Metabarcoding Pipeline Building

An introduction with hands on exercises

CUSO – DPEE Activity

Gerhard Thallinger, PhD, Graz University of Technology

Rachel Korn, PhD, Université de Fribourg

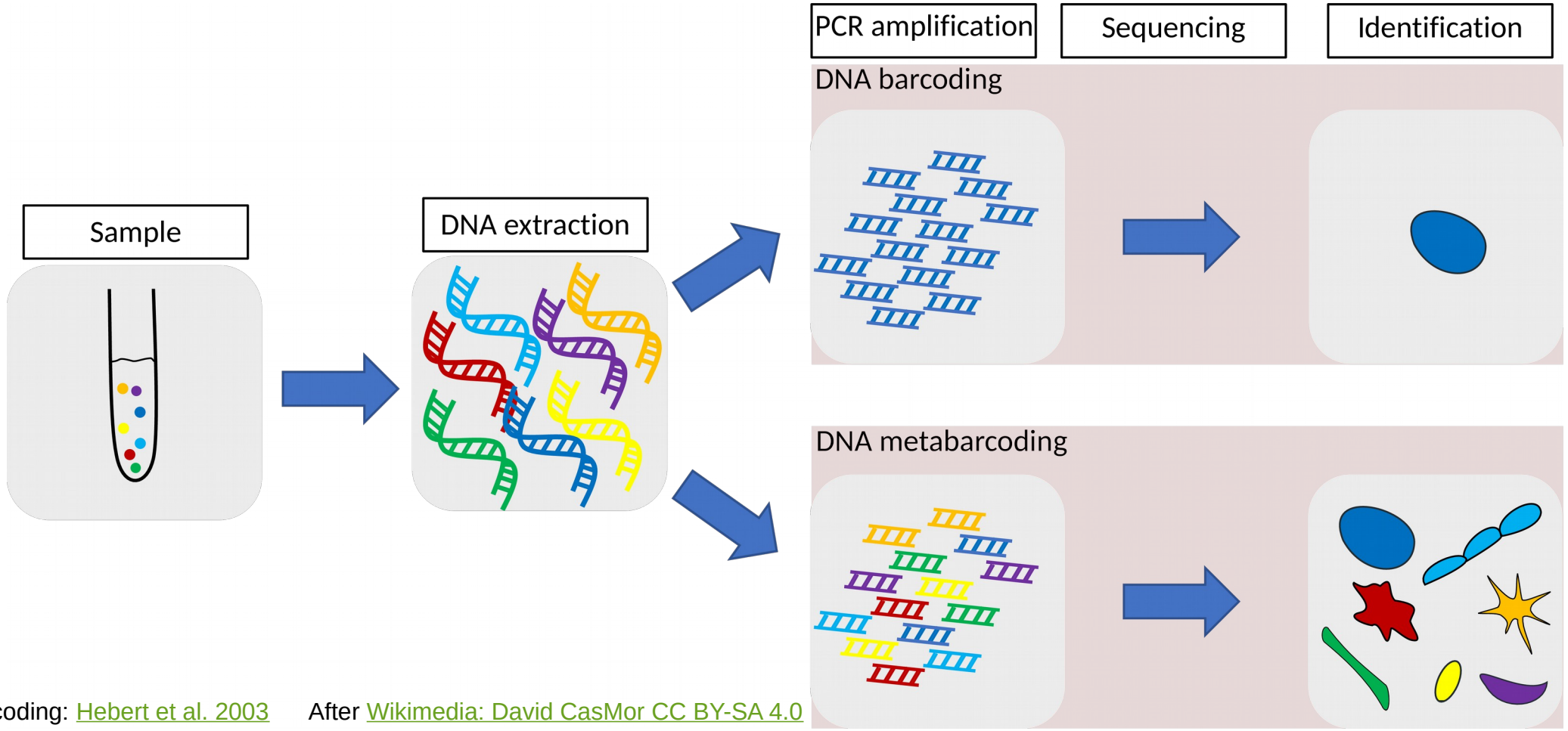
Magdalena Steiner, PhD, Agroscope Wädenswil

CUSO
CONFÉRENCE UNIVERSITAIRE
DE SUISSE OCCIDENTALE

TU Graz

UNI FR

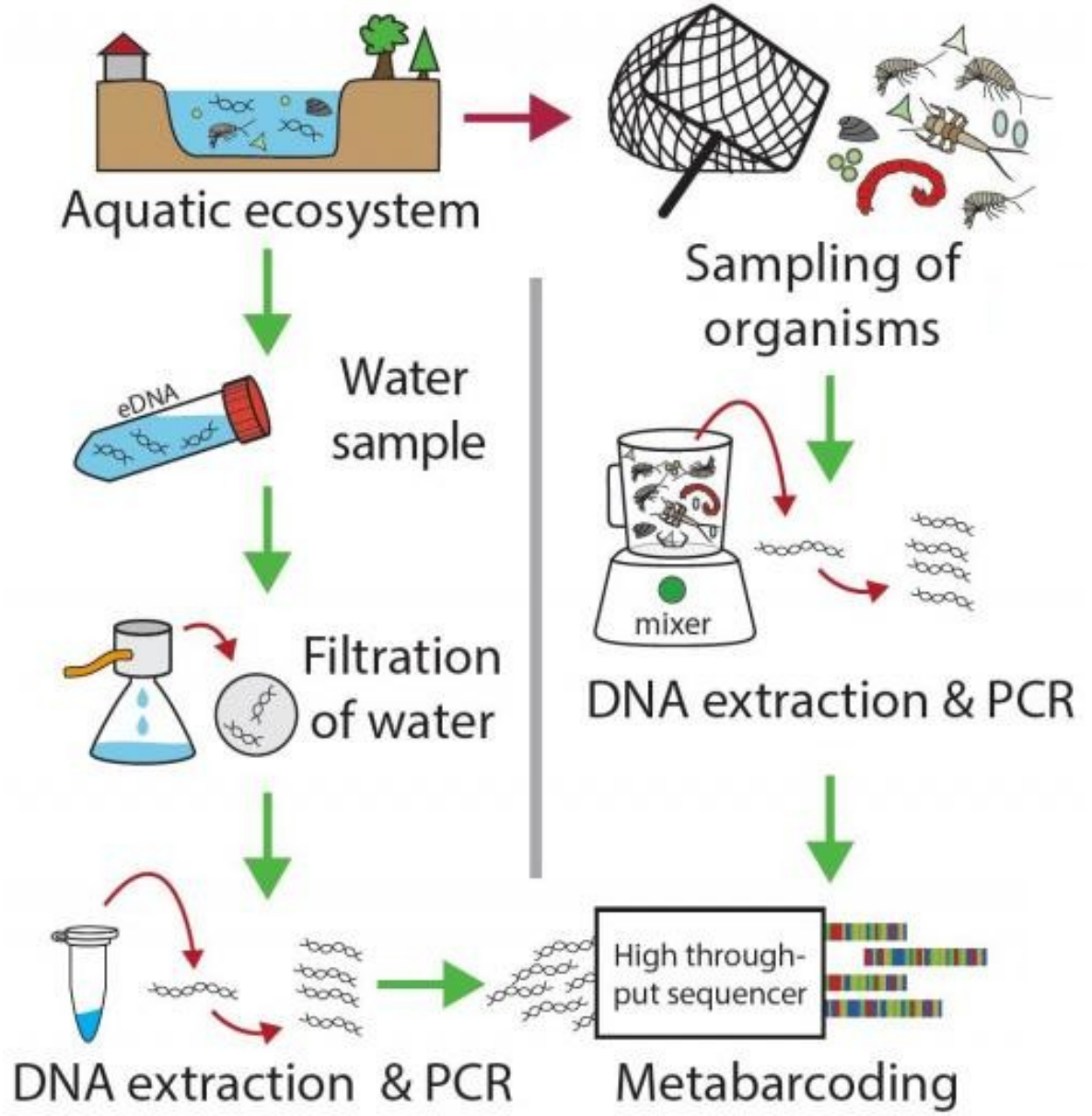
What is metabarcoding?



Barcoding: Hebert et al. 2003 After Wikimedia: David CasMor CC BY-SA 4.0

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

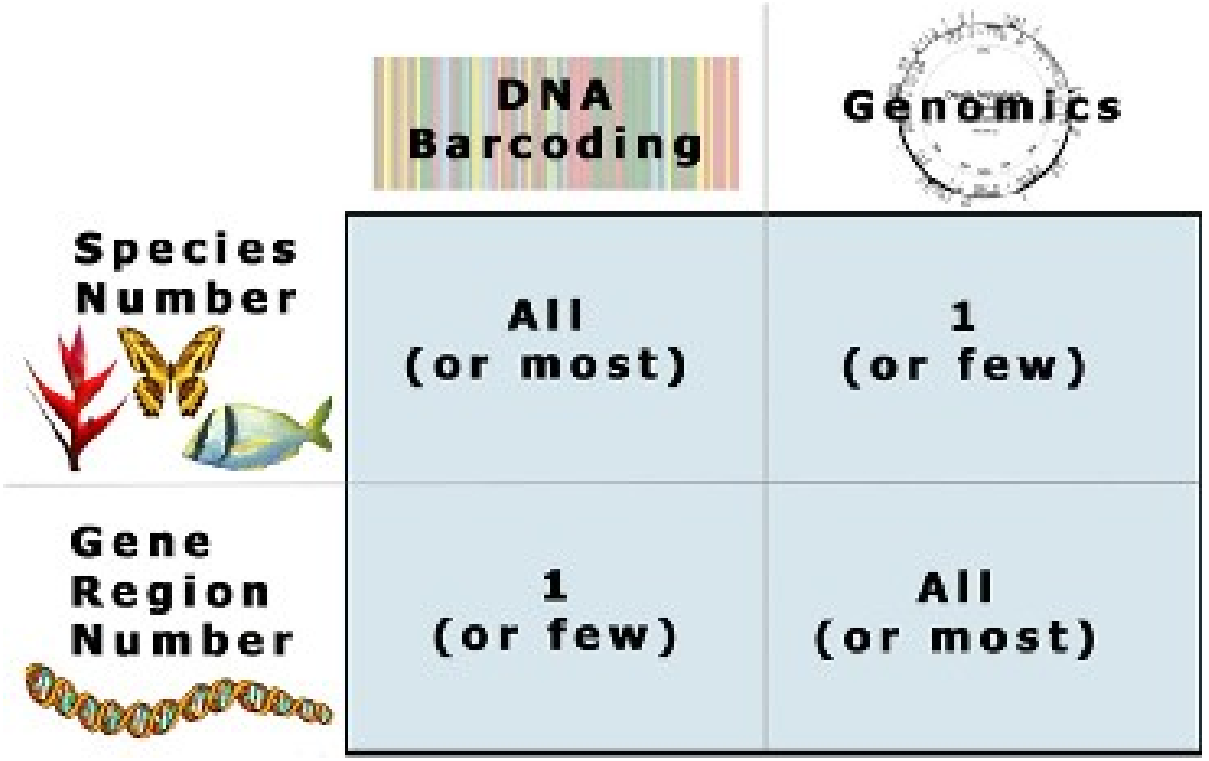
eDNA vs. bulk DNA



After Elbrecht et al. CC-BY 4.0

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Metabarcoding vs. metagenomics



Kress et al. 2008

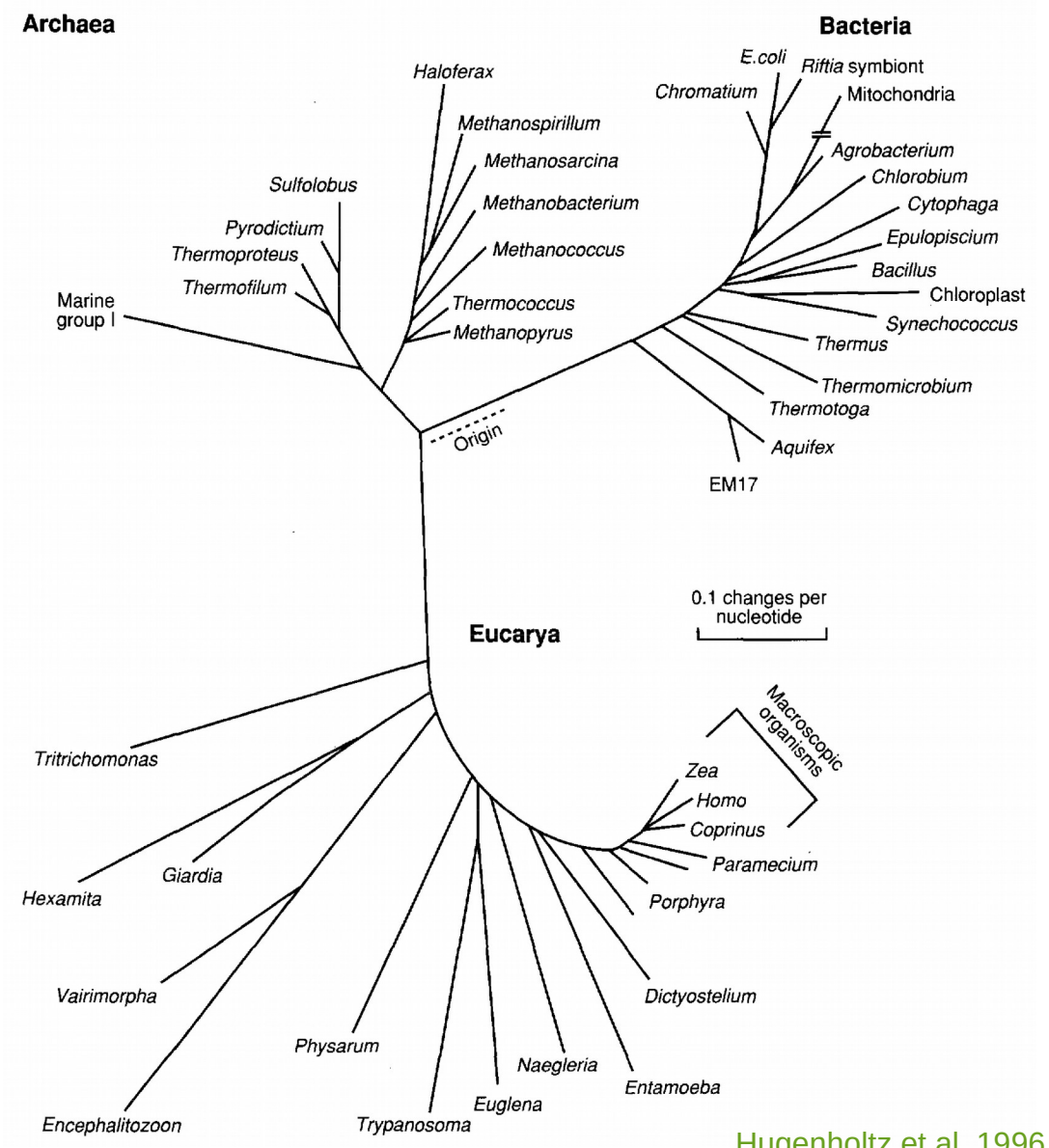
Metabarcoding vs. metagenomics

- Metabarcoding
  - Taxonomic diversity
- Metagenomics
  - Taxonomic & functional diversity

Advantages

- Cheap and relatively quick
- Lab and computer skills instead of taxonomic knowledge
- Independent of culture techniques
- Detection of rare species
- Non-invasive monitoring via eDNA

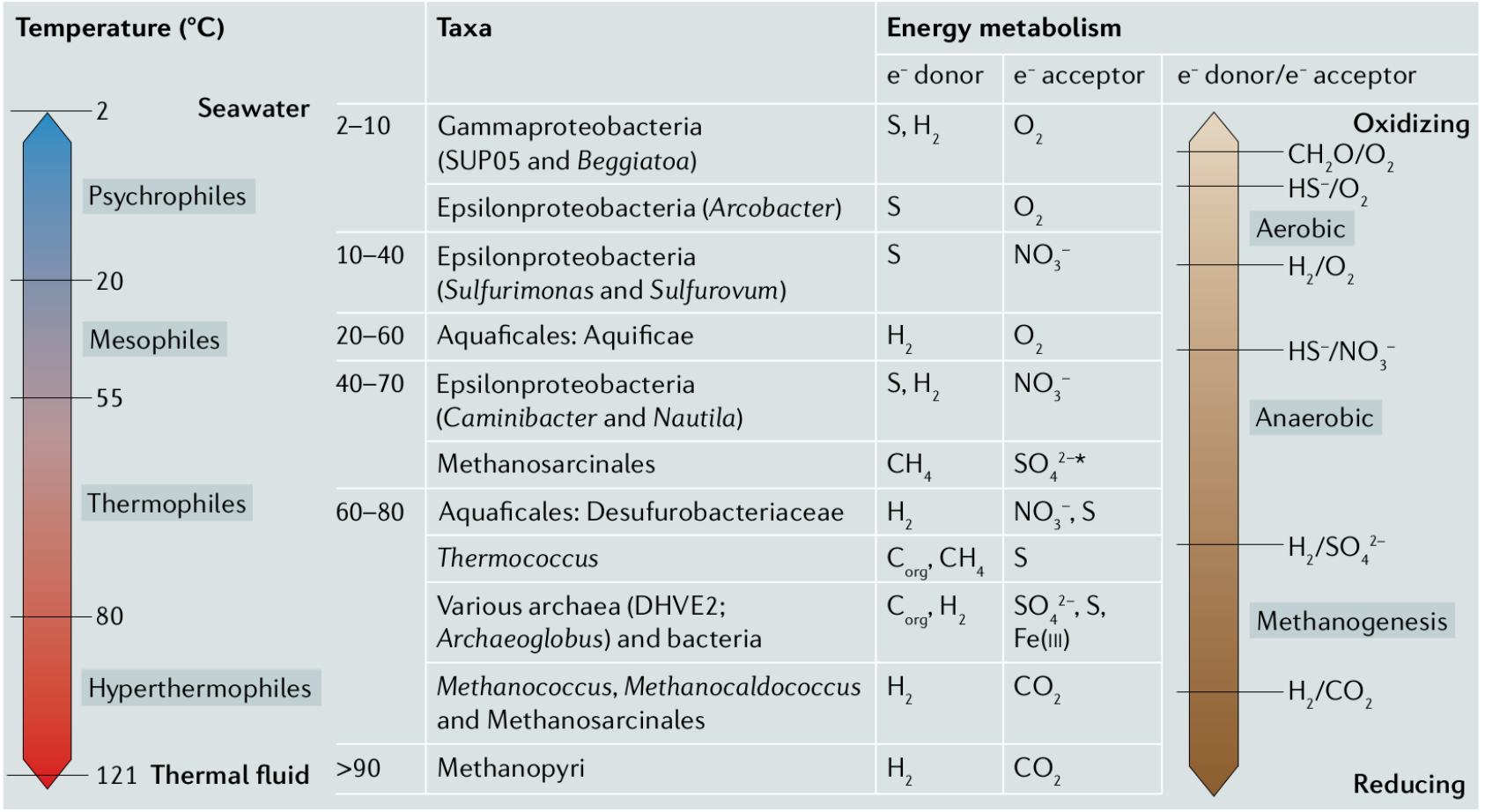
Microbial diversity



Hugenholtz et al. 1996

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Hydrothermal vents

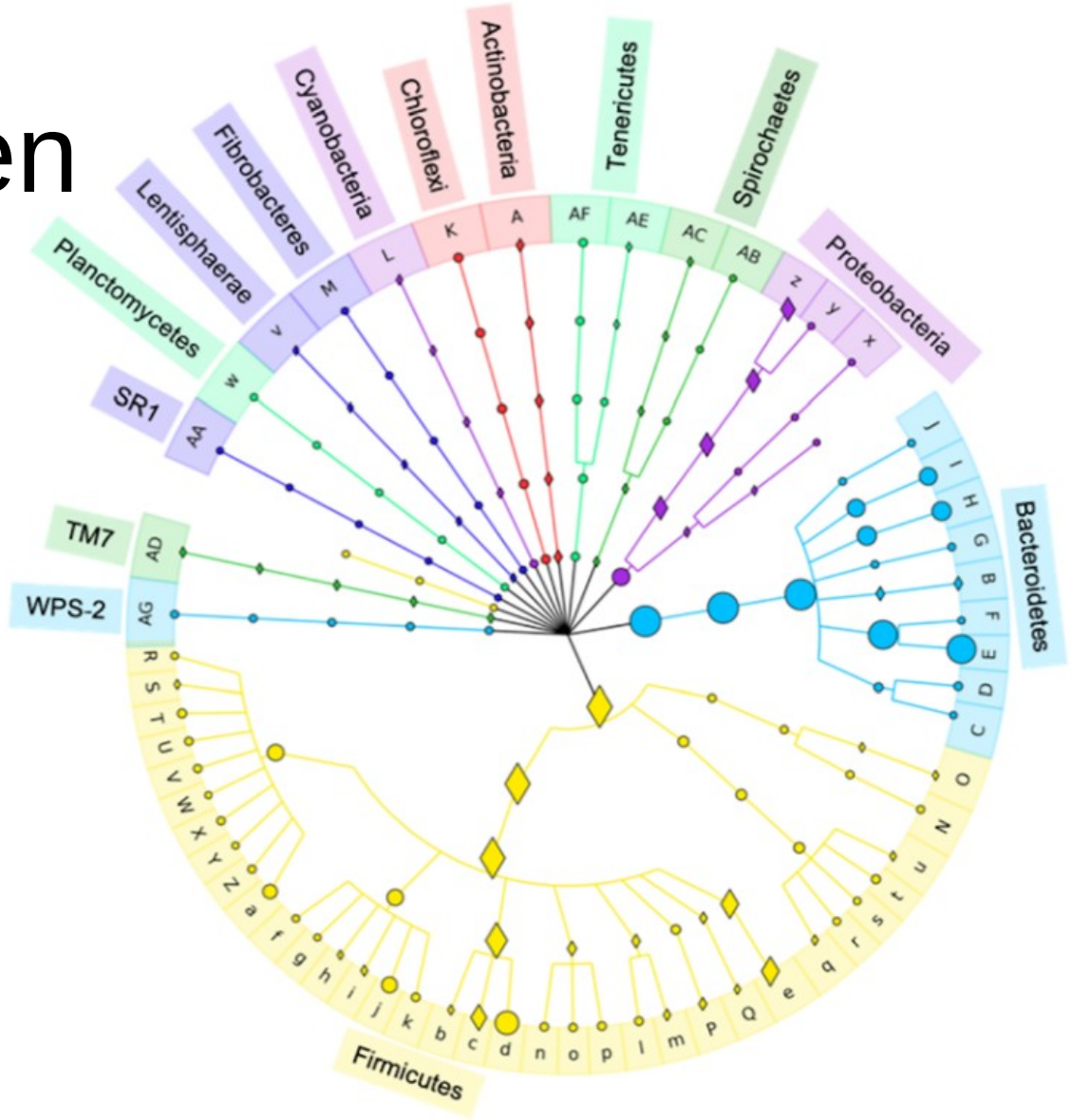


	Taxa	Energy metabolism		
		e⁻ donor	e⁻ acceptor	e⁻ donor/e⁻ acceptor
2–10	Gammaproteobacteria (SUP05 and *Beggiatoa*)	S, H_2	O_2	Oxidizing
	Epsilonproteobacteria (*Arcobacter*)	S	O_2	CH_2O/O_2
10–40	Epsilonproteobacteria (*Sulfurimonas* and *Sulfurovum*)	S	NO_3^-	HS^-/O_2
20–60	Aquificales: Aquificae	H_2	O_2	Aerobic
40–70	Epsilonproteobacteria (*Caminibacter* and *Nautila*)	S, H_2	NO_3^-	H_2/O_2
	Methanosarcinales	CH_4	SO_4^{2-}*	HS^-/NO_3^-
60–80	Aquificales: Desulfurobacteriaceae	H_2	NO_3^-, S	Anaerobic
	Thermococcus	C_{org}, CH_4	S	H_2/SO_4^{2-}
	Various archaea (DHVE2; *Archaeoglobus*) and bacteria	C_{org}, H_2	SO_4^{2-}, S, Fe(III)	Methanogenesis
	Methanococcus, *Methanocaldococcus* and Methanosarcinales	H_2	CO_2	H_2/CO_2
>90	Methanopyri	H_2	CO_2	Reducing

Dick 2019

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

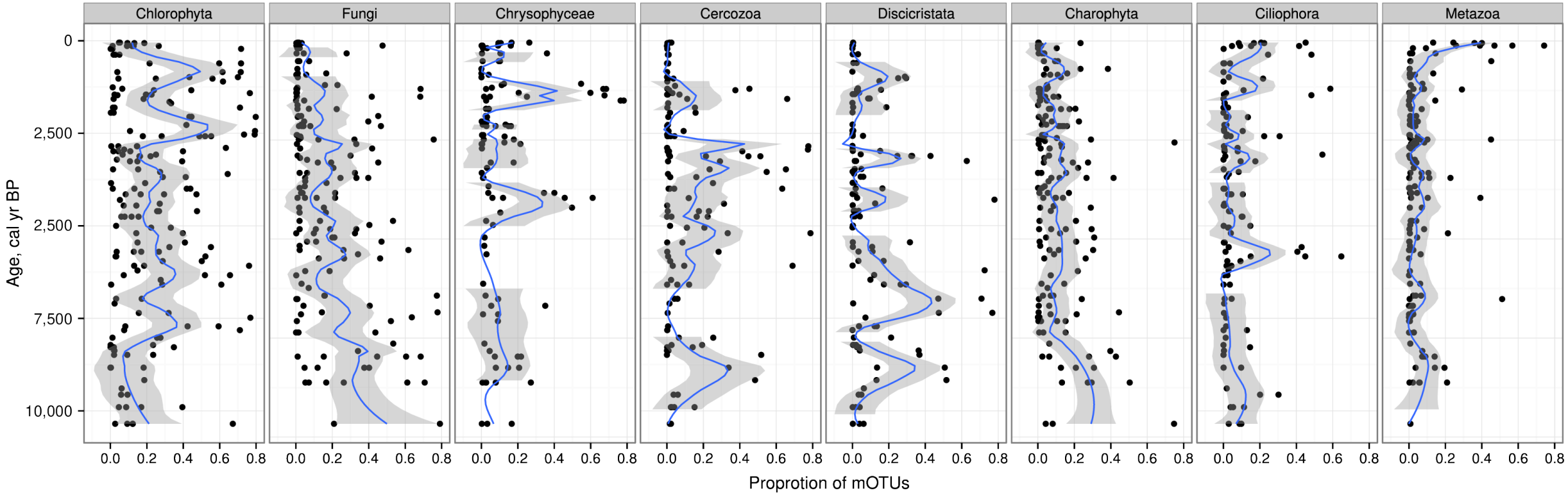
Cow rumen



Li et al. 2019

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

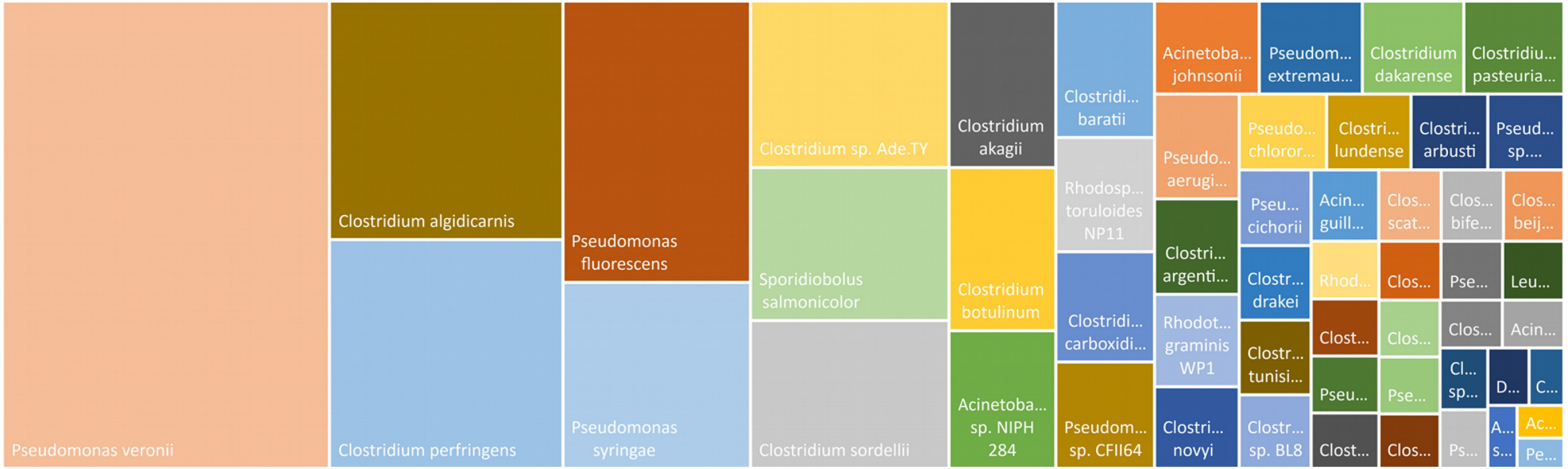
Reconstruction of palaeo ecosystems



Kisand et al. 2018

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Tyrolean Iceman



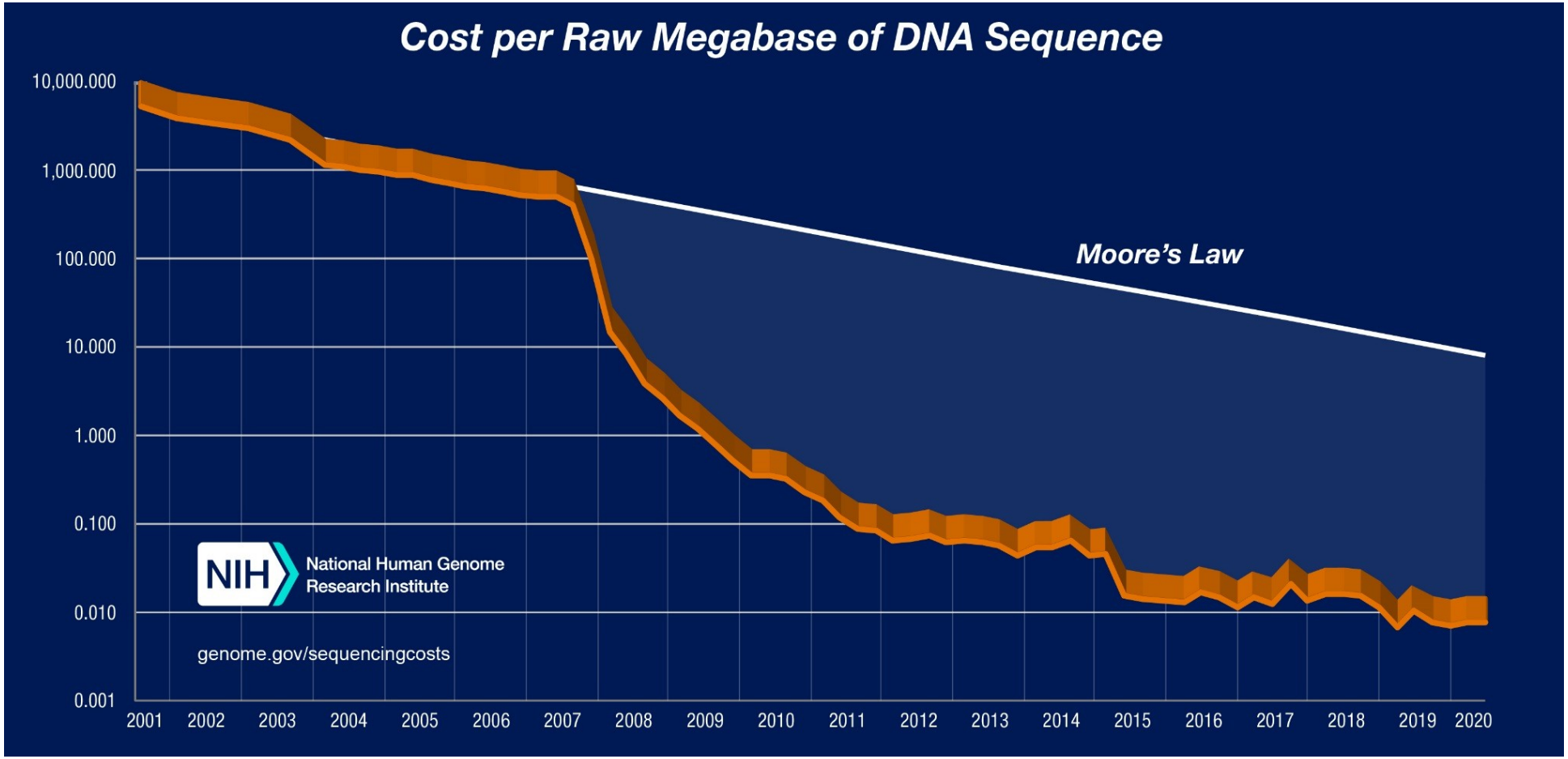
Lugli et al. 2017

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Sequencing platforms: History

- First Generation
  - 1977 Sanger Sequencing (GOLD Standard)
- Second Generation
  - 2005 454 Sequencing (now Roche, end of life 2015)
  - 2005 Solexa Sequencing (now Illumina)
  - 2006 APG SOLiD Sequencing (now Life Technologies, marginal)
- Third Generation
  - 2008 Pacific Biosciences (acquired by Illumina 12/2018)
  - 2011 Ion Torrent (now Life Technologies)
  - 2014 Oxford Nanopore

Sequencing platforms: Cost



NHGRI

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

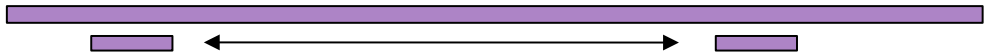
Sequencing: Terms

- Library: Pool of DNA fragments generated from a sample using a specific protocol
- Run: Sequencing of one or more libraries
- Read: A DNA sequence generated in a sequencing run (one of many, see yield)
- Read length: Length of a read (specified in basepairs [bps])
- Yield: Total number of bps generated in a run (in Gbp); Number of reads times the average read length
- Insert size: length of a sequenced DNA fragment

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

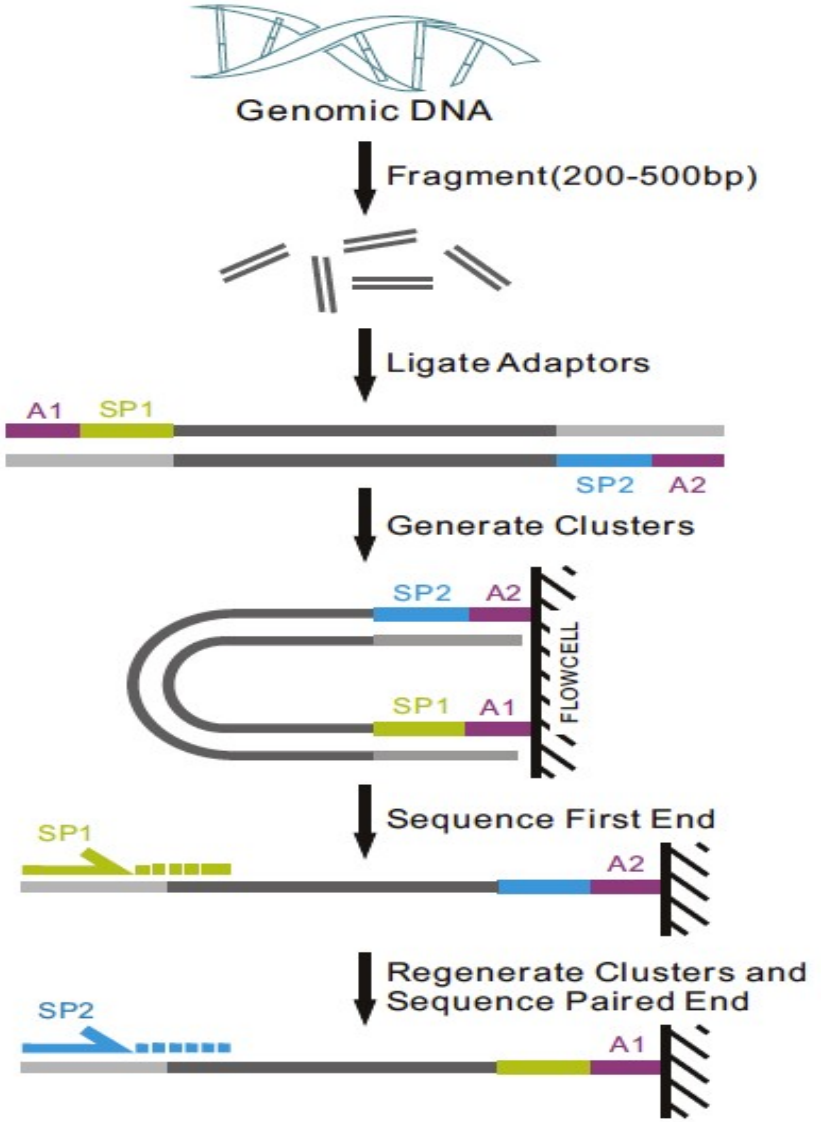
Sequencing: Terms

- Paired-end (mate-pair) reads: two reads representing the ends of a DNA fragments



```
@HWI-EAS225_90320:3:2:1339:1667/1
ATGGCTGAAGTACGGCGACAAGCGCGTCATCGACAC
+
GGGGFGACBB?BBCBBBB?BC=BA;;9?A=>A>:??
```

```
@HWI-EAS225_90320:3:2:1339:1667/2
ATGGCTGAAGTACGGCGACAAGCGCGTCATCGACAC
+
GGGGFGACBB?BBCBBBB?BC=BA;;9?A=>A>:??
```

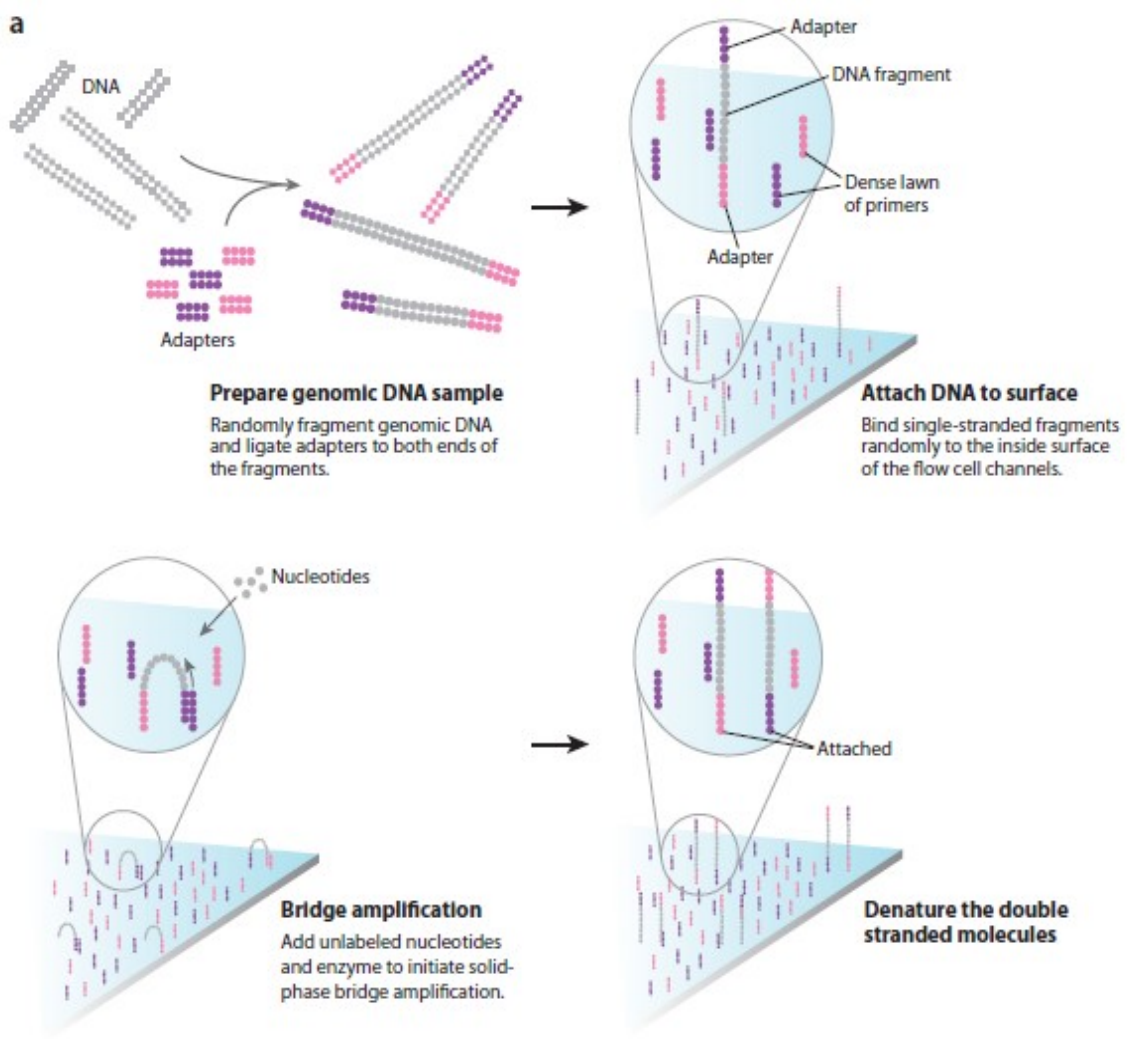


Illumina

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

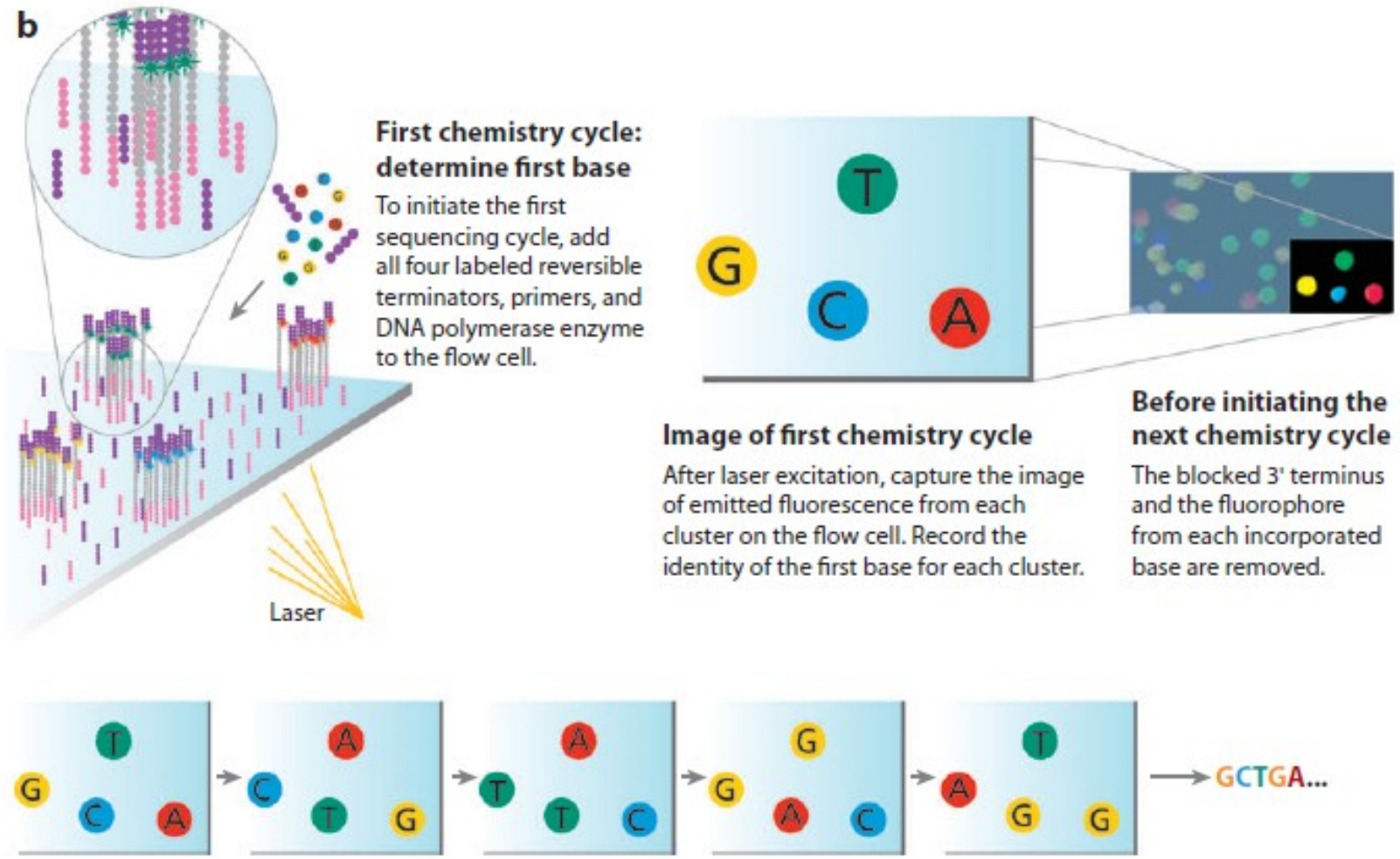
Sequencing platforms: Illumina

- Library generation with bridge PCR
- Read generation with polymerase and “reversible terminators
- Read length 50-350 bps
- Yield 12 – 3,000 Gbps
- Error rate ~ 0.1%



Bentley et al. 2008; Mardis 2008

Sequencing platforms: Illumina

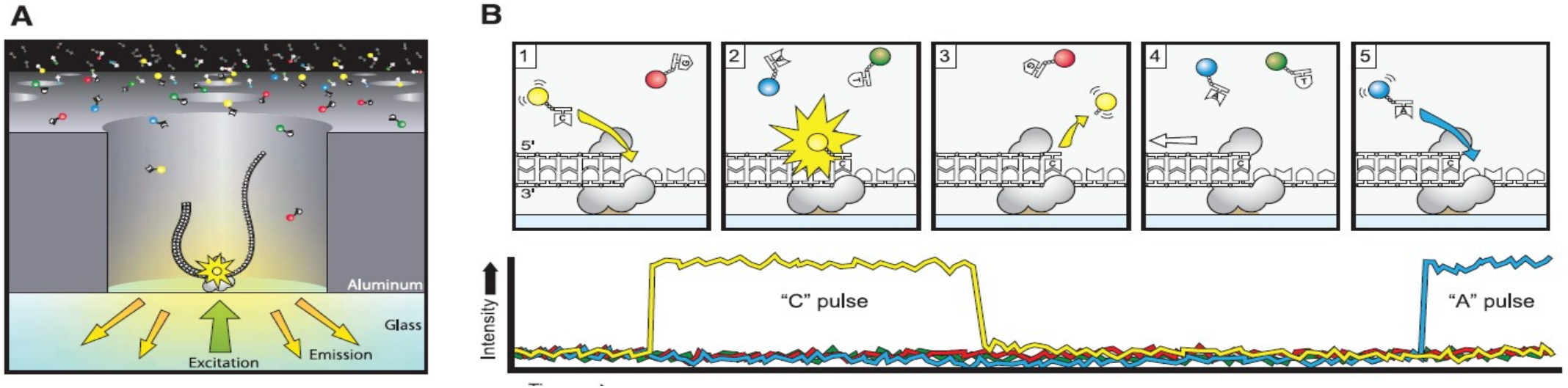


Illumina

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Sequencing platforms: Pacific Biosciences

- “Single molecule sequencing”, no amplification required
- „Realtime sequencing“, bases are determined as polymerase integrates them
- Polymerase speed can be measured, delays can be utilized to detect DNA modifications (e.g. methylation)

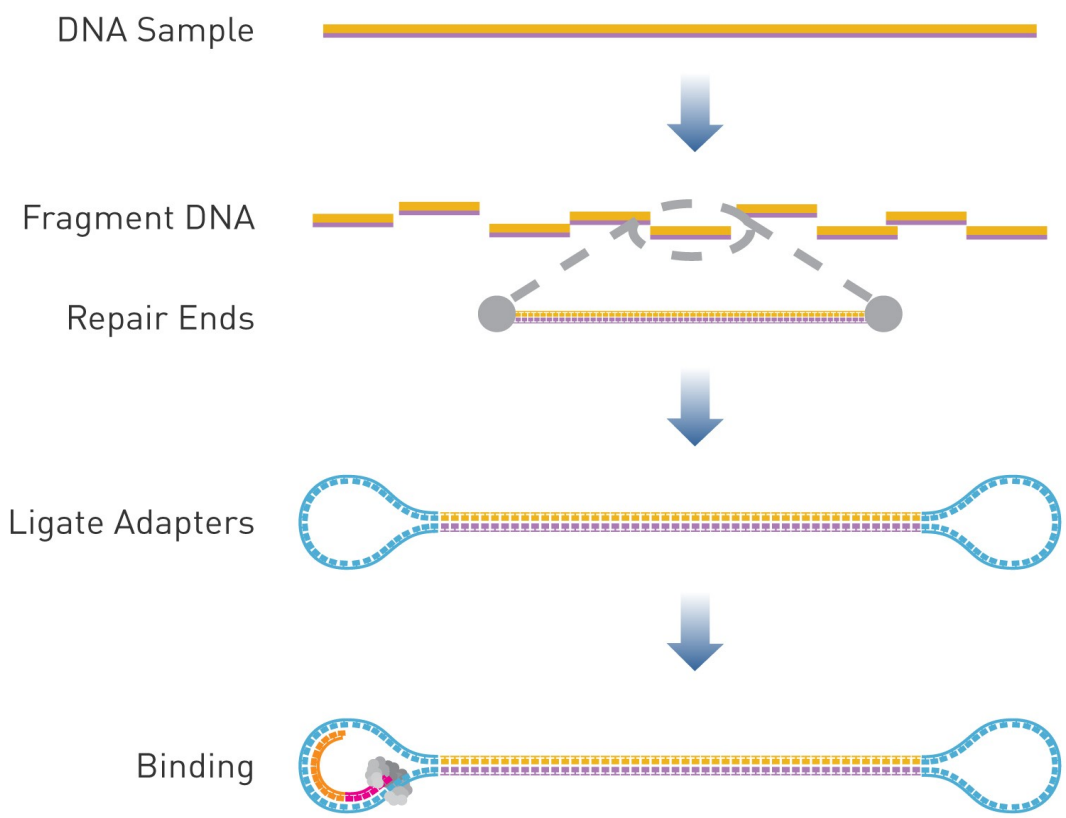


Eid et al. 2009

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Sequencing platforms: Pacific Biosciences

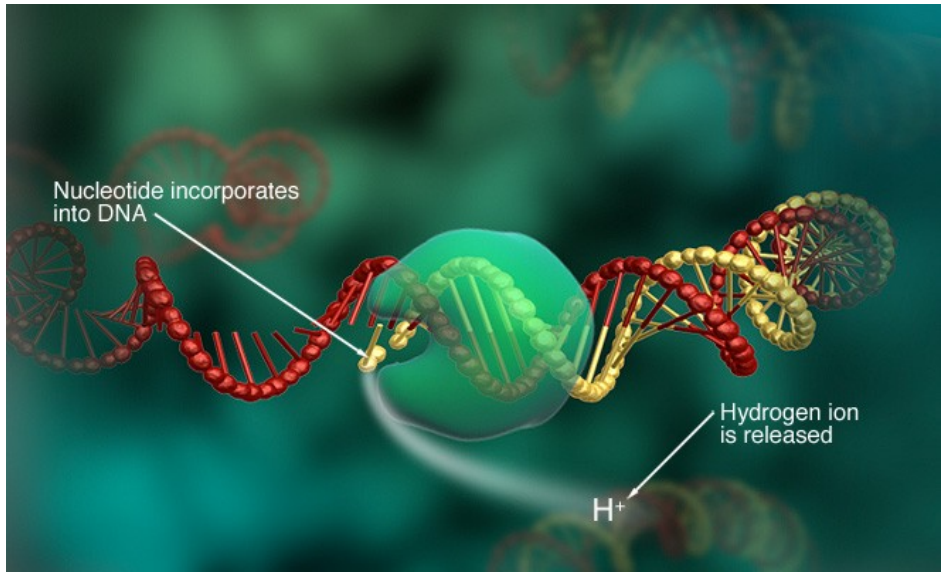
- Read length up to 100 kbps, on average 30 kbps,
  yield: 0.8 - 250 Gbps,
  high error rate: ~ 15 % (mainly insertions and deletions)
  runtime: 1 - 20 h
- „Circular sequencing“, DNA sequence is read several times to improve accuracy
  → build circular consensus sequence (CCS/HiFi)
  ~ 0.01 % error, 16 Gbps yield

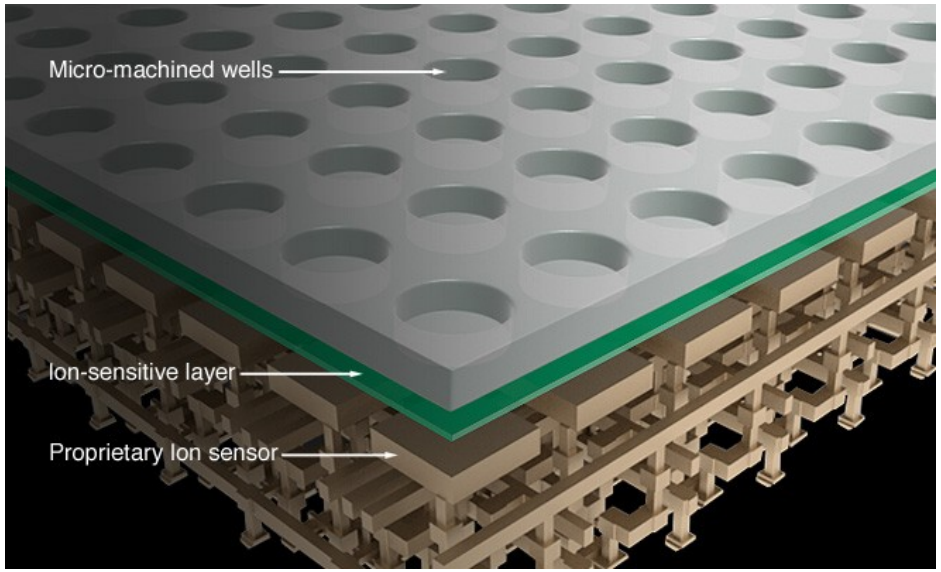


Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Sequencing platforms: Ion Torrent / Proton

- Library generation with emulsion PCR
- Read generation via “pH sensing”
- Read length up to 300 bps; Yield ~ 2 Gbps
- Relatively low error rate: 1 % (mainly homopolymers)



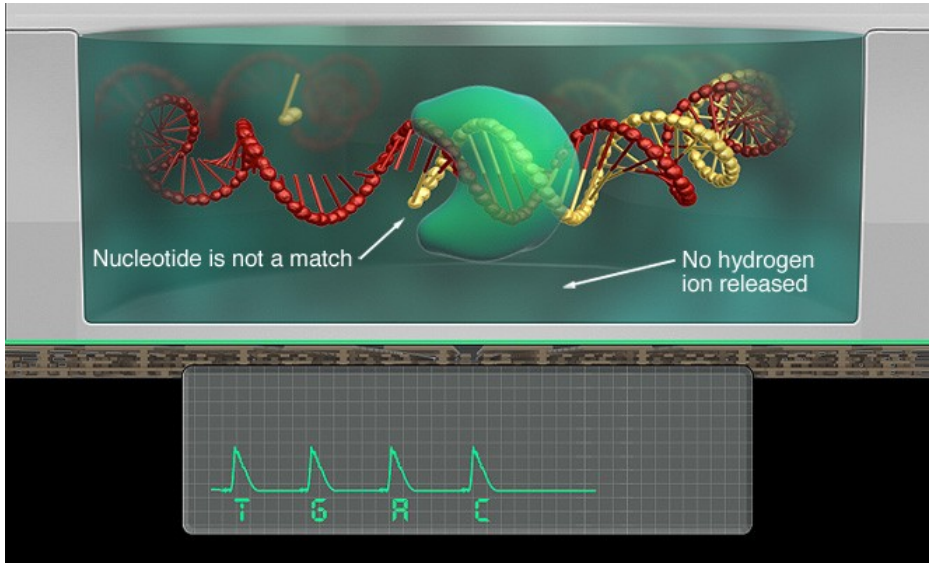


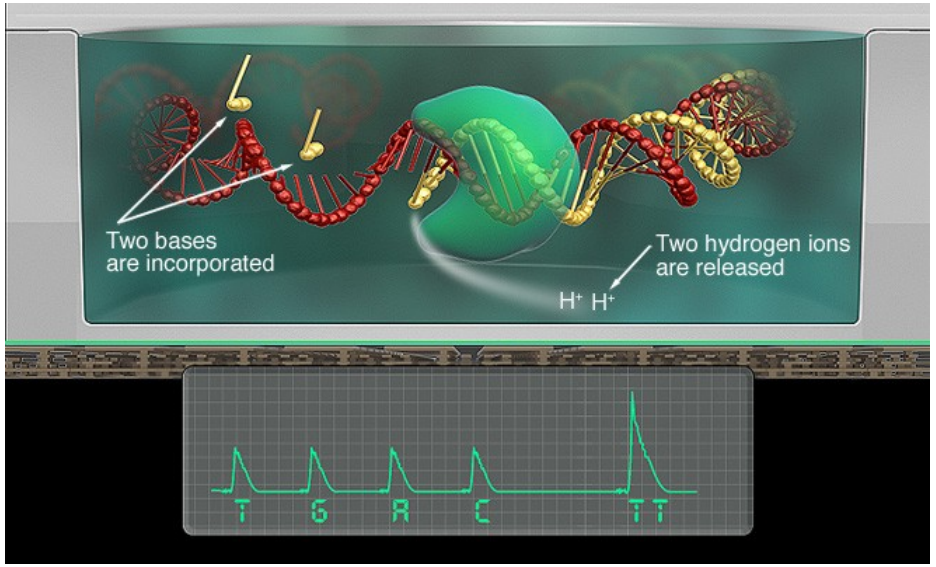
Rothberg et al. 2011

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Sequencing platforms: Ion Torrent / Proton

- Sequencing in cycles with a defined order of nucleotides
- The change in pH of a well is detected
- Well density determines yield



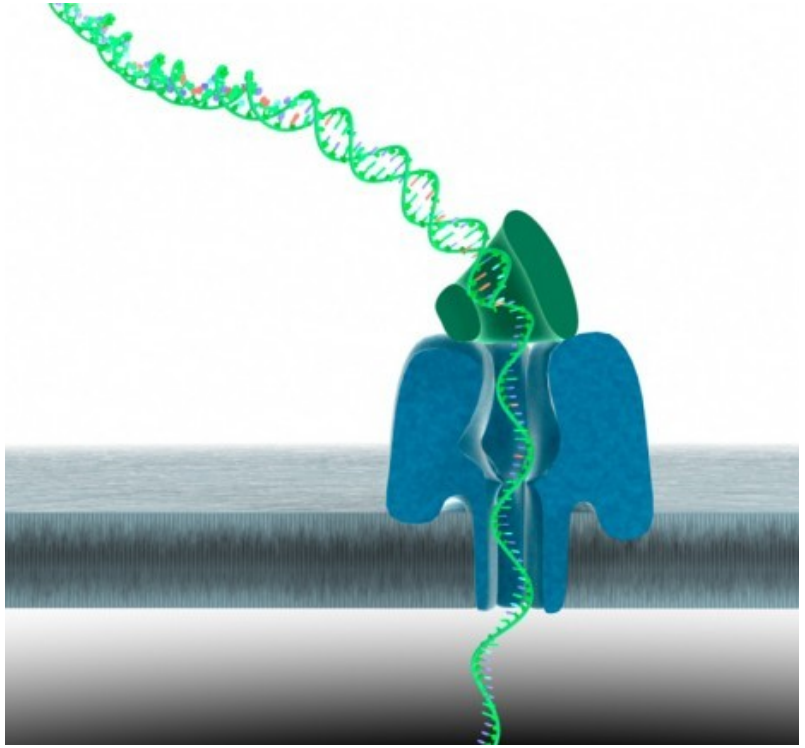


Rothberg et al. 2011

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Sequencing platforms: Oxford Nanopore

- Single DNA strand runs through a nanopore
- Voltage change per nucleotide / nucleotide set (≤ 5) is measured
- Number of pores determines yield
- Read length: up to 4,000 kbps,
  10 kbps on average
- Yield: 1 Gbps (Flongle)
  12 Gbps (Minion)
- High error rate: ~ 15 %



Clarke et al. 2009; Lavera et al. 2015

Vimeo Nanopore sequencing animation

FASTA format

- File format for nucleotide or amino acid sequences
- Two elements: sequence description and sequence
  - Sequence description: single line, starting with “>”
  - Sequence: one or more lines

```
>Saccharomyces_cerevisiae_18S_1800
TTCATAATAACTTTTCGAATCGCATGGCCTTGTGCTGGCGACTTGTGCTGGCGAT
GGTTCATTCAAATTTCTGCCCTATCAACTTTCGATGGTAGGACTTGTGCTGGCGA

>lcl|XM_028624329.1_prot_XP_028475401.1_1|EHS24_009034]
MPDGTQKGSALVLPTAFSAATQINTLLNPSVLIIGAGIGGITLALDLDEKGLTNWL
LVDREDDVGGTWYVNRYPGCRCDIPAIGYSHSRFQNSQWTETHPDHKEIQAYWAKI
```


Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

FASTQ format

- Extension of FASTA format
- Includes quality scores, for nucleotide sequences only
- Description starts with “@”, scores follow after a line with “+”
- Score range may differ between/within technologies

```
@m54067_210309_060726/4260089/ccs
TGGATCACTTGTGCAAGCATCACATCGTAGCCTTCCGCAGGTTCACCTACGGA
GACCTTGTTACGACTTCTCCTTCCTCTAG
+
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~7~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

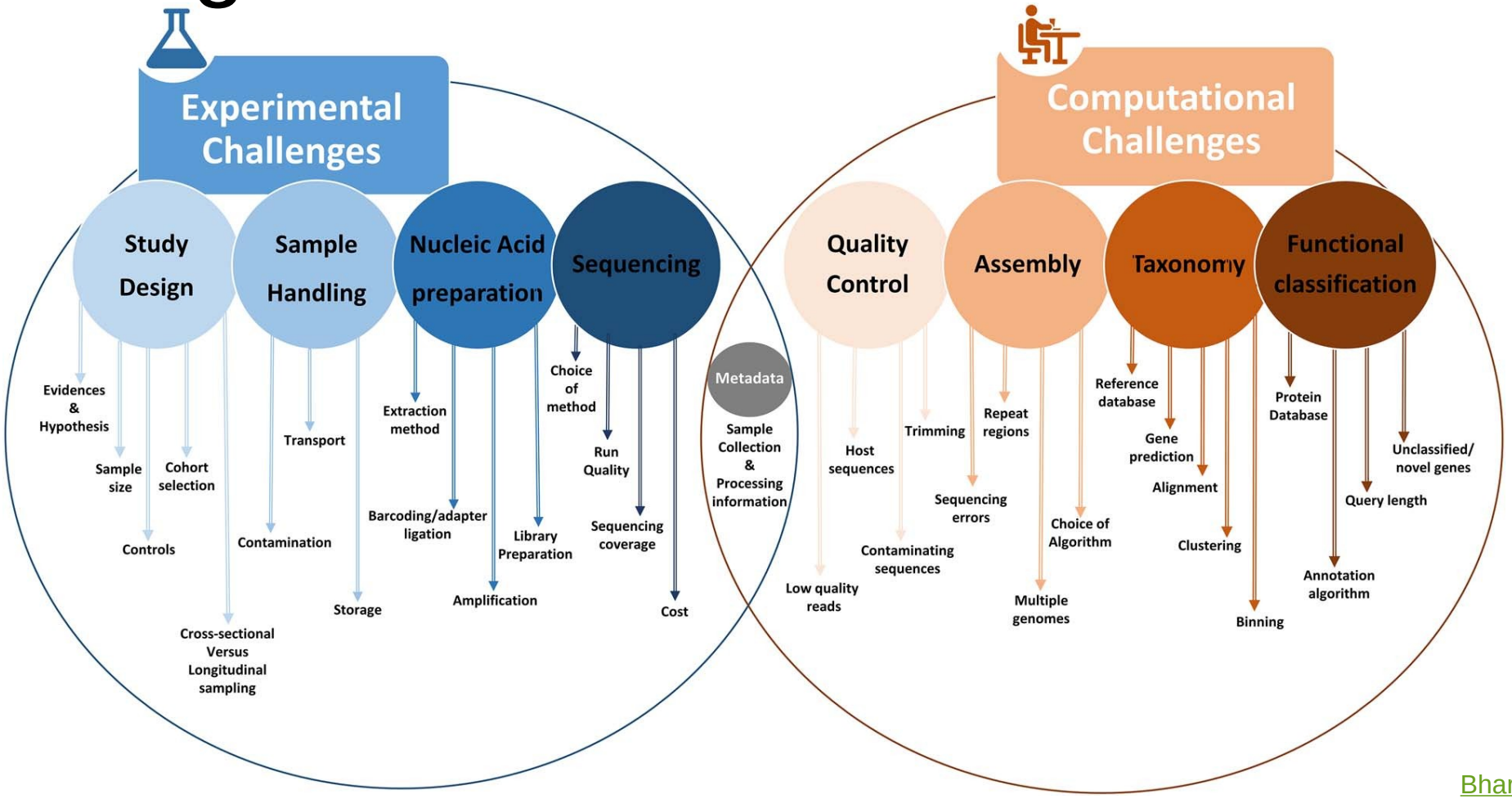

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Reference databases

- RDP (Ribosomal Database Project, 16S)
- SILVA (from Latin *silva* – forest, 16S, 18S)
- GreenGenes (16S, obsolete)
- GTDB (Genome Taxonomy Database, 16S)

- Unite (ITS)
- MycoBank (ITS)

Challenges



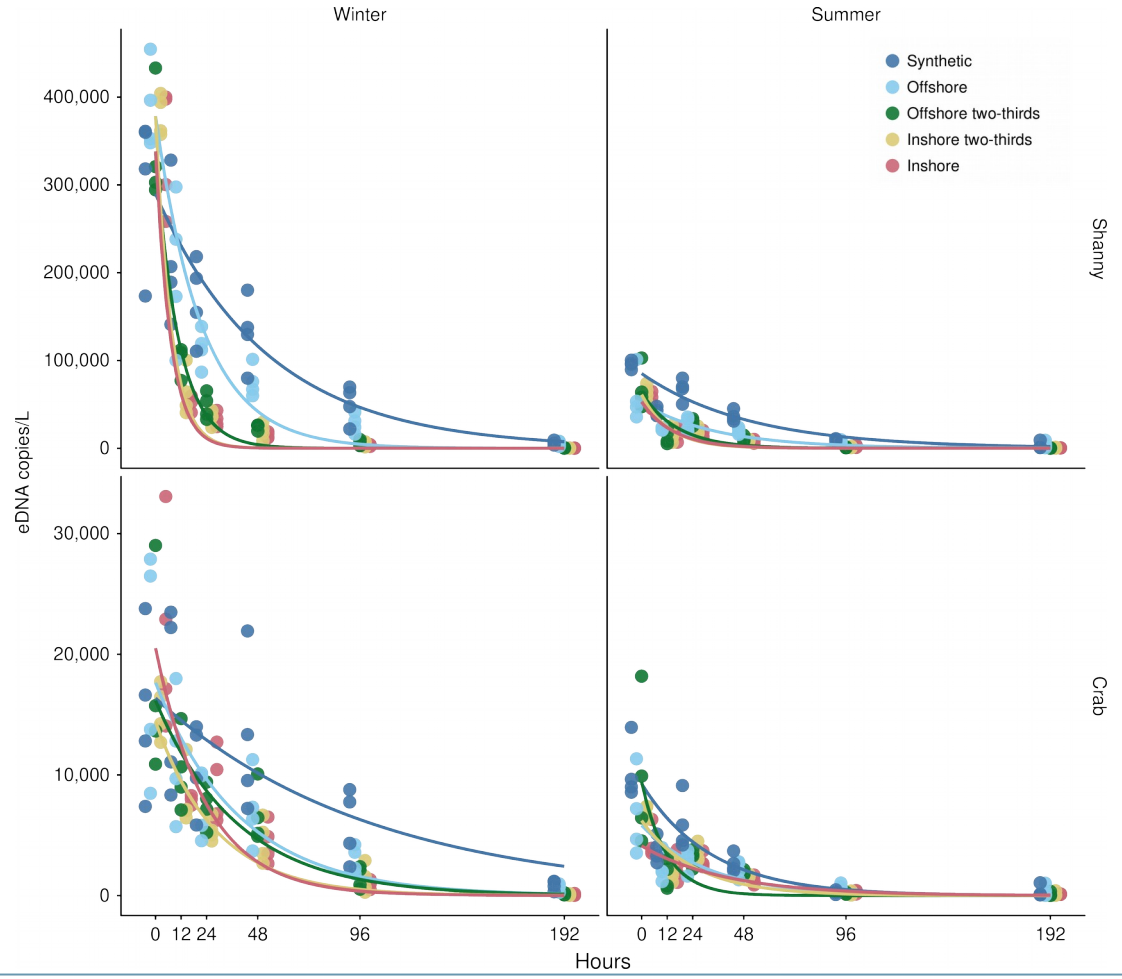
Bharti et al. 2021

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Challenges

- Sampled time frame often uncertain
- The question of abundances
- Species delimitation
- Interspecific genetic variation
- Non-standardized process (sampling, DNA extraction, sequencing platforms, primers…)
- Sequencing errors

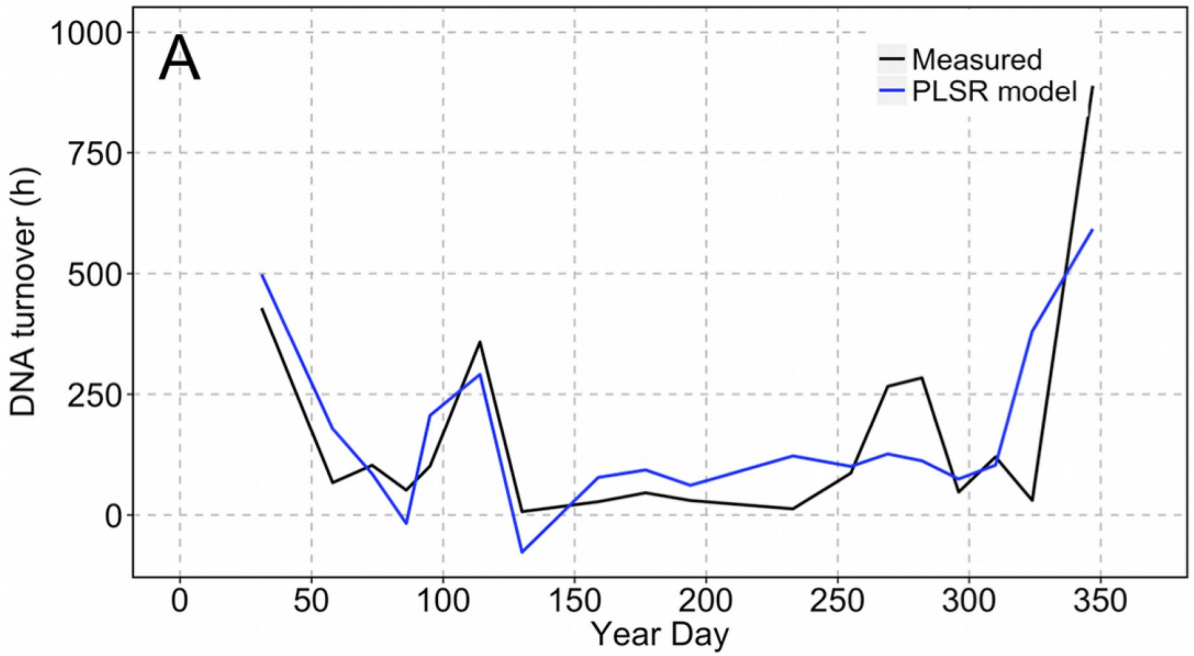
Stability of DNA in environment

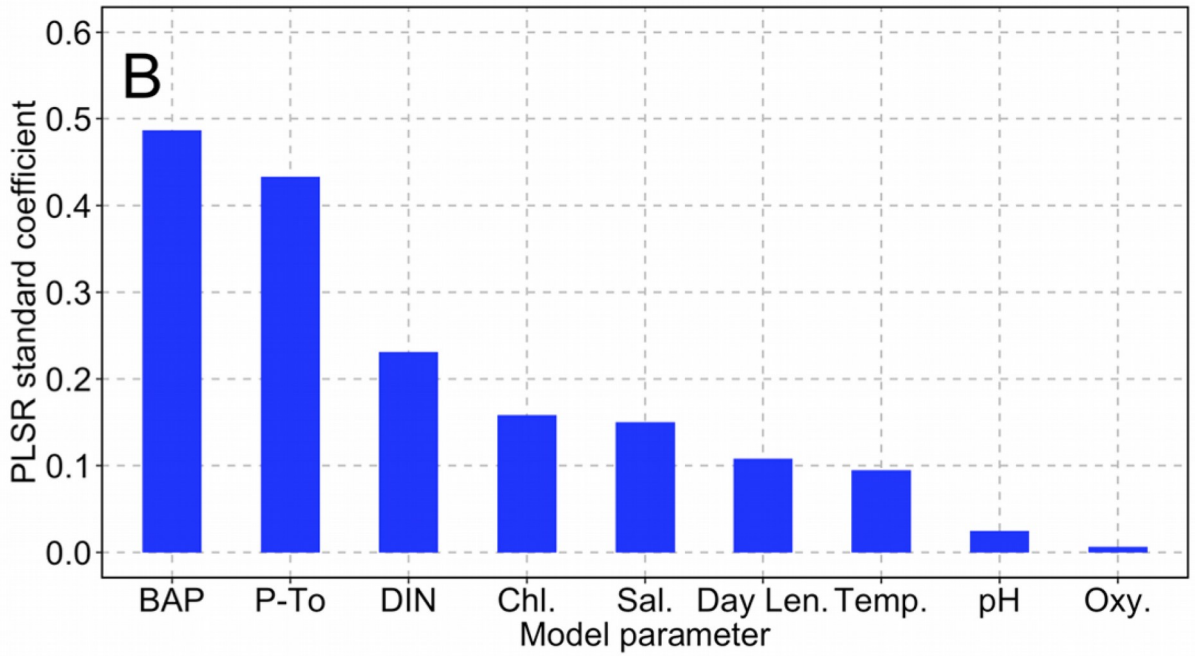


Collins et al. 2018

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Stability of DNA in environment



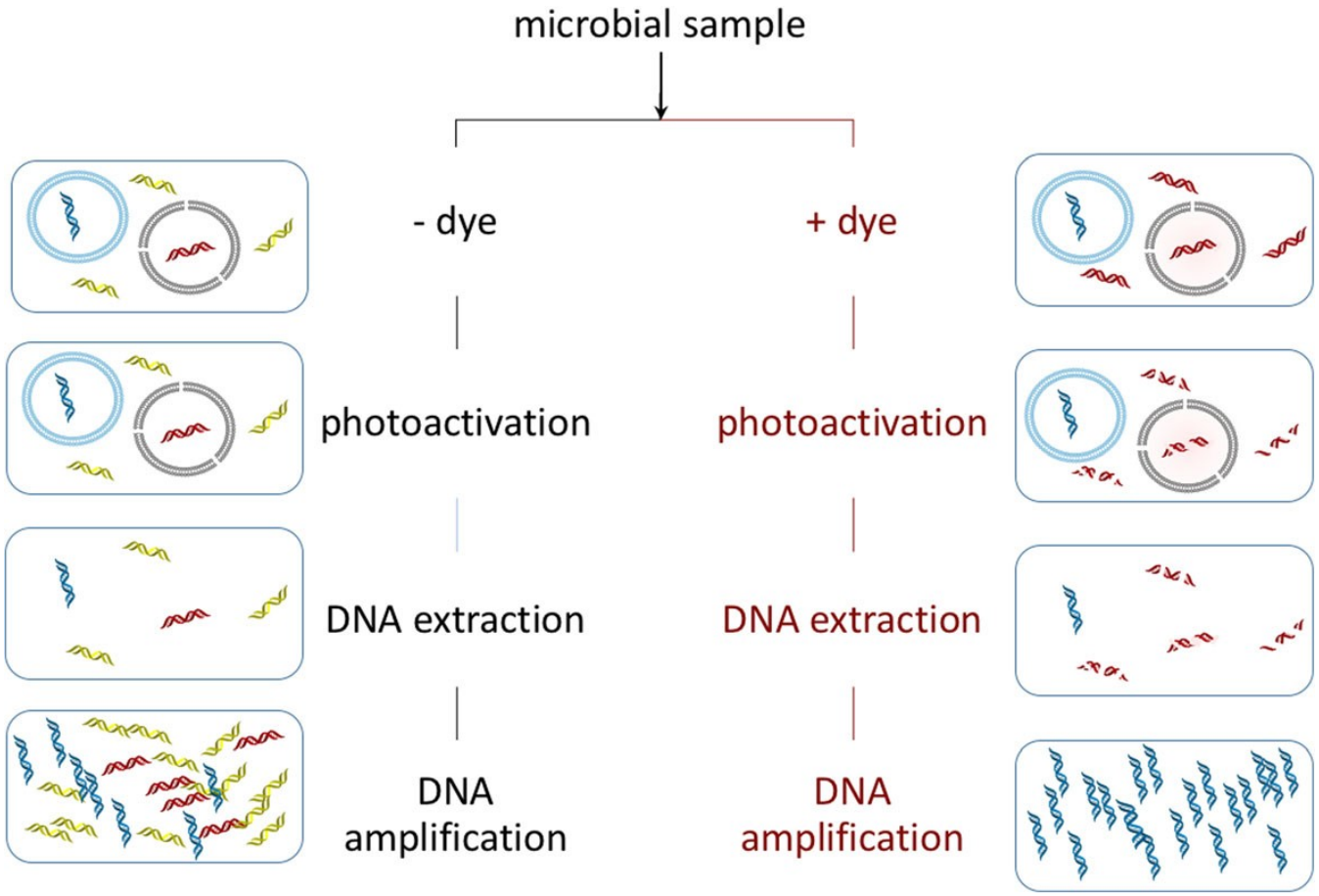


BAP - biologically available phosphate (k + Sn); **P-T0** - phosphate turnover; **DIN** - dissolved inorganic nitrogen; **Chl.** - chlorophyll; **Sal.** - salinity; **Day Len.** - day length; **Temp.** - temperature; **Oxy** - dissolved oxygen.

Salter 2018

Schrödinger's microbes: viability PCR

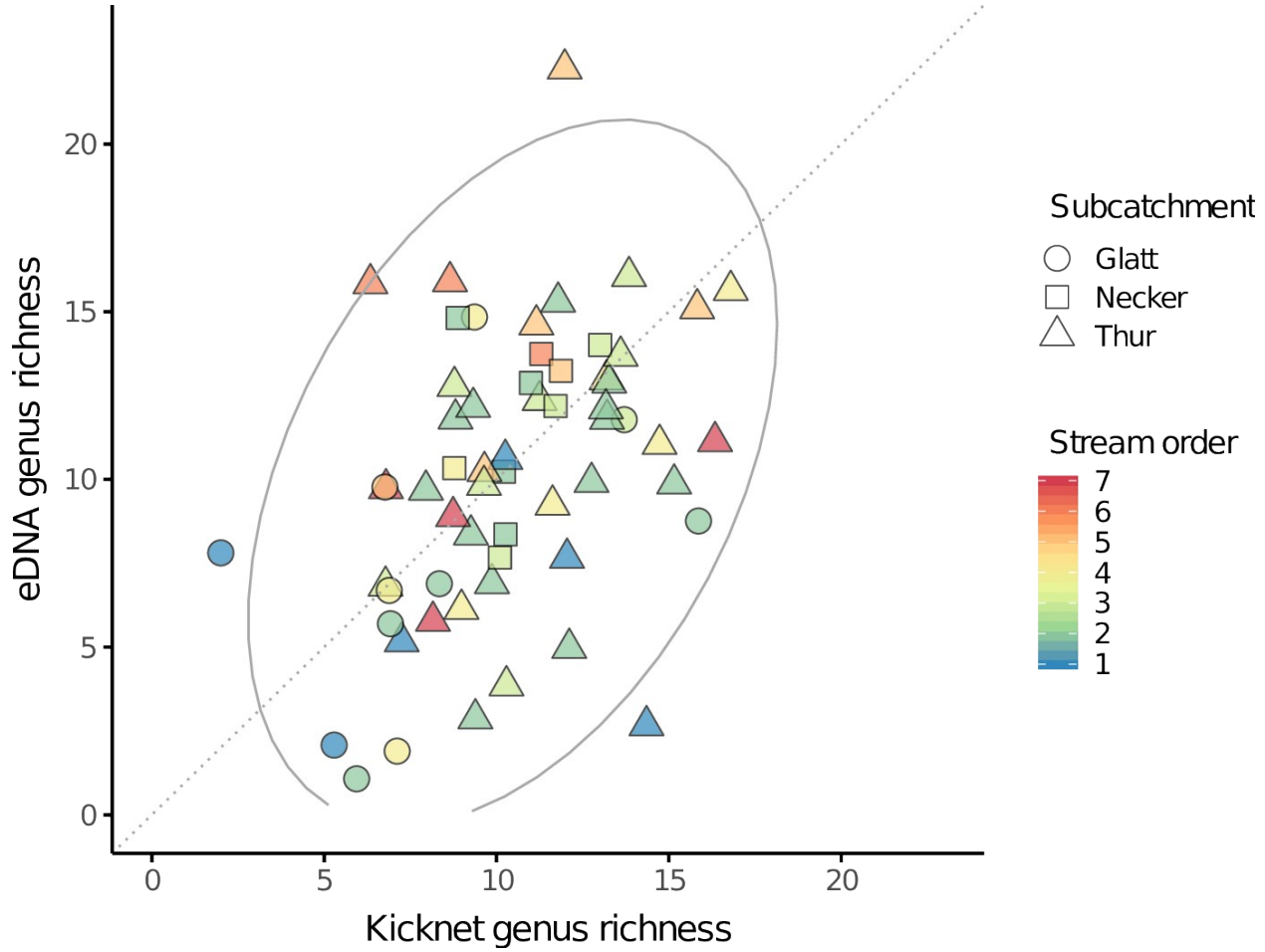
- Ethidium monoazide (EMA) or propidium monoazide (PMA) treatment



Emerson et al. 2017

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

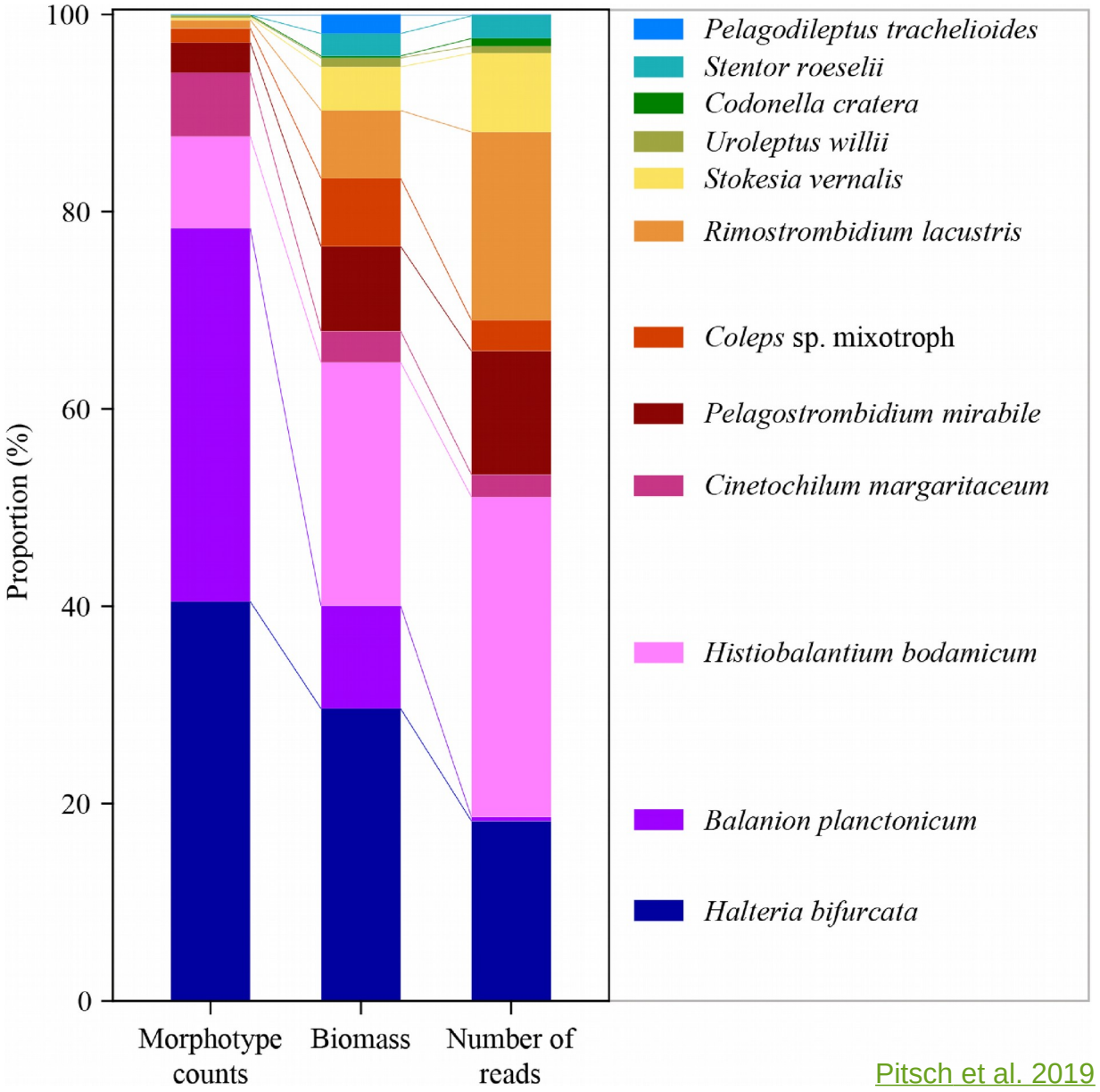
Community recovery



Mächler et al. 2019

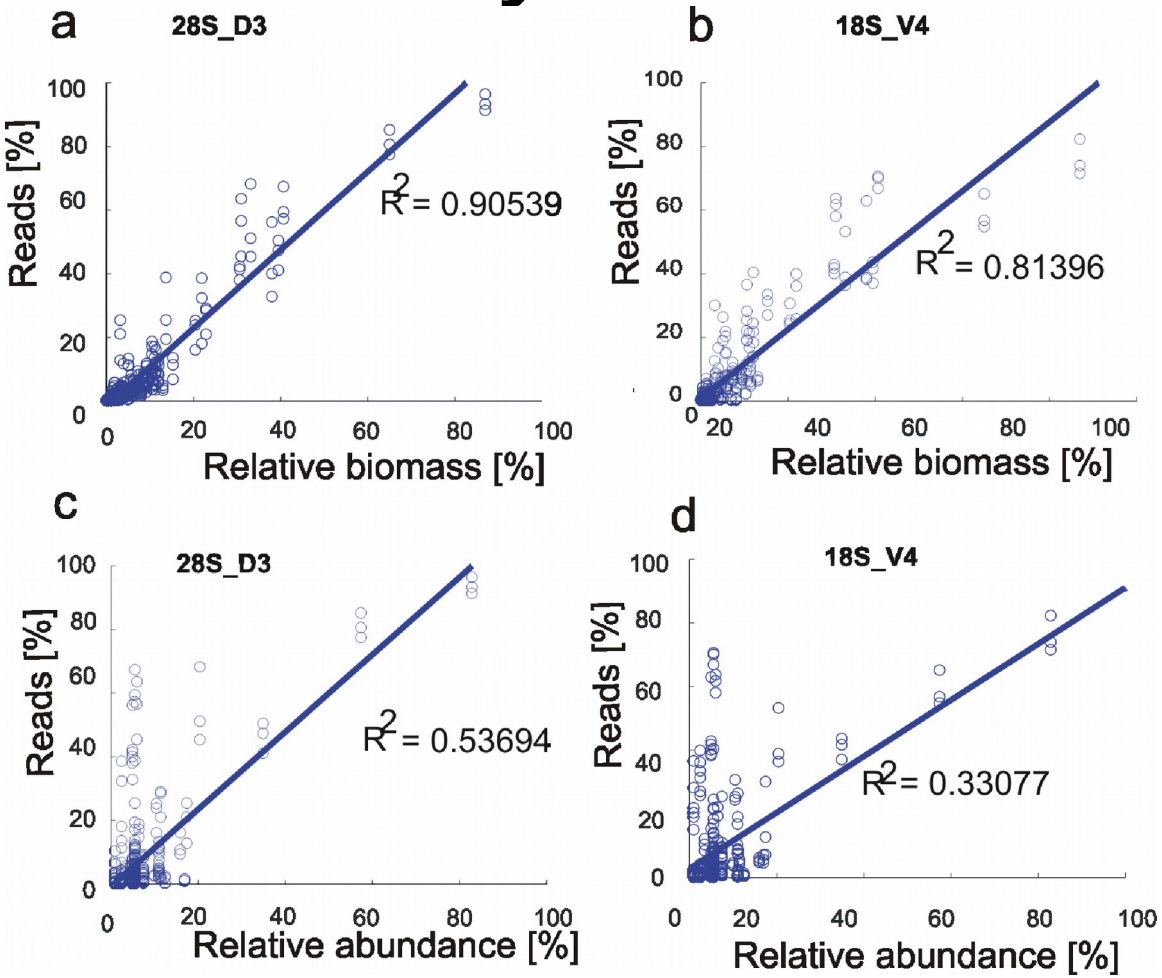
Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Counts – biomass – reads



Pitsch et al. 2019

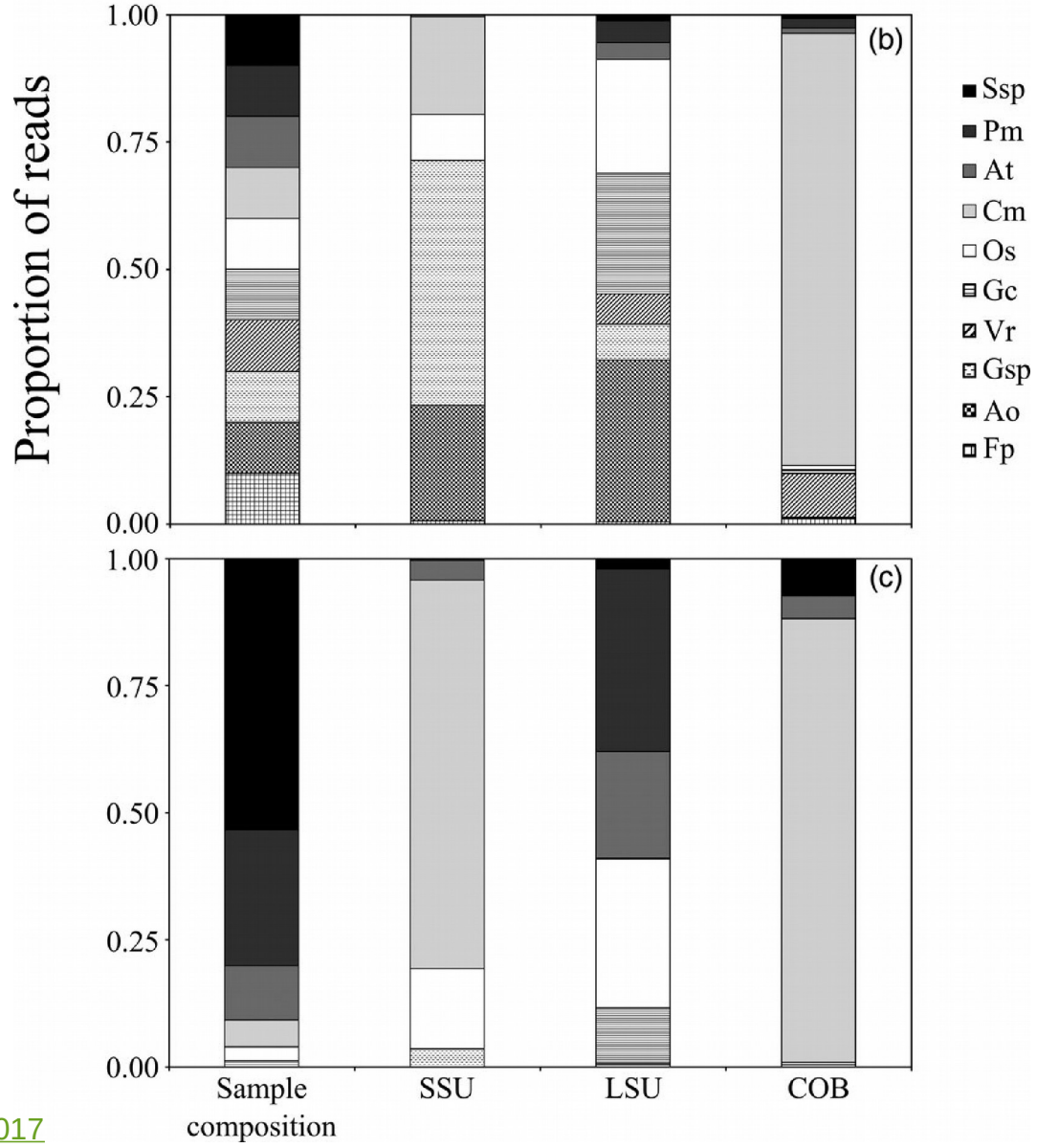
Biomass recovery



Schenk et al. 2019

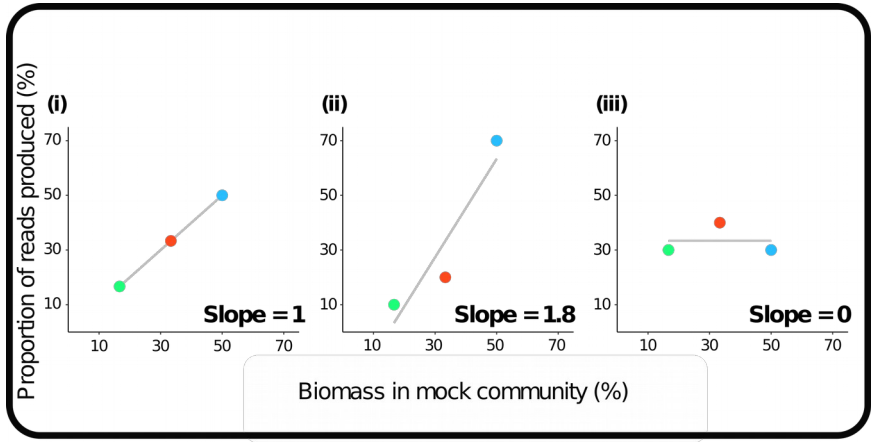
Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

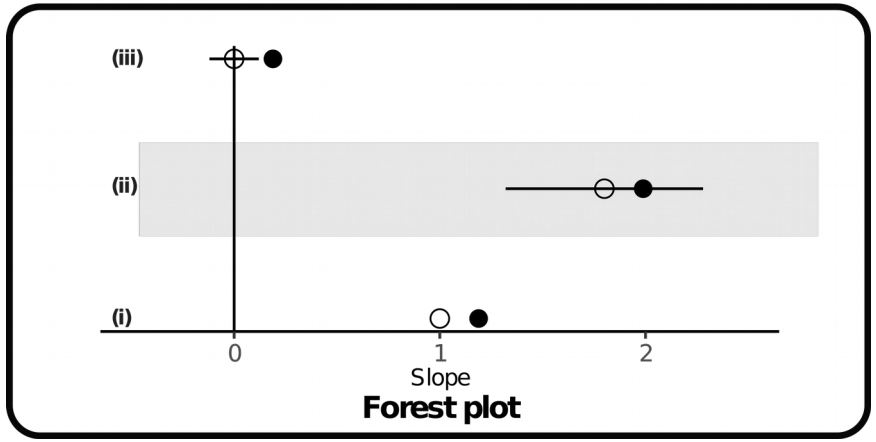
Abundance recovery

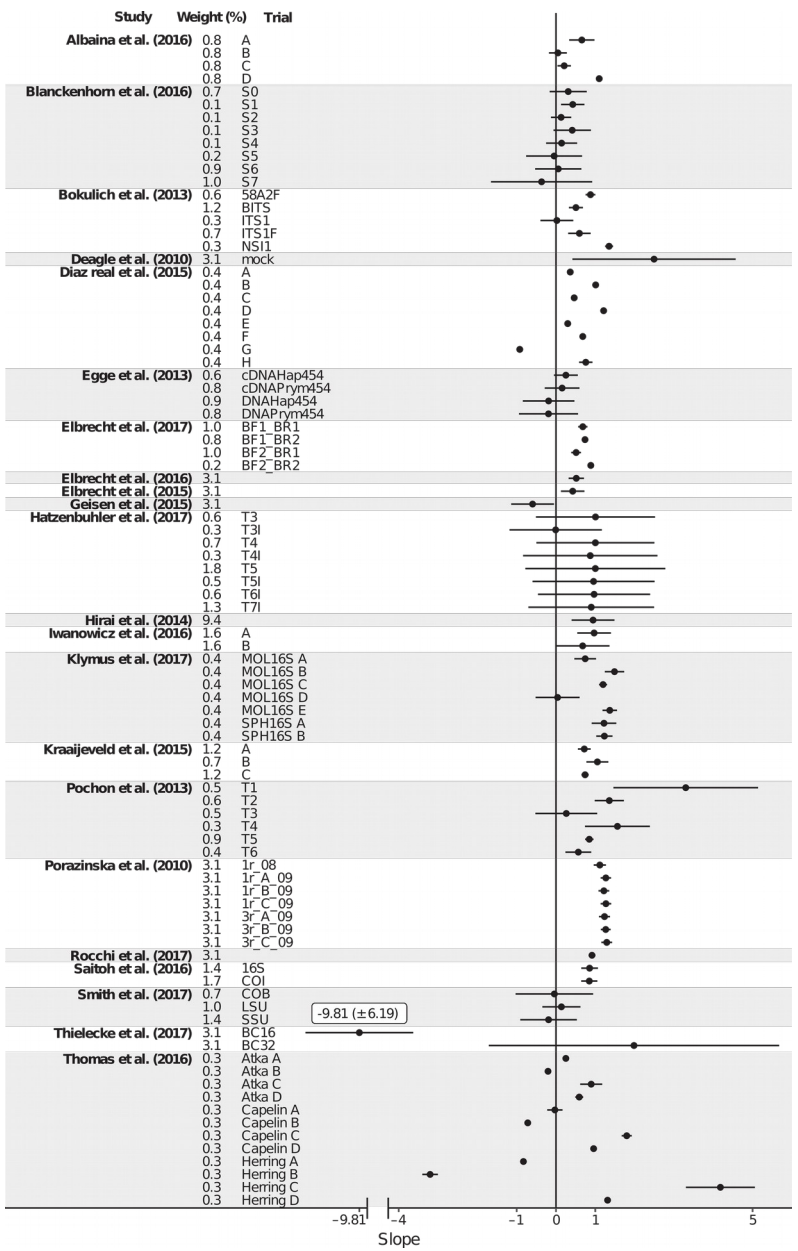


Smith et al. 2017

Abundance recovery



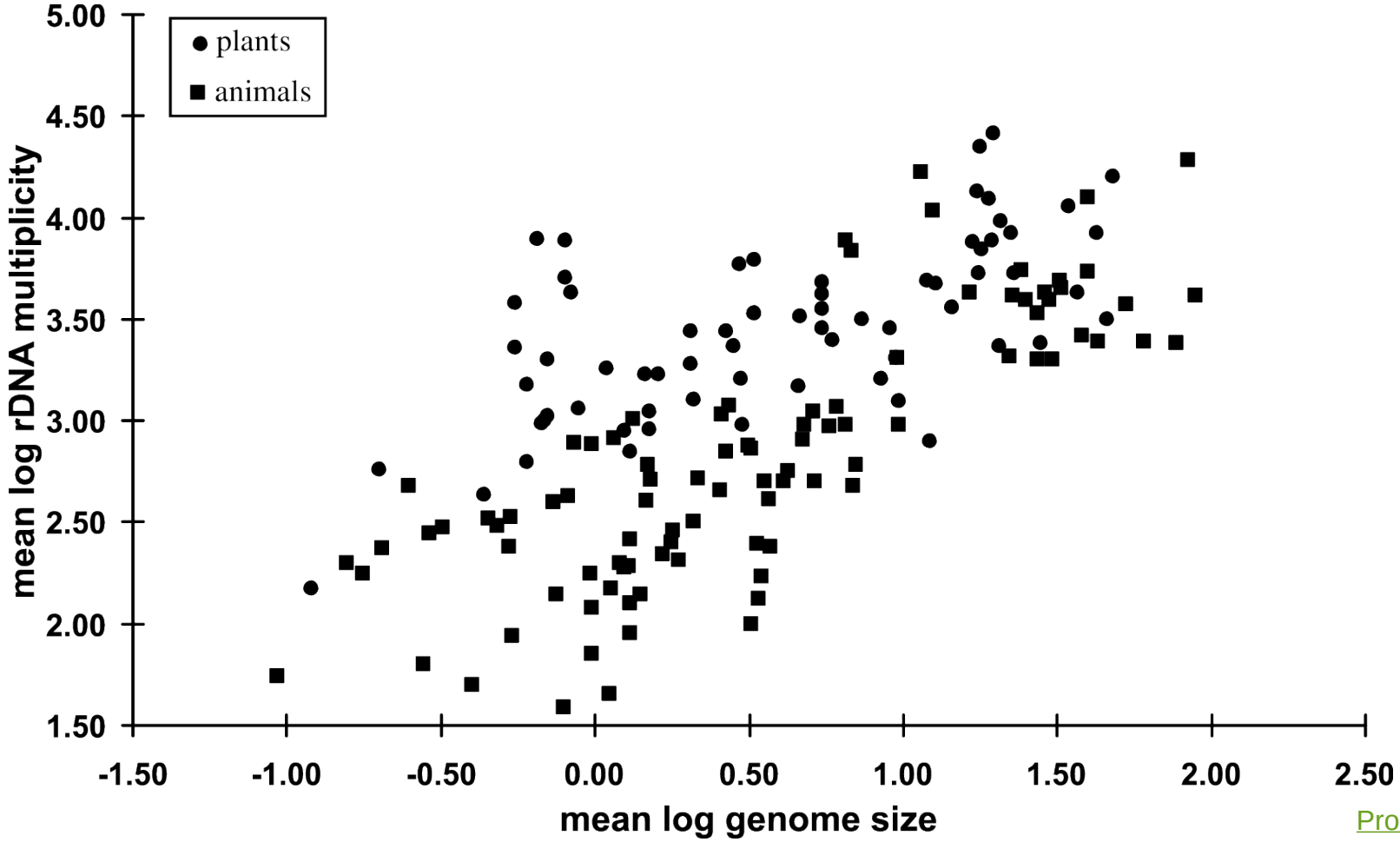




Lamb et al. 2019

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

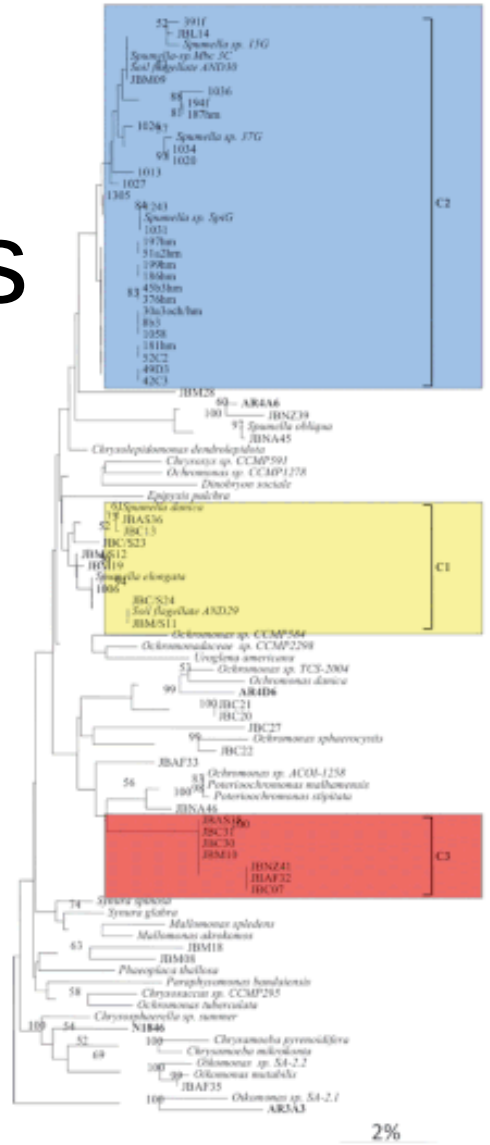
rRNA gene copies

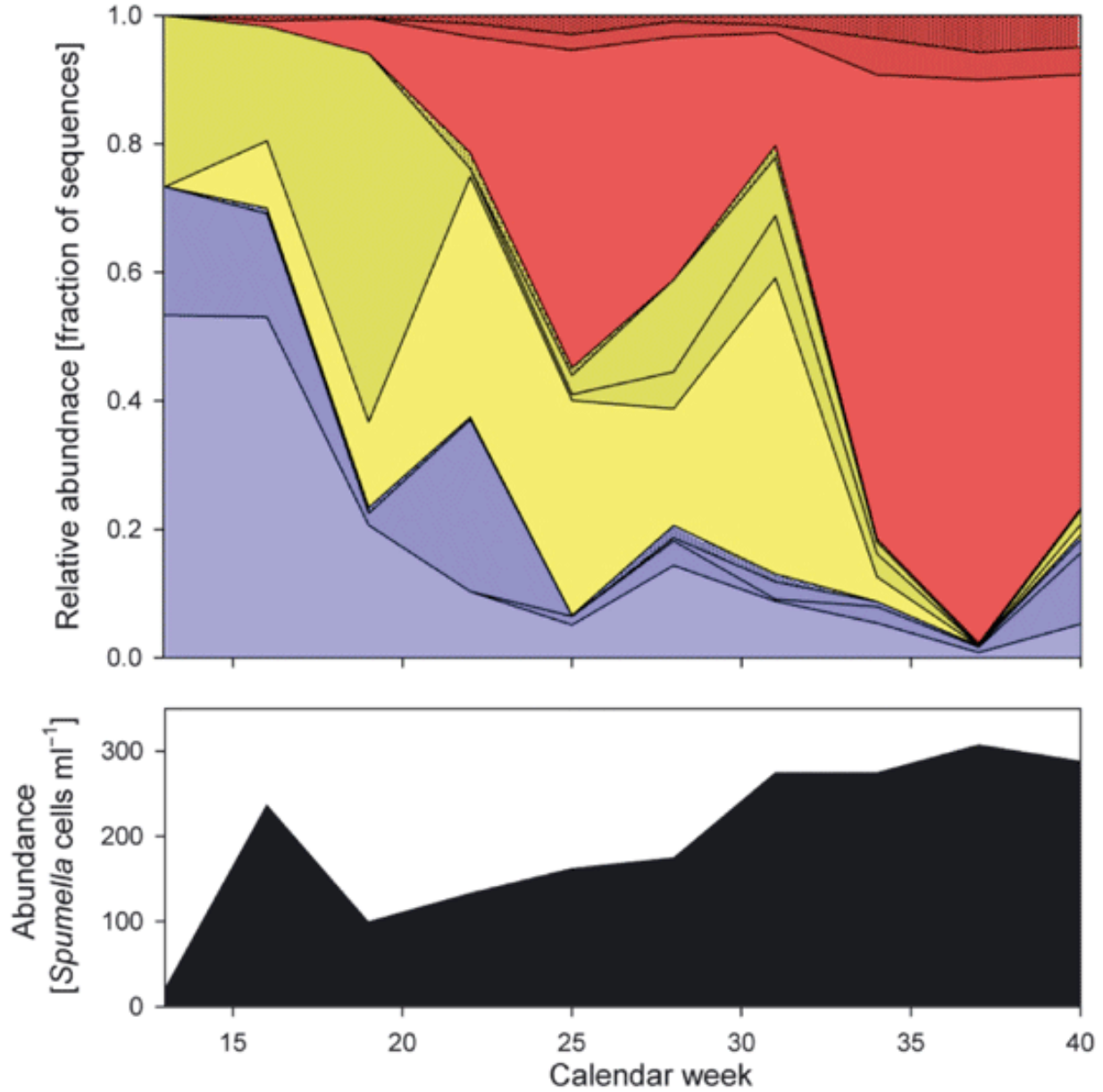


Prokopowich et al. 2003

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Seasonal fluctuations





Nolte et al. 2010

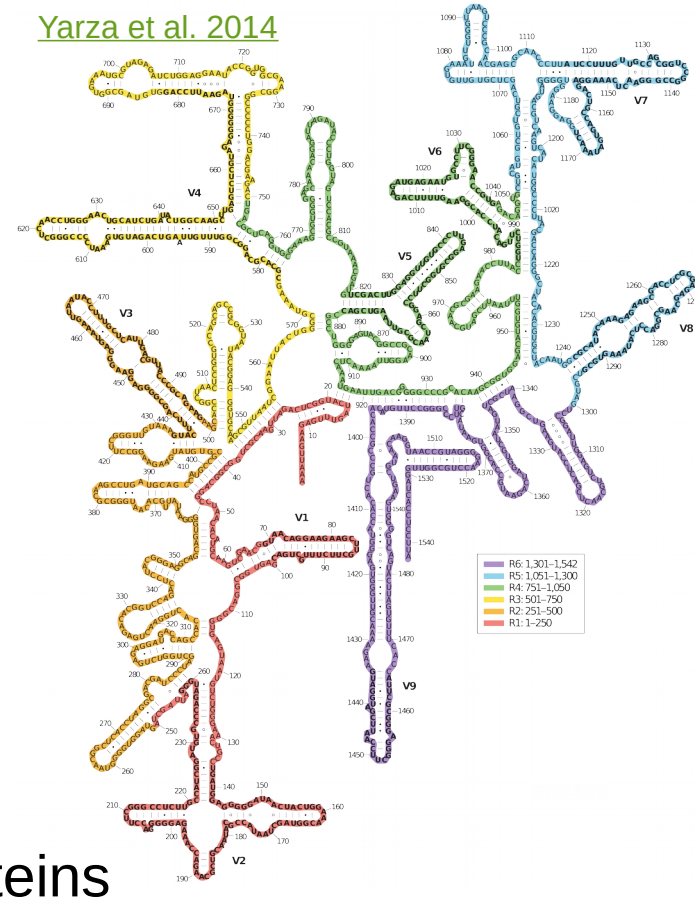
rRNA gene copies of ciliates

Estimated rDNA copy numbers in single ciliate cells and in other eukaryotic groups.

Organism	Accession No. (SSU)	Copies per cell	Standard deviation
Oligotrichs			
Favella sp.	JX178773	46,498	555
Pseudotontonia sp.	JX178769	172,889	9,832
Strombidinopsis sp.	JX178771	30,247	6,576
Strombidium sp.	JX178772	34,647	3,465
Tintinnopsis sp.	JX178770	126,372	1,368
Peritrichs			
Epistylis sp. iso.1	JX178765	64,865	15,089
Epistylis sp. iso.2	JX178766	88,161	20,699
Vorticella sp.1	JX178760	161,355	11,498
Vorticella sp.2	JX178761	315,786	7,100
Vorticella sp.3 iso.3	JX178762	99,376	5,482
Vorticella sp.3 iso.4	JX178763	61,226	2,417
Vorticella sp.5	JX178764	82,194	4,927
Zoothamnium sp.1	JX178767	40,675	4,145
Zoothamnium sp.2	JX178768	3,385	392
Oligohymenophorea			
Tetrahymena thermophila[#]	-	170-200	-
Tetrahymena thermophila	-	~9,000	-
Spirotrichea			
Oxytricha nova[*]	-	200,000	-
Stylonychia lemnae[*]	AF164124	400,000	-
Prostomatea			
Cryptocaryon irritans	AB608054	~3,000	-
Other groups			
Microalgae	-	1-12,000	-
Diatoms	-	61-36,896	-
Dinoflagellates	-	200-1,200	-
	-	1,057-12,812	-
Fungi	-	60-220	-
Animals	-	39-19,300	-
Plants	-	150-26,048	-

Gong et al. 2013

16S/18S rRNA gene

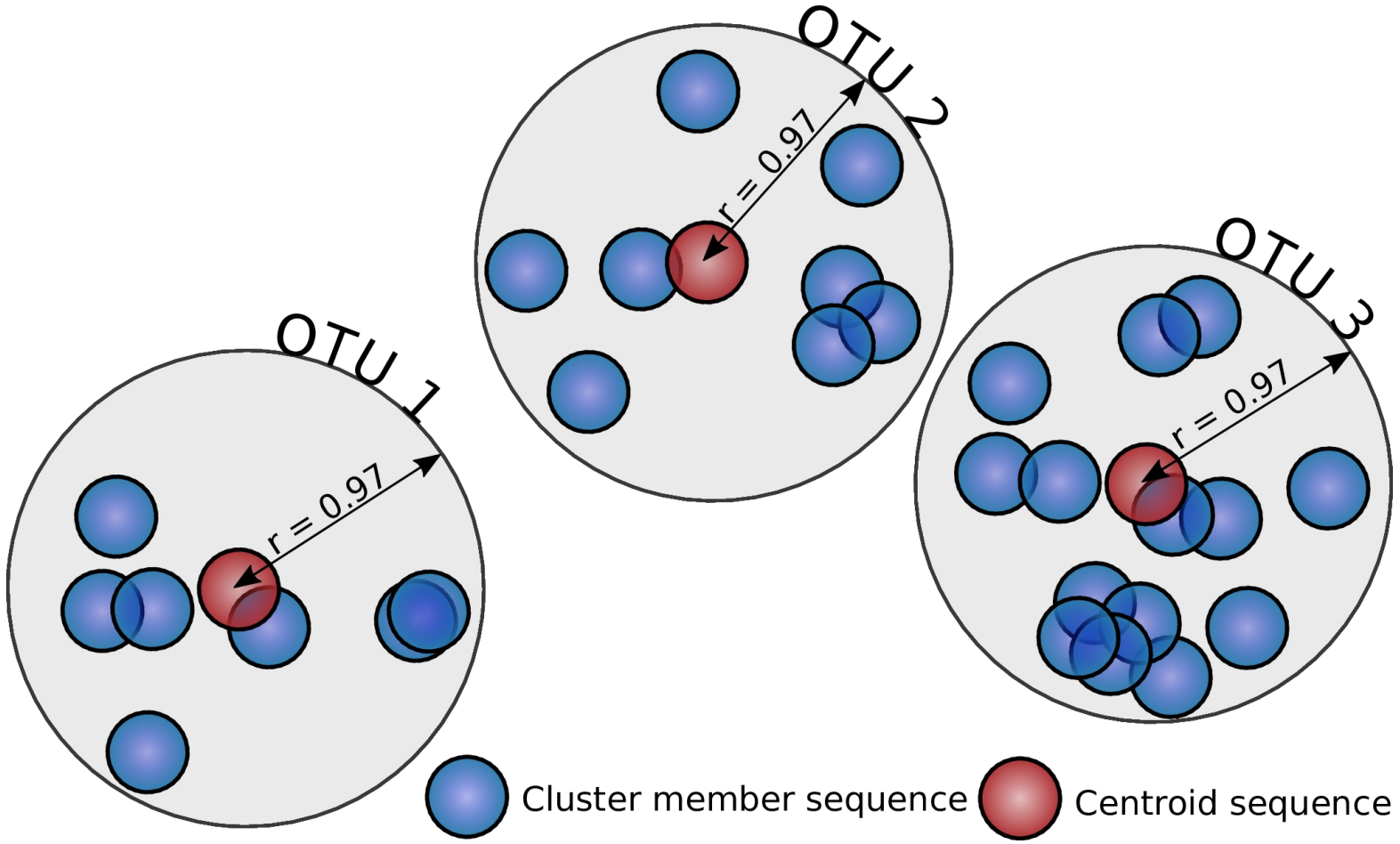


Yarza et al. 2014

- Standard marker for pro-/eukaryote diversity
  - Unlikely to be laterally transferred
  - Contains fast and slowly evolving portions
  - Universally conserved structure
  - Ancient and essential
  - Interacts with many other co-evolved cellular RNAs and proteins
  - Sequence is long enough to provide meaningful phylogenetic information
  - Sequence is small enough for a feasible analysis

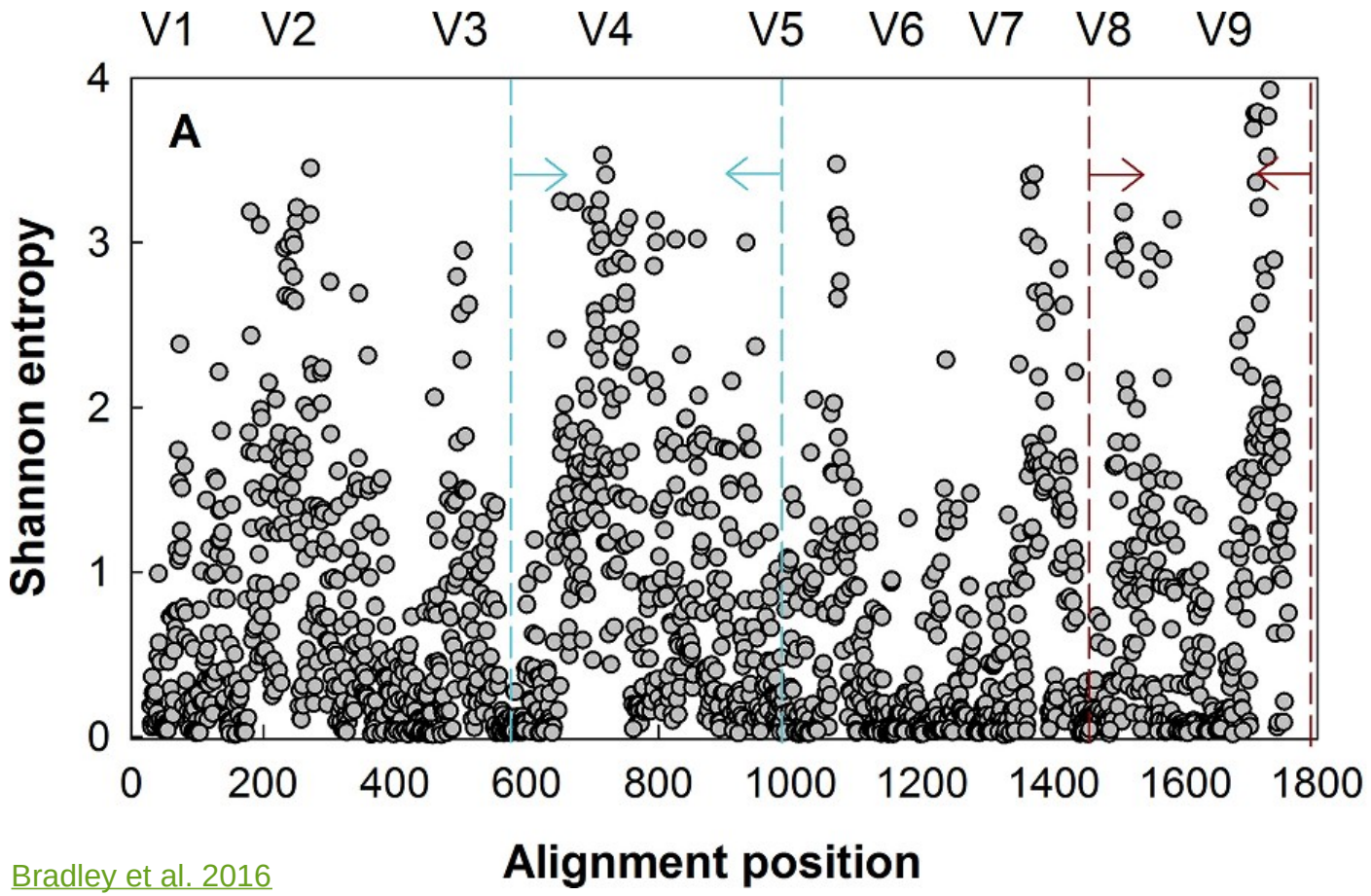
Fox 1977

Operational taxonomic unit (OTU)

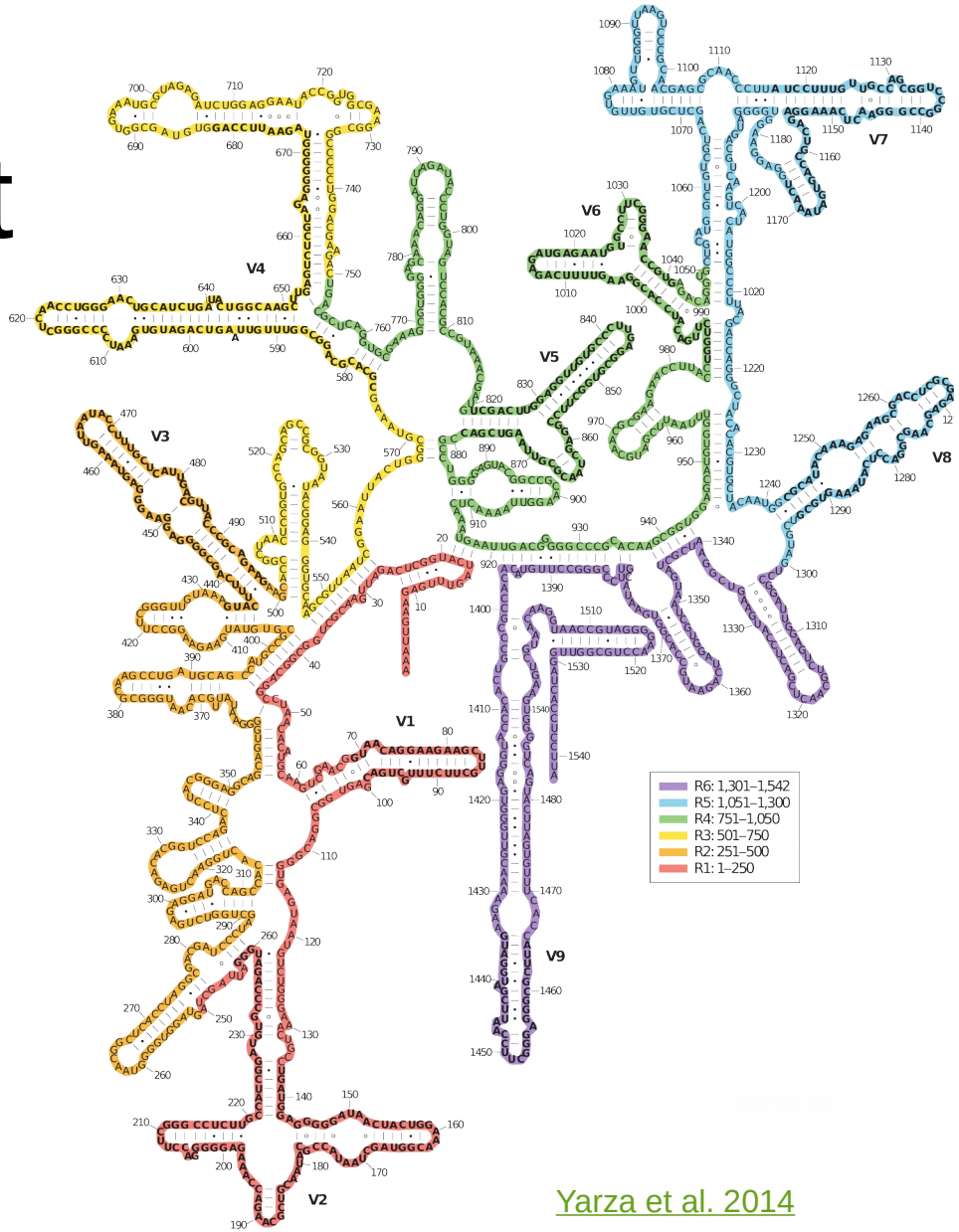


Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Operational taxonomic unit



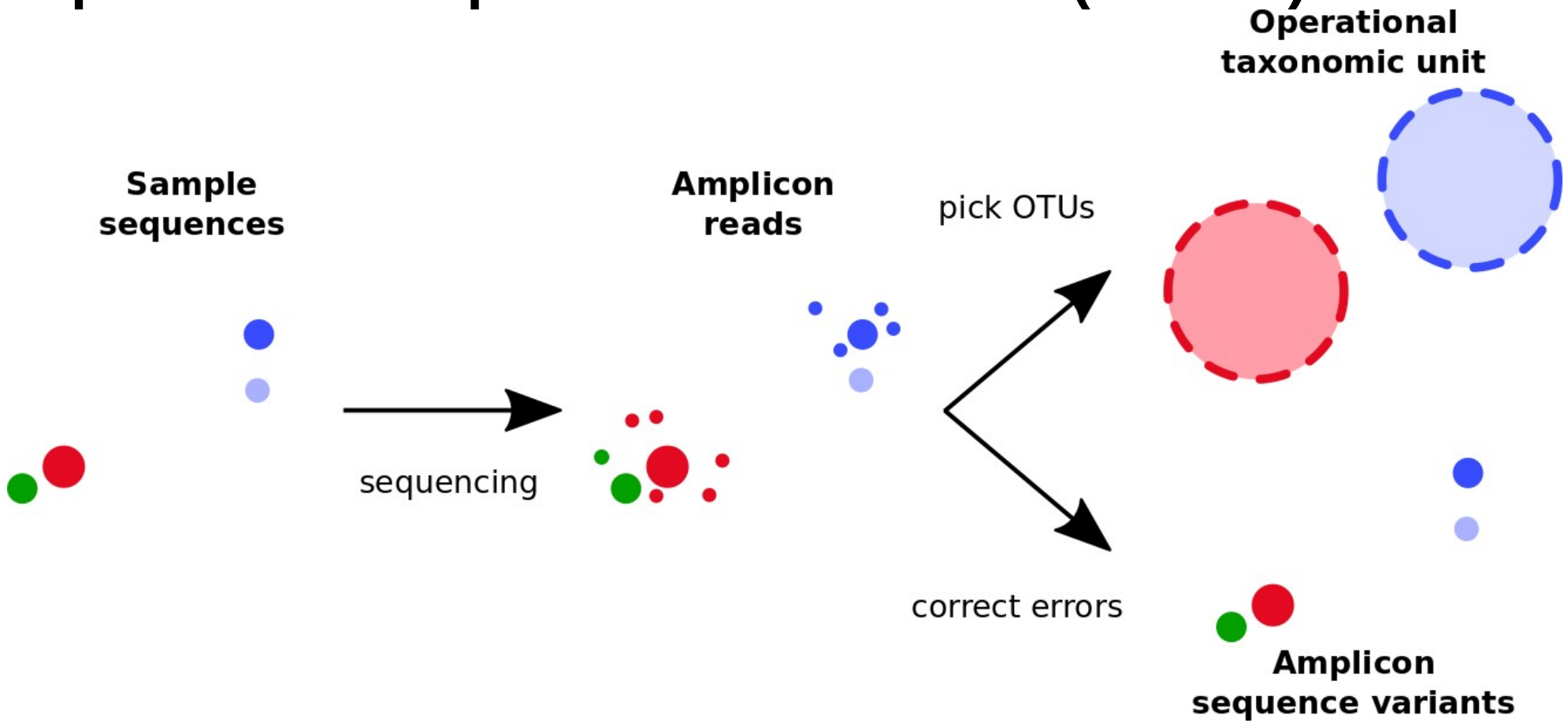
Bradley et al. 2016



Yarza et al. 2014

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Amplicon sequence variant (ASV)



After Wikimedia CC BY-SA 4.0 Dr. Benjamin John Callahan & Wikimedia CC BY-SA 4.0 Dr. Benjamin John Callahan

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

OTU vs. ASV: general issues

- The same species may contain differing marker sequences (16S, 18S)
- Different species may have very similar marker sequences (below distance cut-off) see Edgar 2018
- Amplicon sequences contain errors (error rate depends on technology)
- PCR of similar amplicons may lead to chimeras

OTU vs. ASV: Pros and Cons

Challenge	OTU	ASV
Species differing by a small number of nucleotides are discriminated		
Different marker sequences within a species are combined		
Not influenced by sequencing errors		
OTUs / ASVs can be compared across studies		
Low abundance species are retained		

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

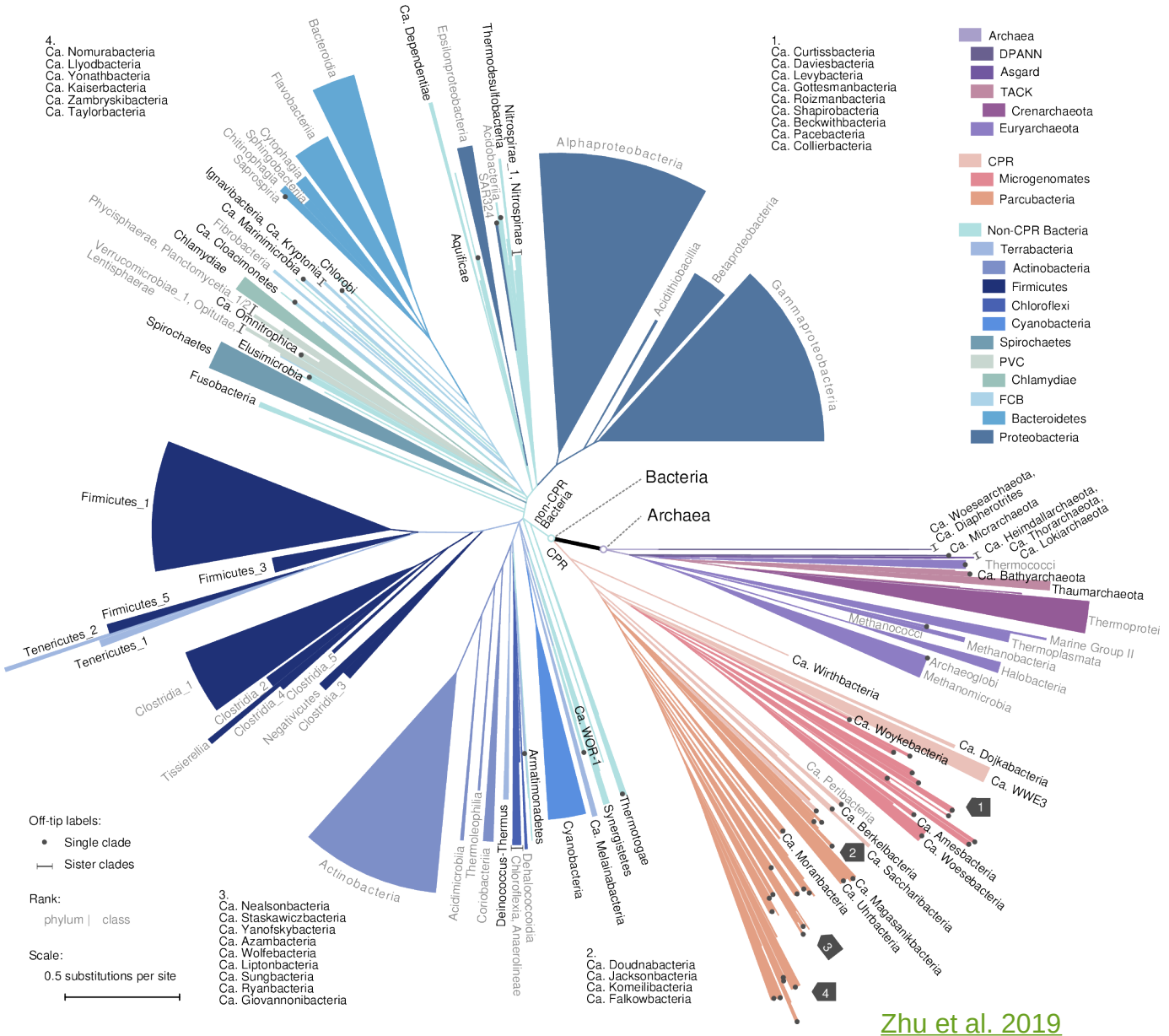
Species concepts

> [...] I was much struck how entirely vague and arbitrary is the distinction between species and varieties
>
> — Charles Darwin, *On the Origin of Species*[4]

Wikipedia [23.8.2020]

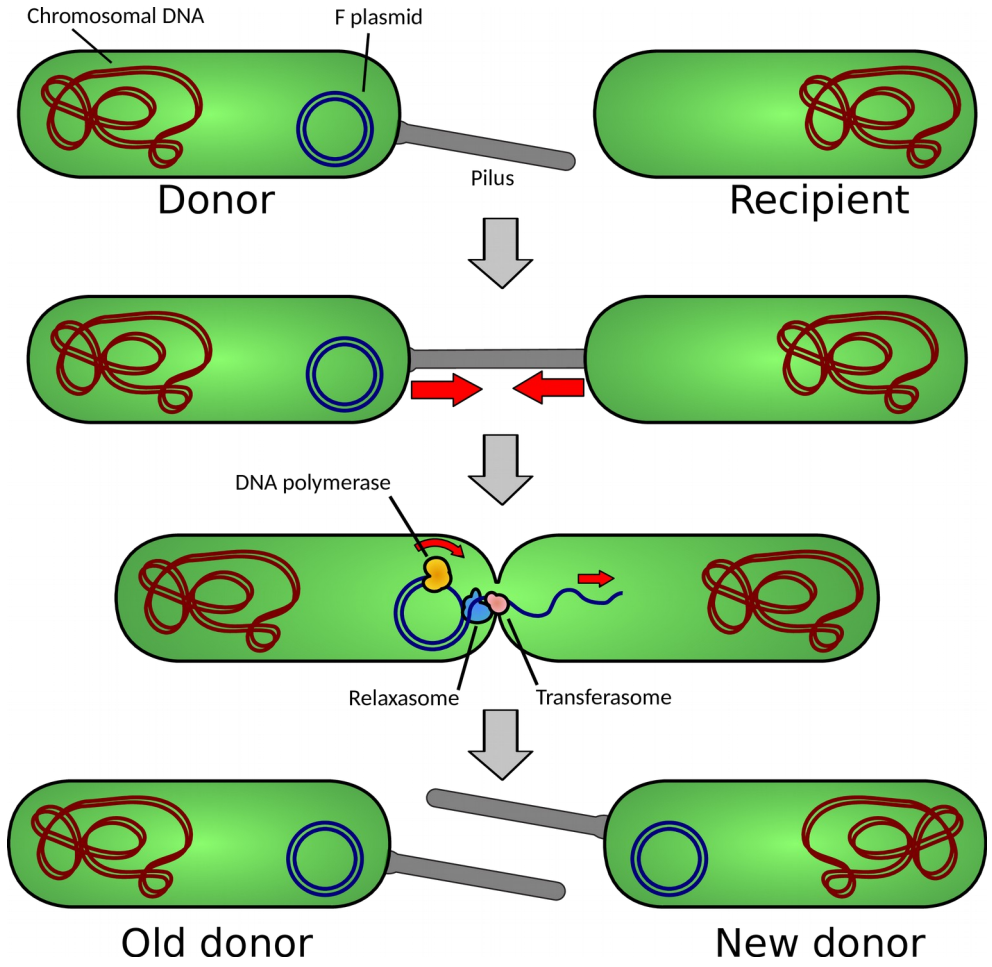
Prokaryotes

- Up to 10^{12} bacteria species (Locey et al. 2016)



Zhu et al. 2019

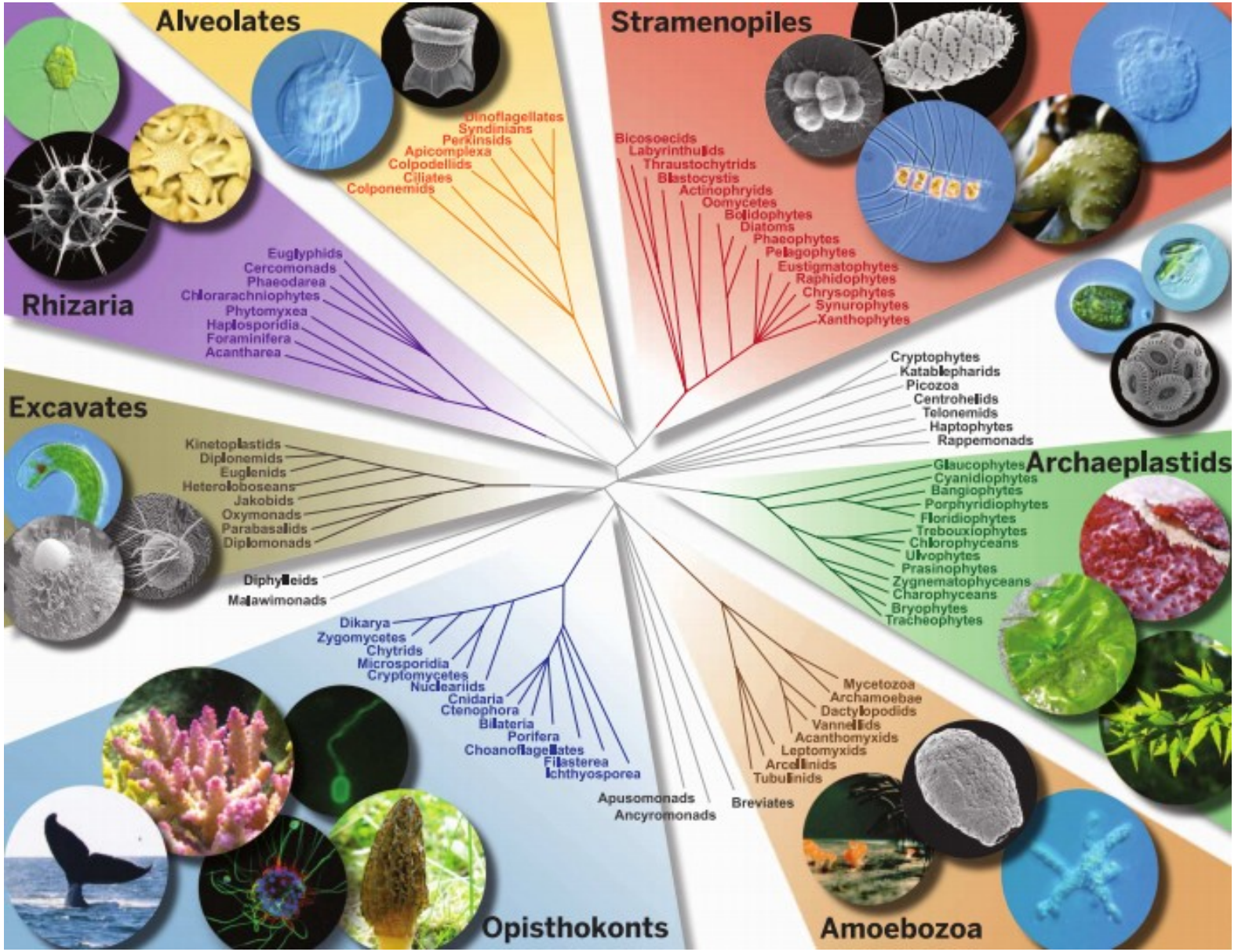
Bacterial conjugation



Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Eukaryotes

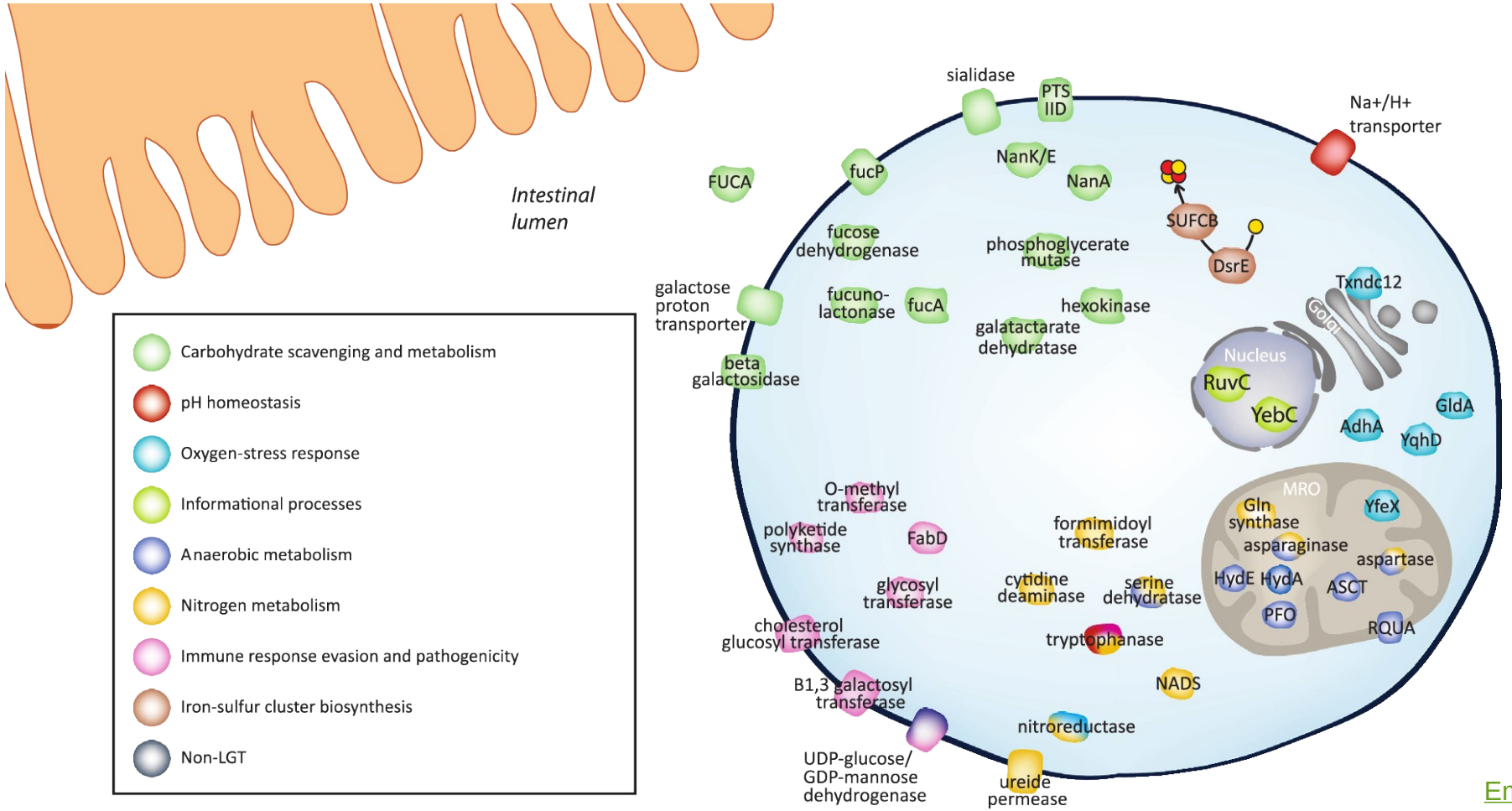
- > 10^6 protist species
  (Pawlowski et al. 2012)
- All trophic levels



Worden et al. 2015

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

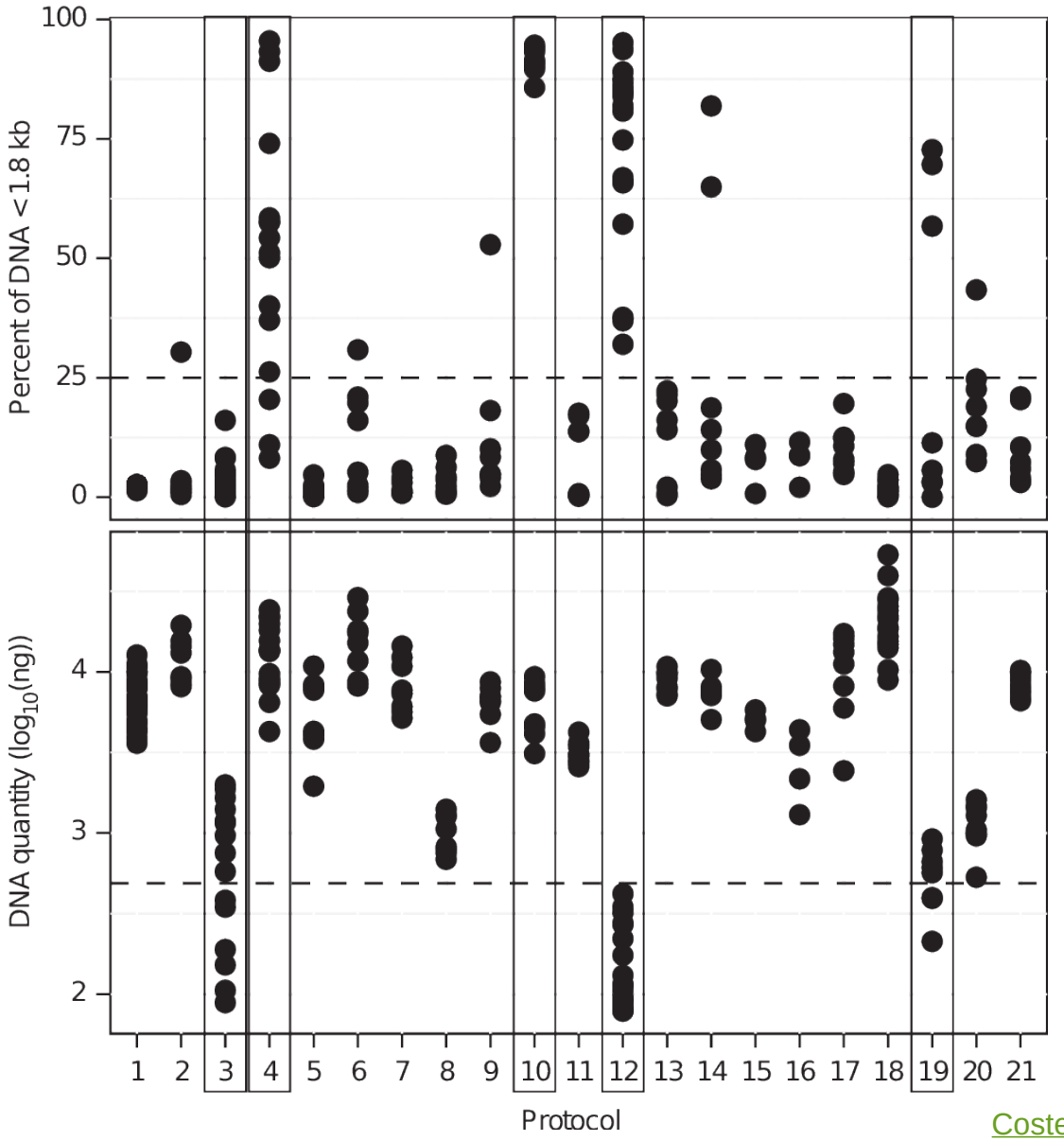
Horizontal gene transfer in *Blastocystis*



Eme et al. 2017

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

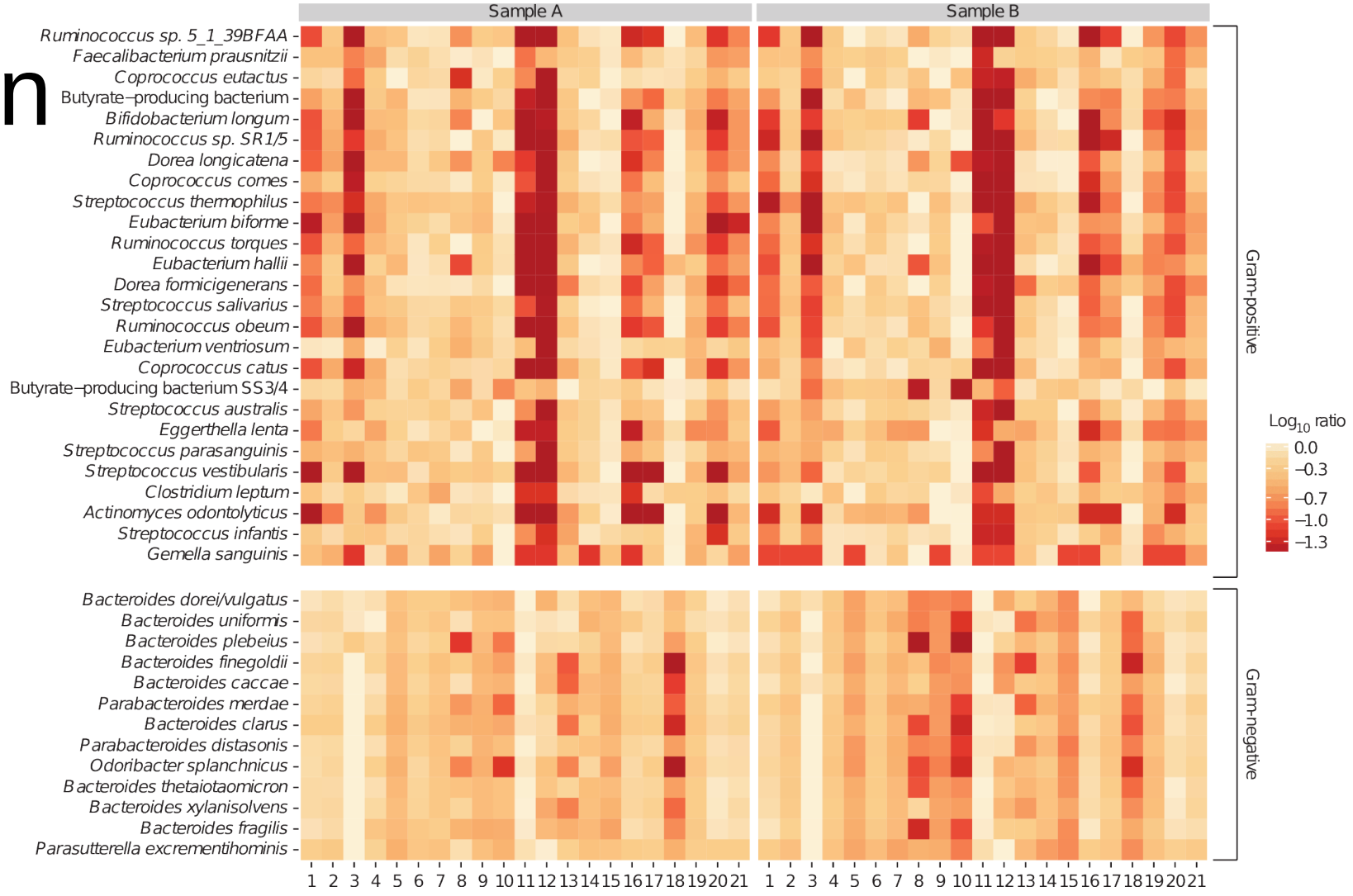
DNA extraction



Costea et al. 2017

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

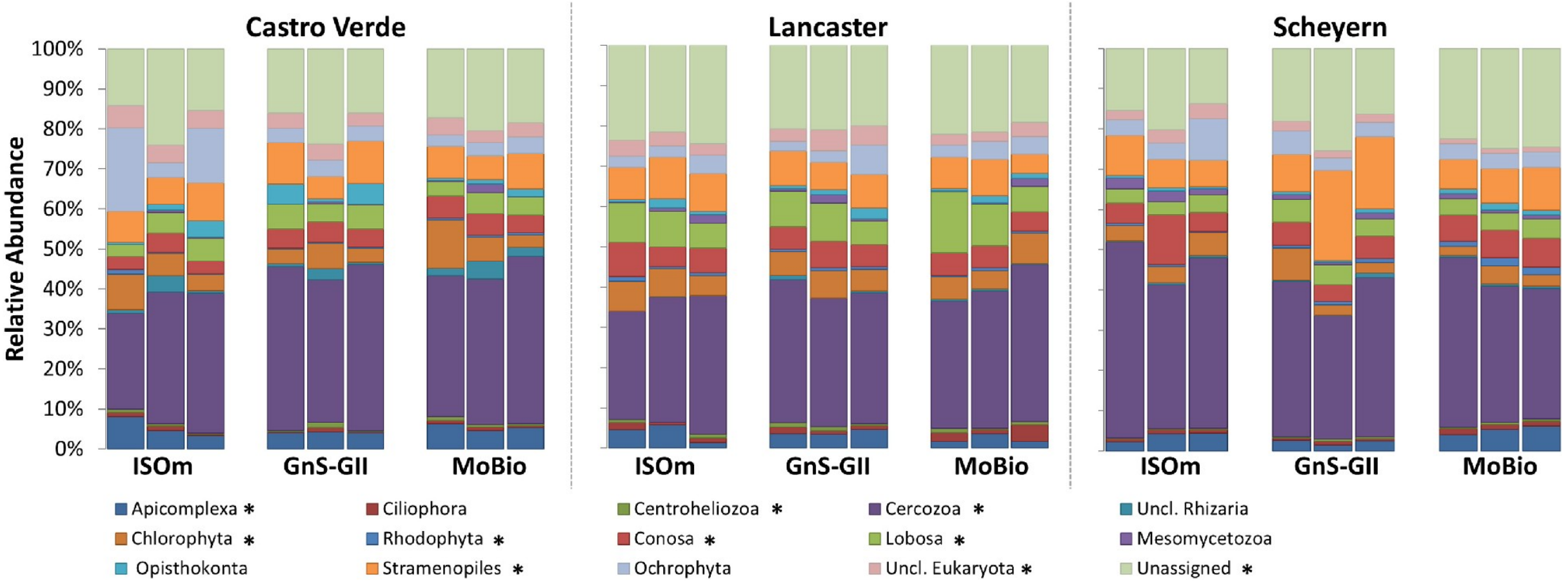
DNA extraction



Costea et al. 2017

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

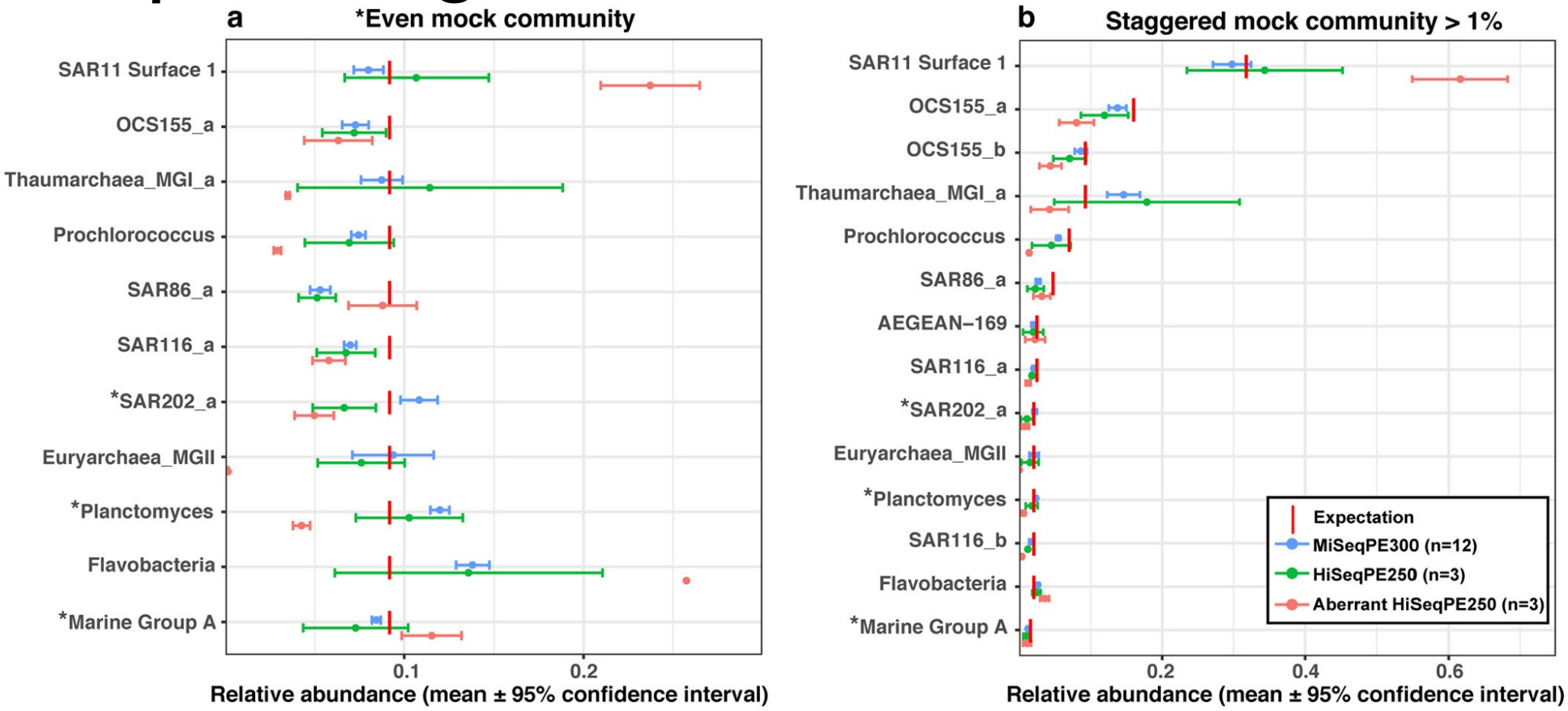
DNA extraction



Santos et al. 2017

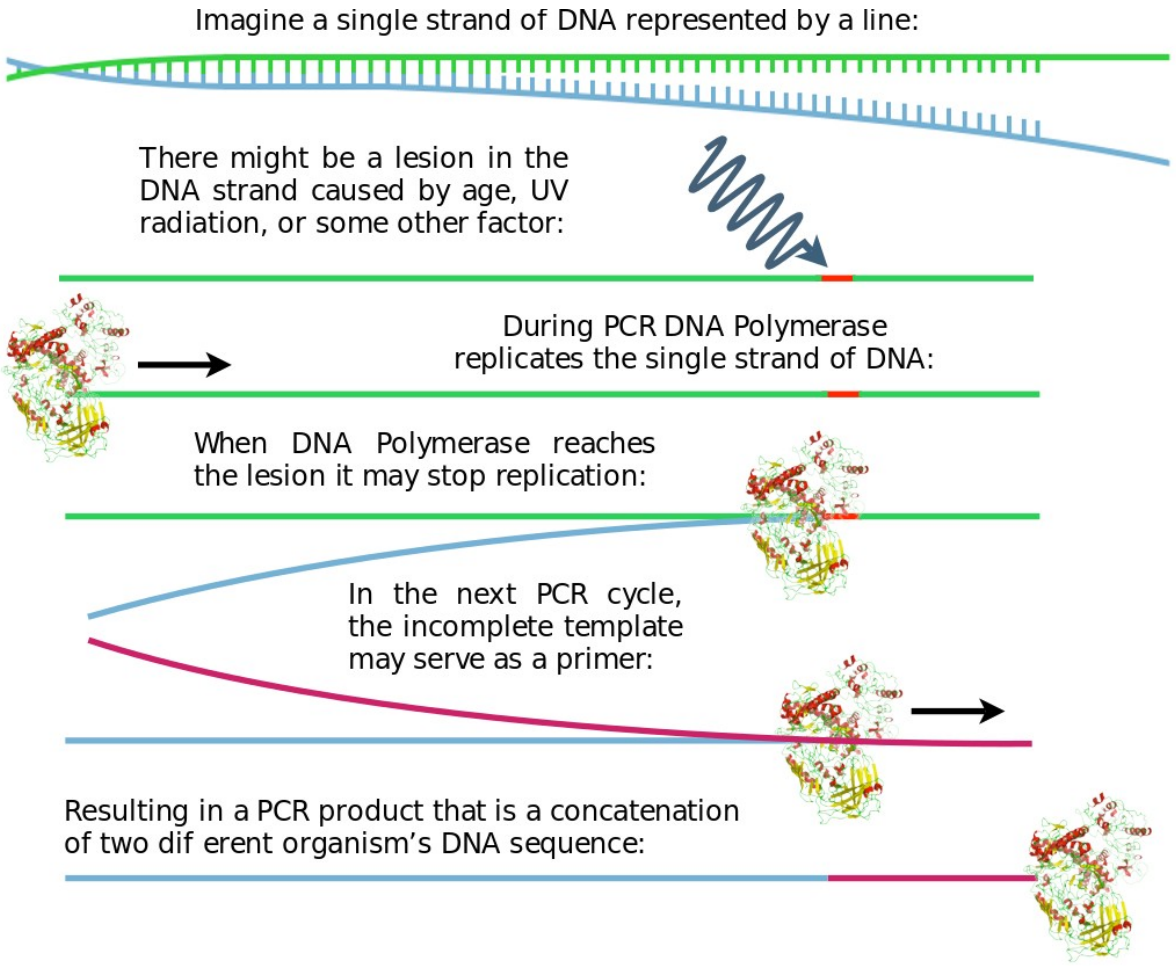
Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

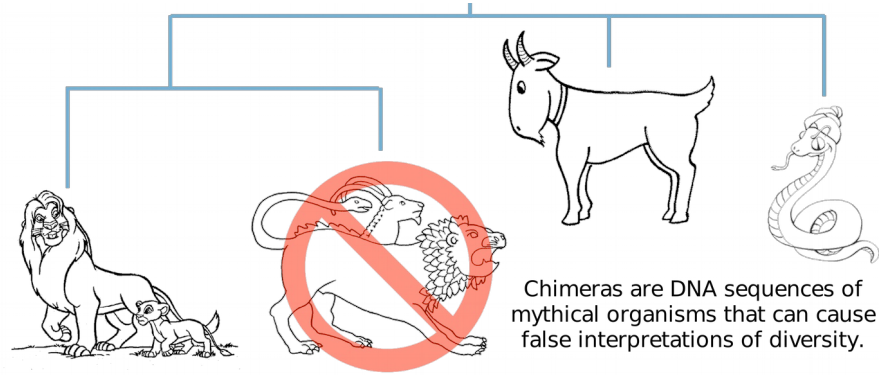
Sequencing error



Mukherjee et al. 2015

Chimeras

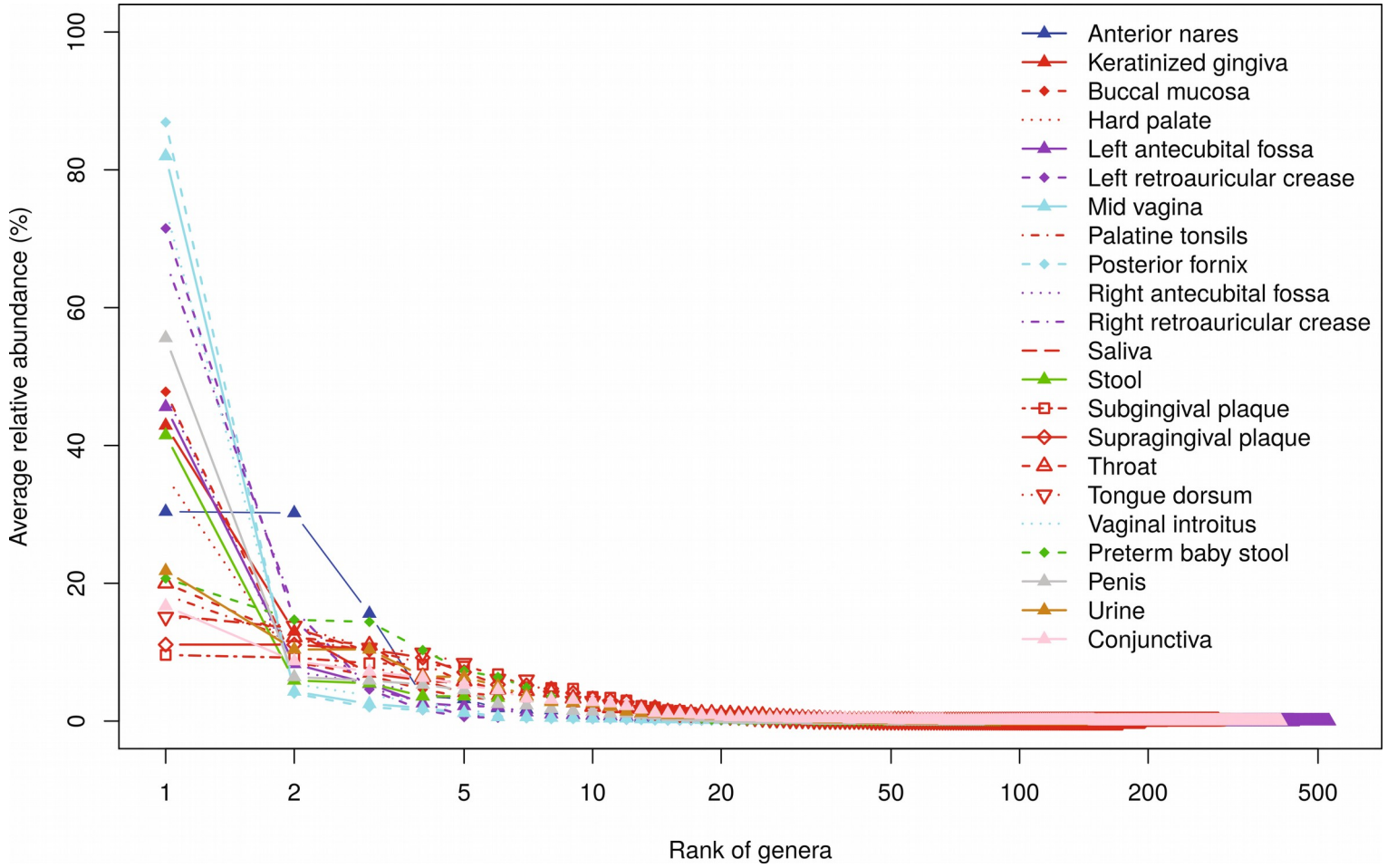




Wright et al.

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Spurious or rare taxa?

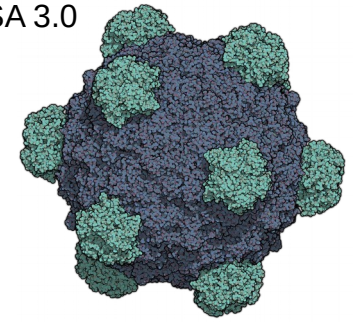


Zhou et al. 2013

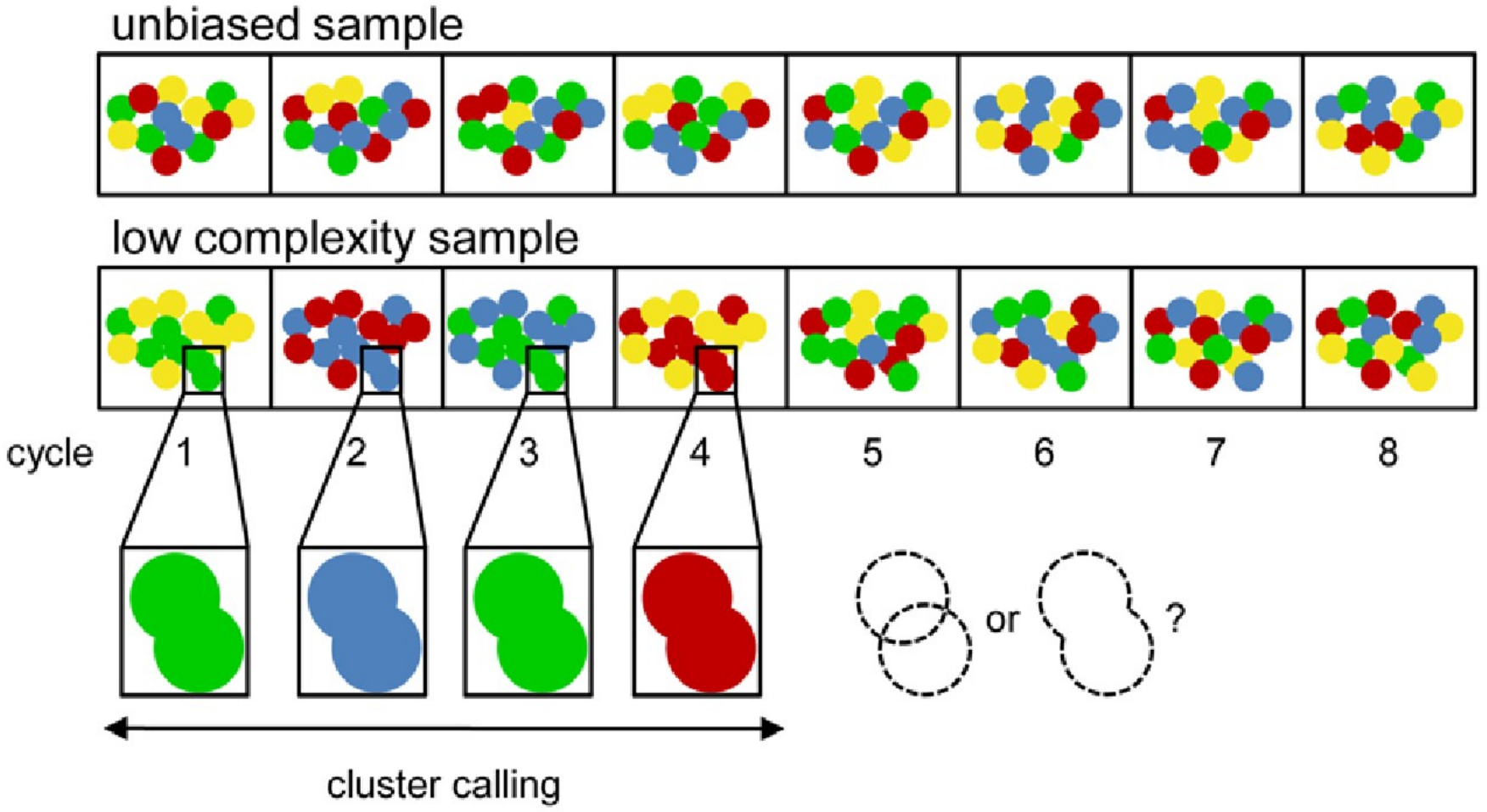
Spurious taxa

- Reason: PCR errors
- Arise especially from high abundant taxa
- Solution: abundance filtering

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

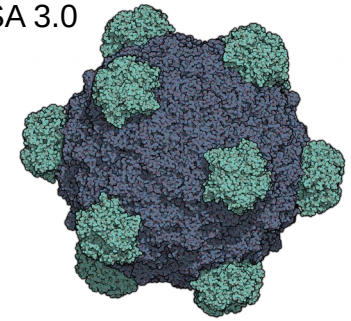


PhiX spike-in

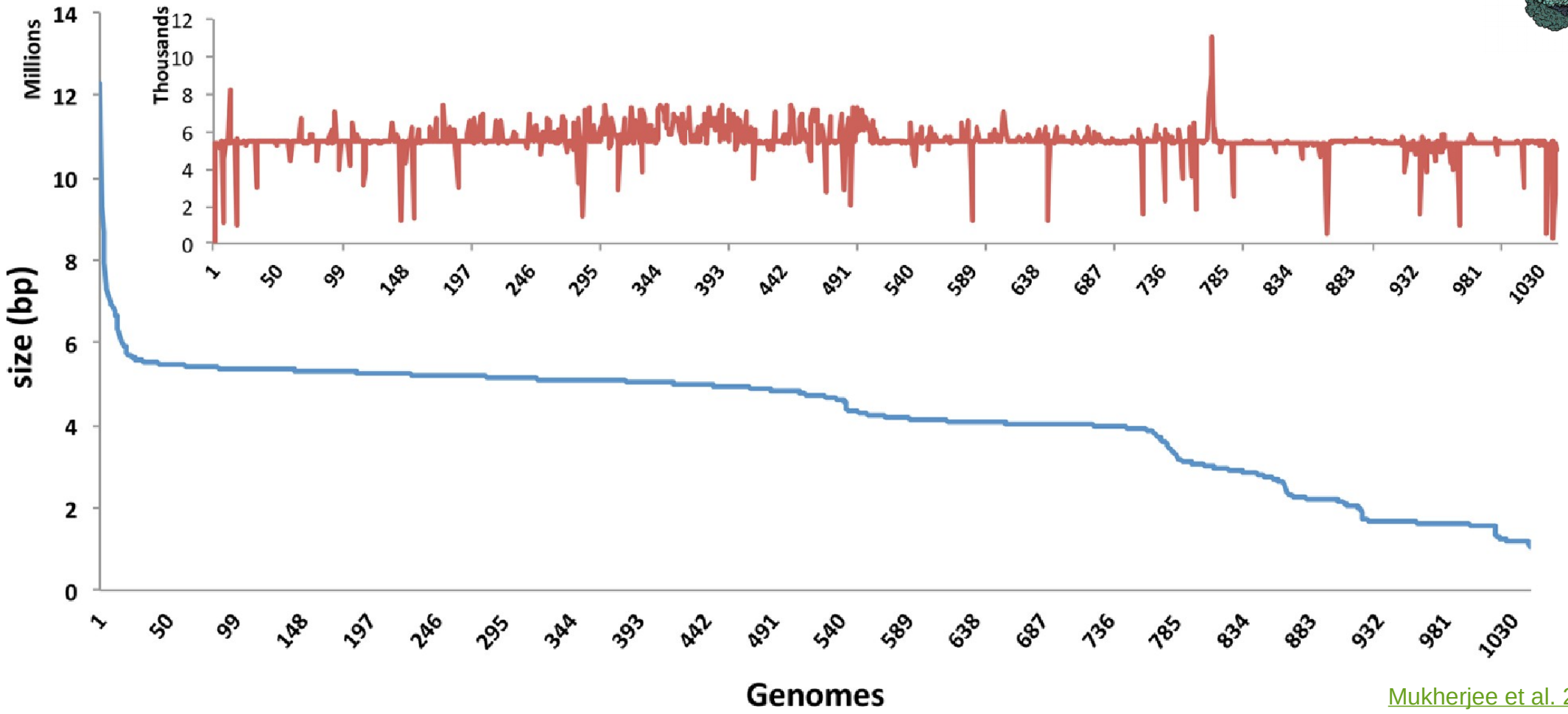


Krueger et al. 2011

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil



PhiX contamination



Mukherjee et al. 2015

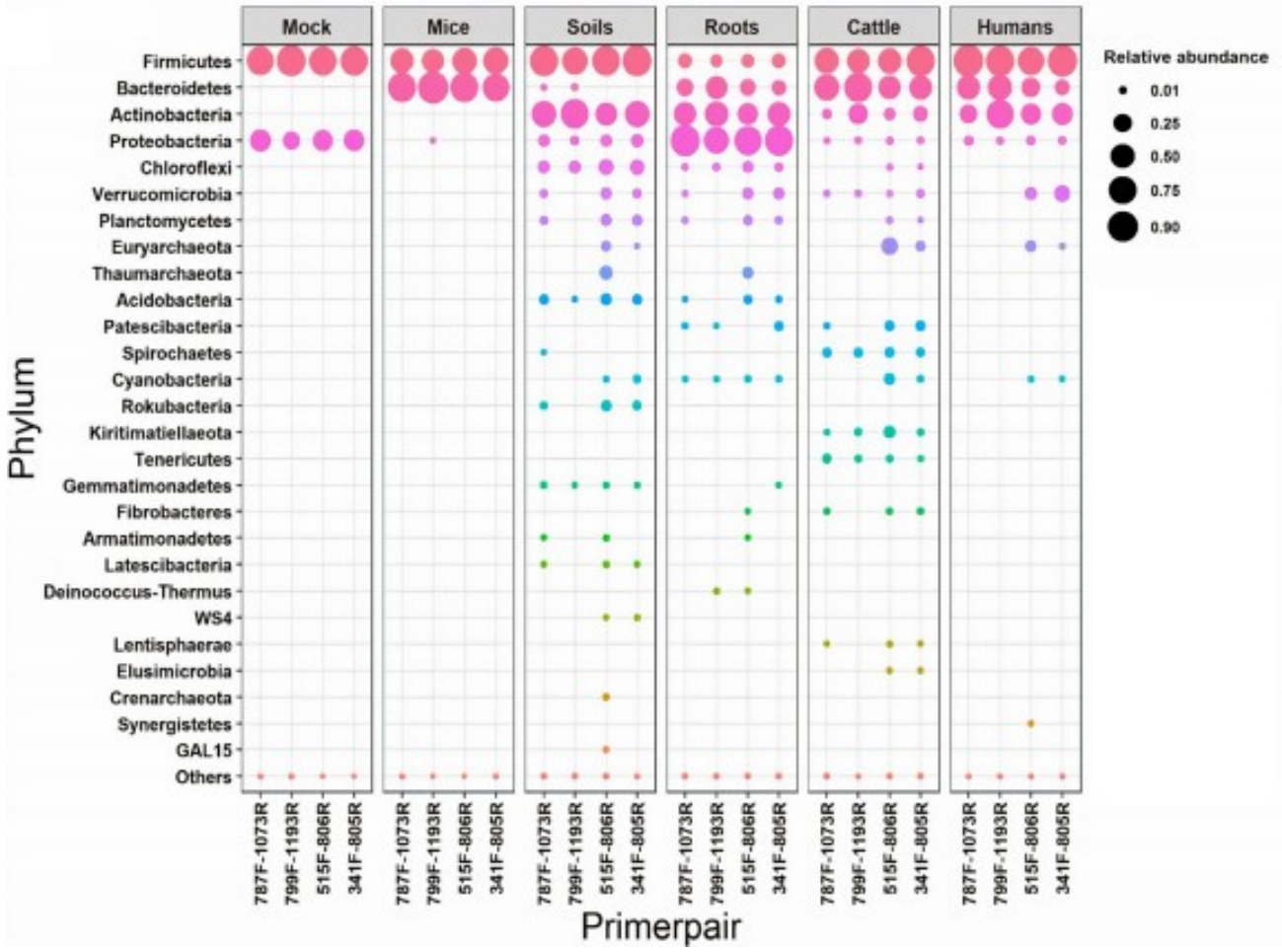
Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Primer COI

Hajibabaei et al. 2019

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

Primer prokaryotes



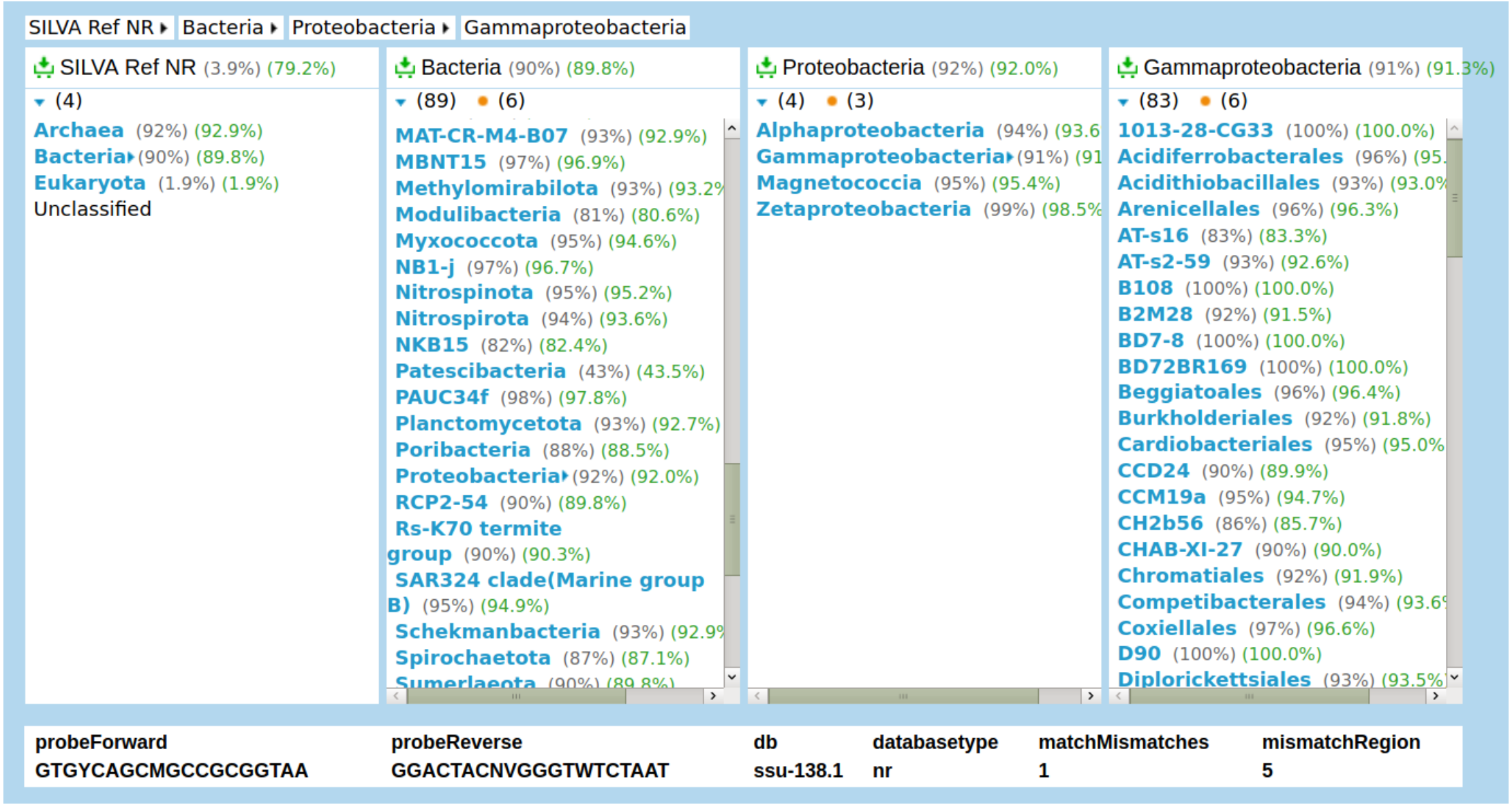
Wasimuddin et al. 2019

Gerhard Thallinger, PhD, Graz University of Technology
Rachel Korn, PhD, Université de Fribourg
Magdalena Steiner, PhD, Agroscope Wädenswil

SILVA TestPrime

- In-silico PCR on SILVA db

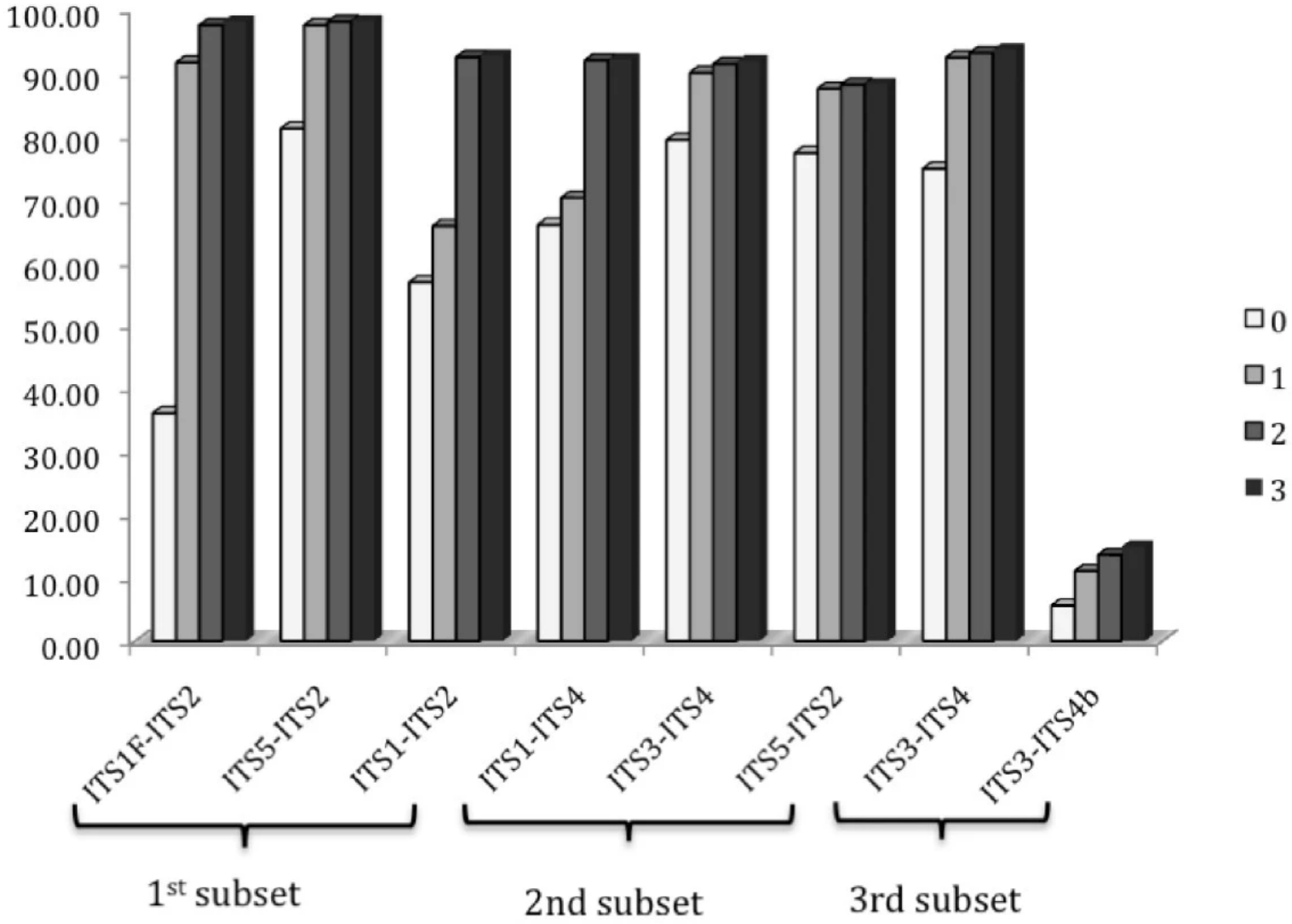
SILVA test prime



probeForward	probeReverse	db	databasetype	matchMismatches	mismatchRegion
GTGYCAGCMGCCGCGGTAA	GGACTACNVGGGTWTCTAAT	ssu-138.1	nr	1	5

ecoPCR

- In-silico PCR on custom db



EcoPCR on GitLab

Bellemain et al. 2010; Ficetola et al. 2010

Take home

- Remain consistent within a project (sample treatment, DNA extraction, sequencing platforms and even runs, primers & downstream analysis)
- Replicate samples (and think of sample size too)
- Include negative (reagent) and positive (mock) controls
- If possible, pilot study
- If stuck, contact the developers (but search the web before!)