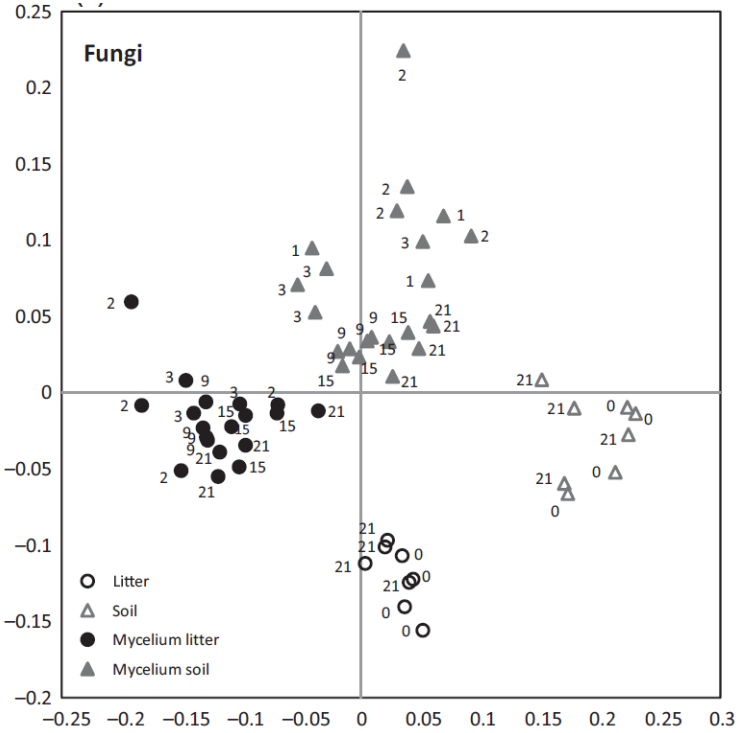
a) *Total RNA sequencing and analysis of the SSU sequences represents biomass content. Good coverage of all taxa,* ***excellent taxonomic resolution when sequencing long amplicons (PacBio HiFi reads)***

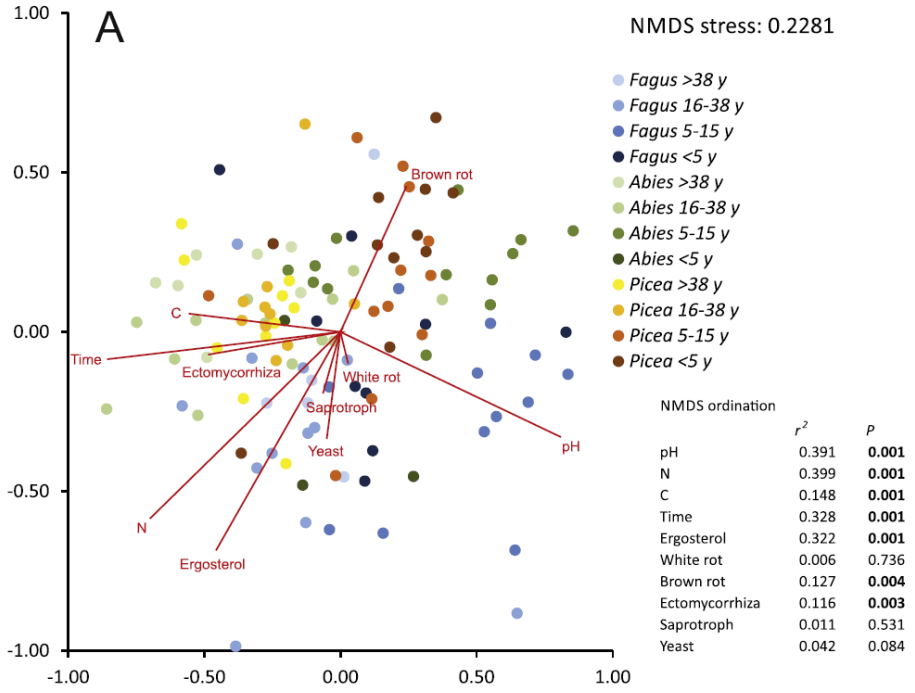
Tláskal et al. 2021 mSystems

Environmental microbiomes:

Data visualization and statistics

Variation of environment (not a bias): stochasticity plays higher or smaller role





Some part of variability is unexplained

You can fight stochasticity using higher replication

Field sampling

Representativity

- select sufficiently large sampling plot (10-25 m^2 in forest)
- take as large sample as possible (>50 g)
- take as many spatial replicates (at least 5 spatially independent soil cores); soil cores in forests are spatially independent if taken (approximately) 2 m from each other

Replication - treatments

- <3 replicate plots: no statistical tool available to assess p value of differences
- 3 replicates: can be analysed by ANOVA or other tests that assume normal distribution (microbial community composition is NEVER normal)
- 4 replicates: minimum to test for p value differences if distribution is not normal
- 5-6 replicates: minimum replication for treatments with low variation
- 10-12 replicates: reasonable compromise between replication and costs in most treatment

Replication – effect of environmental parameters

- 15 replicate plots minimum, 25+ plots reasonable for correlative statistics
- 60+ replicates needed for structural equation modelling (identify causal relationships)

Data visualization and statistics

Comparing treatments	*Can the samples be divided into cathegories?*
Basic data visualization	
Barplots	*What is the community composition in a treatment?*
Box-and-whisker	*What is the variability (median value, variation and extremes)?*
Heatmaps	*Visualize community composition at a small area of the graph*
Finding patterns in data	*Is the distribution of data normal or not? How to normalize data?*
Non-metric multidimensional scaling	*How do treatments differ (no normal distribution)*
Principal component analysis	*How do treatments differ (normal distribution)*
Testing of statistical significance	
Analysis of variance	*Are there statistically significant treatment effects (normal data)?*
PERMANOVA	*Are there statistically significant treatment effects (any data)?*
Mann-Whitney U-test	*Which treatments differ significantly from each other (any data)?*
Post-hoc correction tests	*Which treatments differ significantly from each other?*
Variation paritioning	*What is the contribution of multiple effects to variation?*
Factors behind significance	
Indicator species analysis	*What microbial taxa are characteristic for a treatment?*

Data visualization and statistics

Finding environmental effects	*Treatments are not defined, looking for effects of envir. variables*
Similarity	*How much do two samples differ from each other?*
Bray-Curtis similarity	
Correlations	*Apply with caution! Outlier samples affect results. Correlation among factors may prevent explanations.*
Pearsons linear correlations	*What is the relationship between two variables?*
General linear models	*What are the relationships among multiple variables?*
Mantel test	*What are the relationships between two data matrices?*
Random forest models	*How do variables change across gradients?*
Spatial statistics	*What is the effect of spatial distance?*
Autocorrelation	*How do samples differ across spatial scales?*
Geostatistics	*What models of spatial variation to apply?*
Finding causal relationships	*Why do we observe what we observe?*
Structural equation modelling	*What variables cause observed effects?*