Bioinformatics analysis using metagenome

Today's class

- Metagenome and bioinformatics
- Bioinformatics analysis using metagenome
  - DNA extraction
  - Library preparation
  - High-throughput sequencing
  - Data pretreatment
  - Bioinformatics analysis

Metagenomics / Bioinformatics

- Metagenomics*: study of genetic materials recovered directly from environmental or clinical samples by sequencing
  - For prokaryotes, commonly performed by analyzing 16S rRNA**

** "meta-" means more comprehensive*

*** Strictly speaking, DNA that encodes 16S rRNA*

- Bioinformatics: study of developing methods and tools for understanding biological data

- Why metagenomics?
  - Microbial community evolves in response to its environment
  - Different microbial communities may function differently
  - To understand microbial community we need information on its members
  - Many of the microorganisms in environmental samples are viable but not culturable

ACS Sustainable Chemistry & Engineering

pubs.acs.org/journal/ascecg Research Article

Long-Term Stability of High-*n*-Caproate Specificity-Ensuring Anaerobic Membrane Bioreactors: Controlling Microbial Competitions through Feeding Strategies

Byung-Chul Kim, Changyu Moon, Yongju Choi, and Kyoungphile Nam*

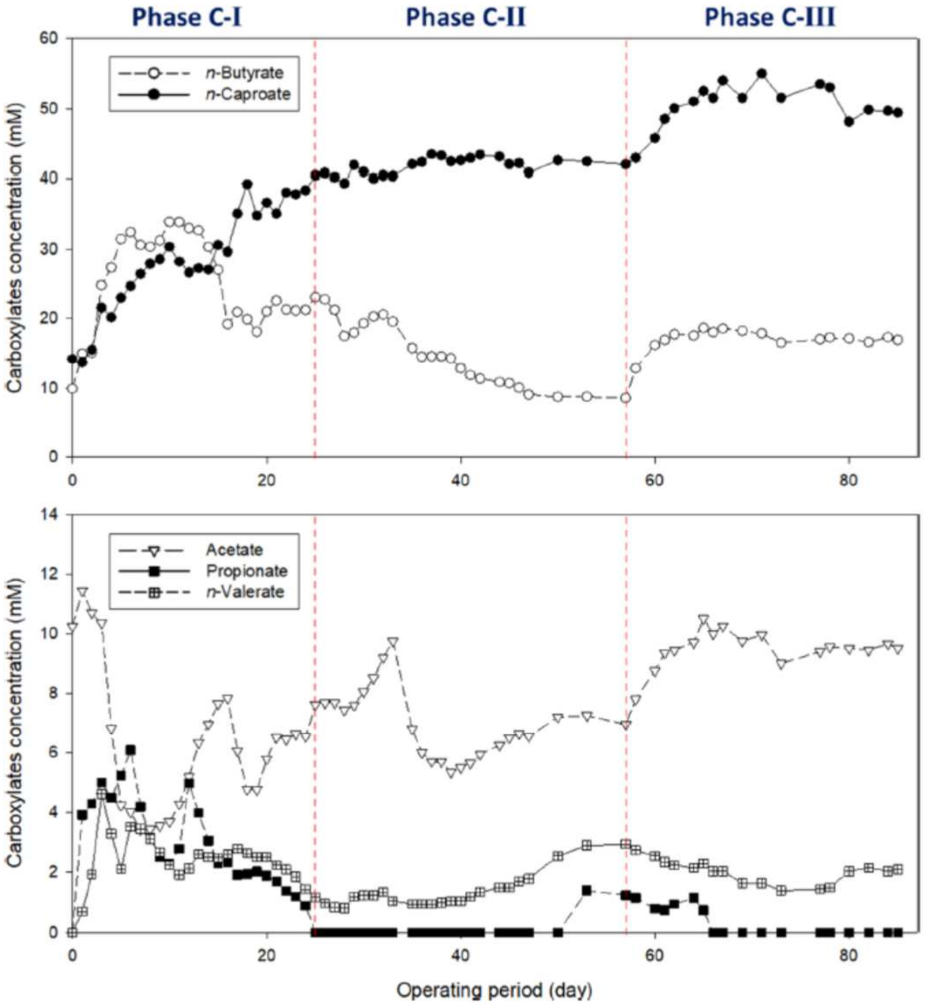


Figure 4. Concentration (mM) of carboxylates in the effluent during the C-AnMBR operating period. The operating period was divided into phase C-I (initial operation period, day 1–24), phase C-II (high *n*-caproate specificity period, day 25–56), and phase C-III (increased electron acceptor concentration period, day 57–85).

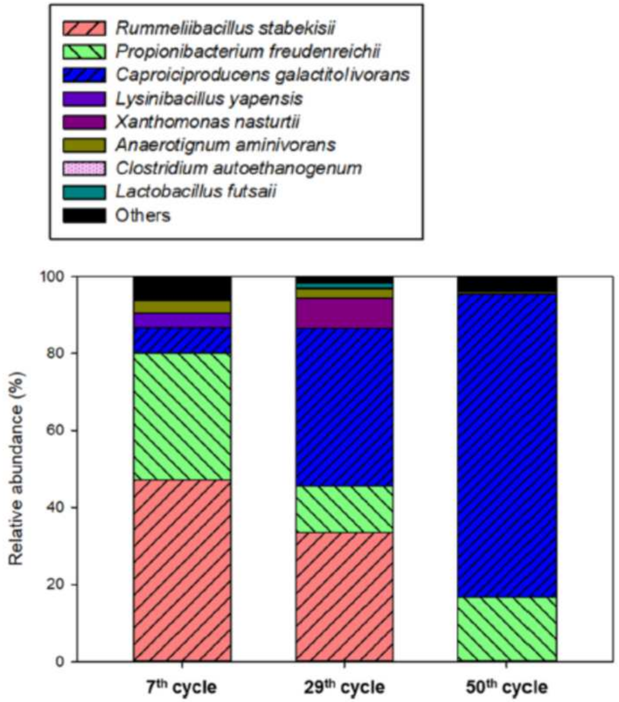
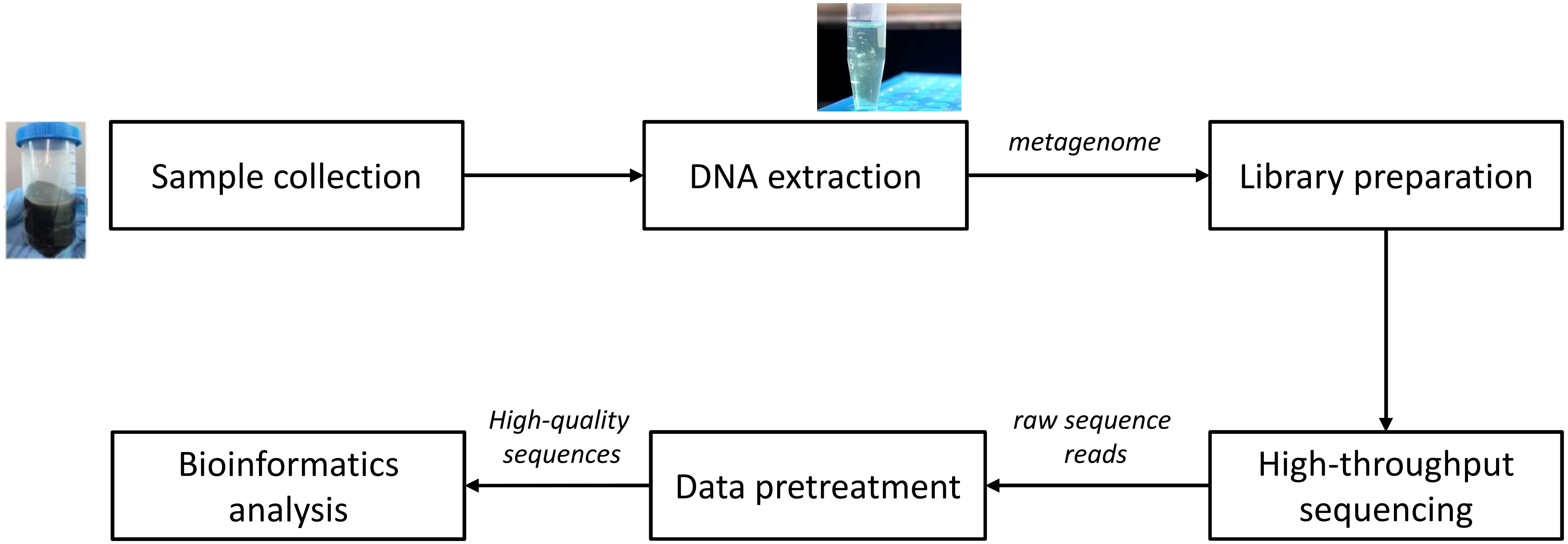


Figure 2. Genus-level relative abundance of samples collected at the 7th, 29th, and 50th operating cycles during the Sc-AnMBR operating period.

Bioinformatics analysis using metagenome: General workflow



Sample collection

DNA extraction

metagenome

Library preparation

High-throughput sequencing

raw sequence reads

Data pretreatment

High-quality sequences

Bioinformatics analysis

- Microbial composition
- Diversity/similarity
- Interaction networks
- Metabolic networks and pathways
- Functional gene profile

- Removing adapters and linkers
- Excluding chimeras and replication
- Demultiplexing of barcoded samples

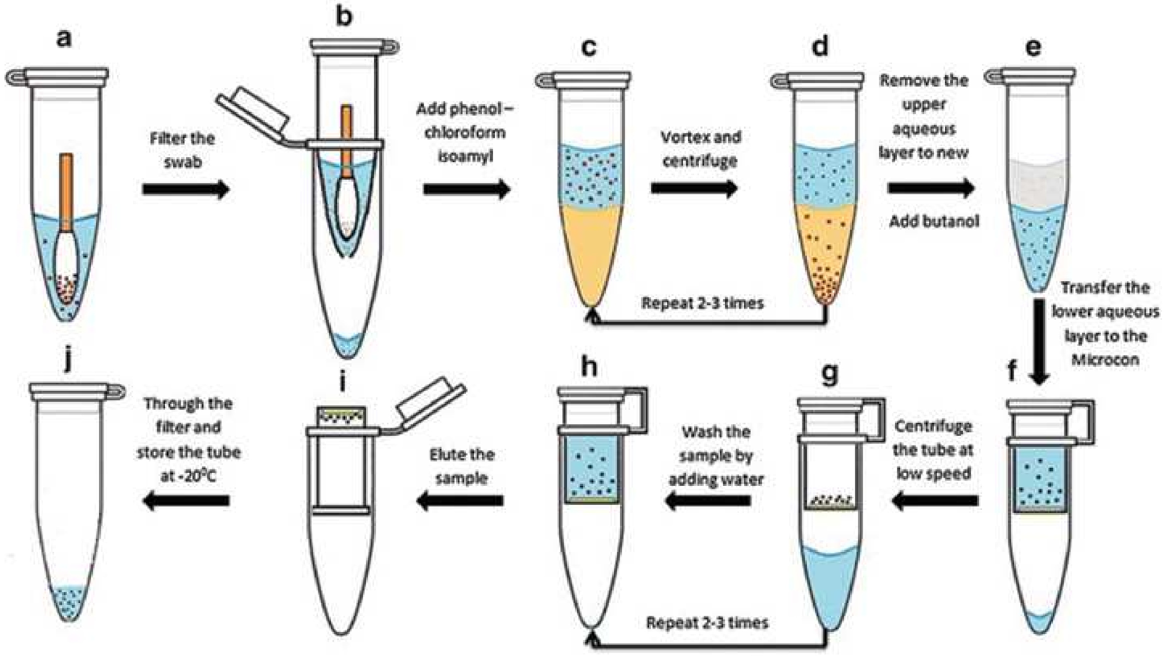
- 454 sequencing (pyrosequencing)
- Illumina sequencing
- Single molecule, real-time (SMRT) sequencing
- Nanopore sequencing
- …

DNA extraction: general procedure

1) Cell lysis (break open the cell)
2) Separate the DNA from other cell components
3) Isolate the DNA

Procedure example

- May be different for different types of samples/extraction methods


a
Filter the swab
b
Add phenol–chloroform isoamyl
c
Vortex and centrifuge
d
Repeat 2-3 times
Remove the upper aqueous layer to new
Add butanol
e
Transfer the lower aqueous layer to the Microcon
f
Centrifuge the tube at low speed
g
Wash the sample by adding water
Repeat 2-3 times
h
Elute the sample
i
Through the filter and store the tube at -20°C
j

Library preparation

- Metagenomic library: fragments of genetic material extracted from environmental or clinical samples and cloned* into specific vectors**

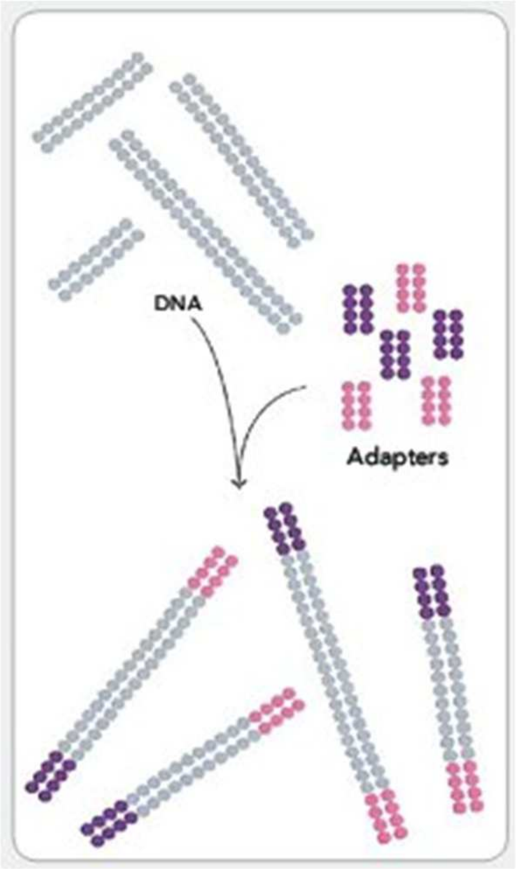
** copied*

*** a DNA molecule used to carry a particular DNA segment*

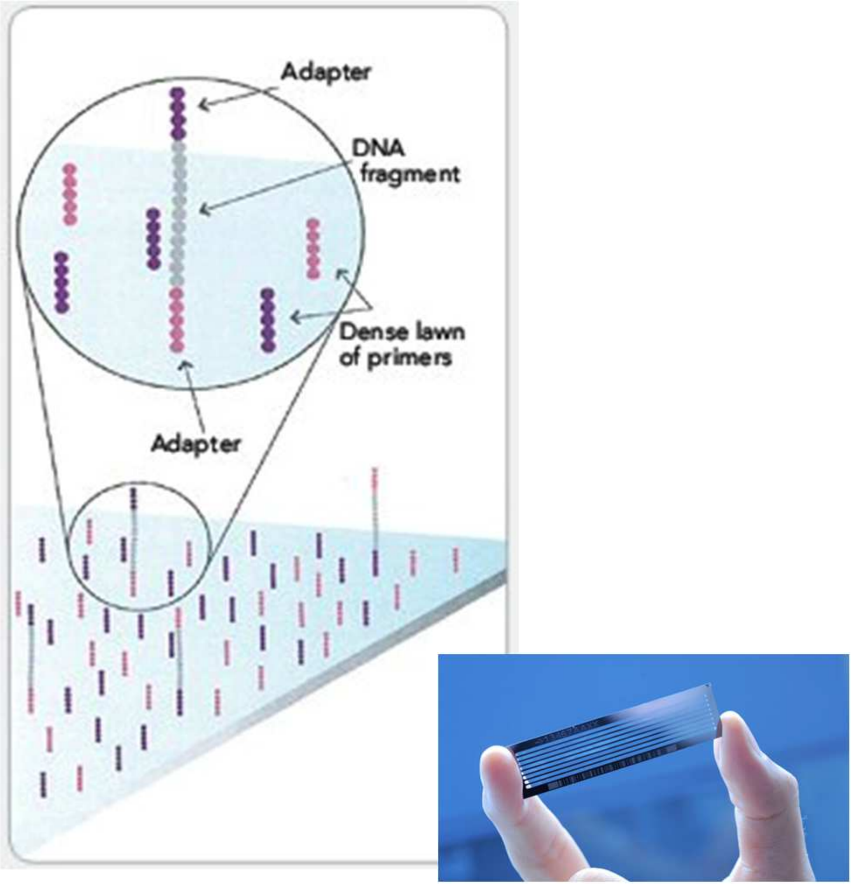
- Needed for high-throughput sequencing

- Specific library prep methods needed for each sequencing technique

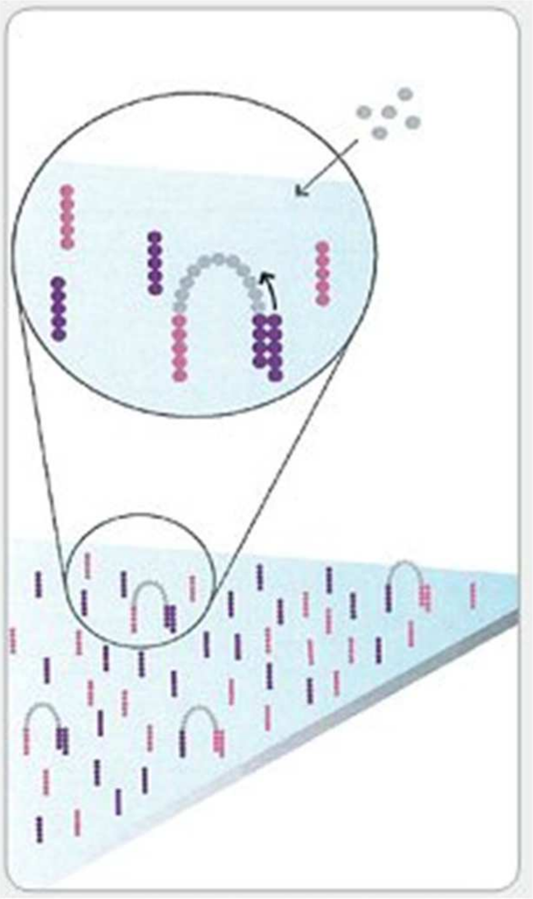
Library prep example – Illumina seq.



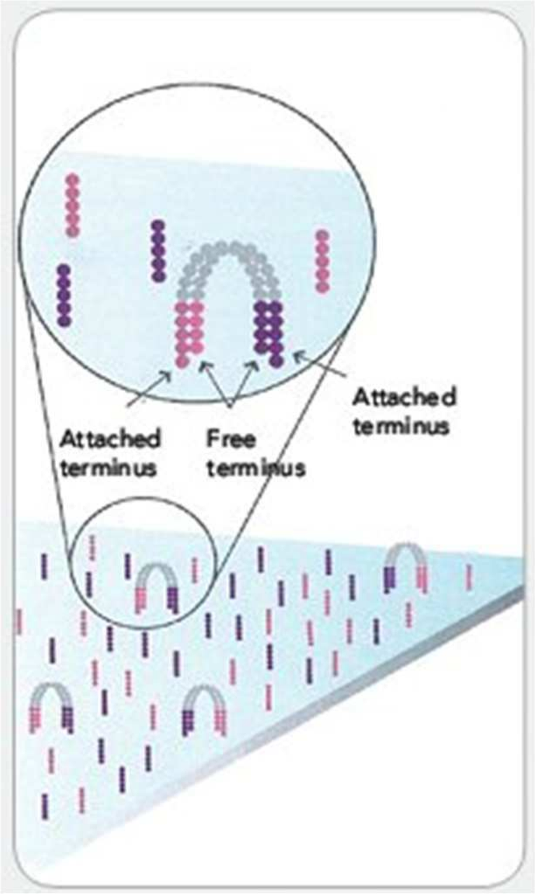
Adapters are attached to randomly fragmented DNA



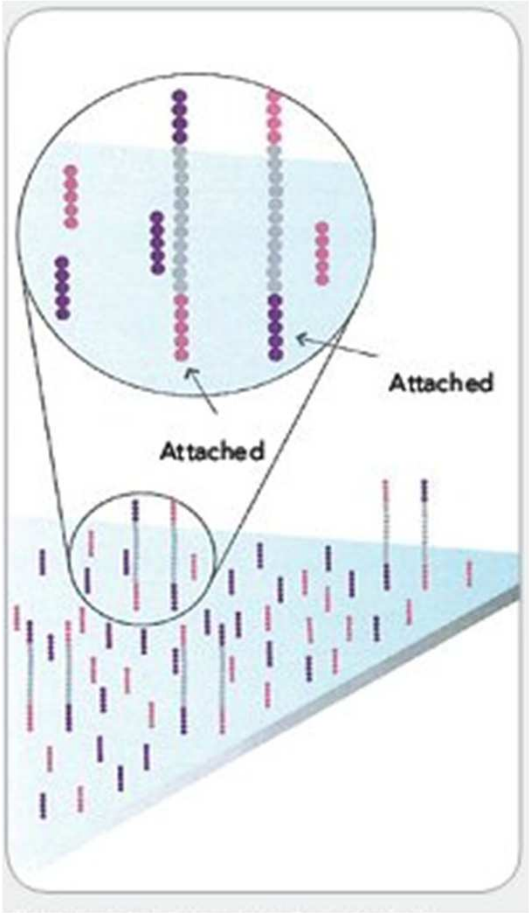
Denatured single-stranded fragments are bound to the surface of the flow cell



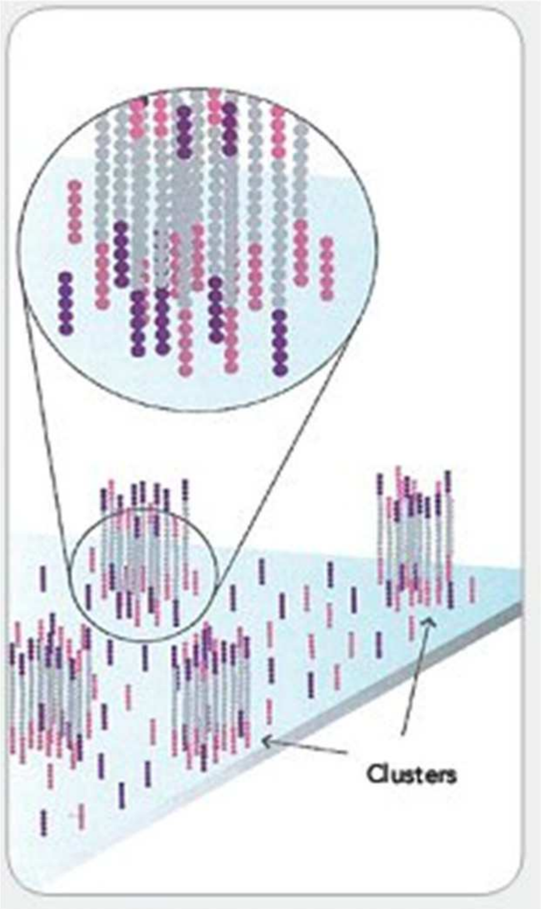
Unlabeled nucleotides and DNA polymerase are added



Polymerases uses nucleotides to build double stranded bridges on the surface



Bridges are denatured



The single stranded DNA are amplified to millions of identical single-stranded DNA

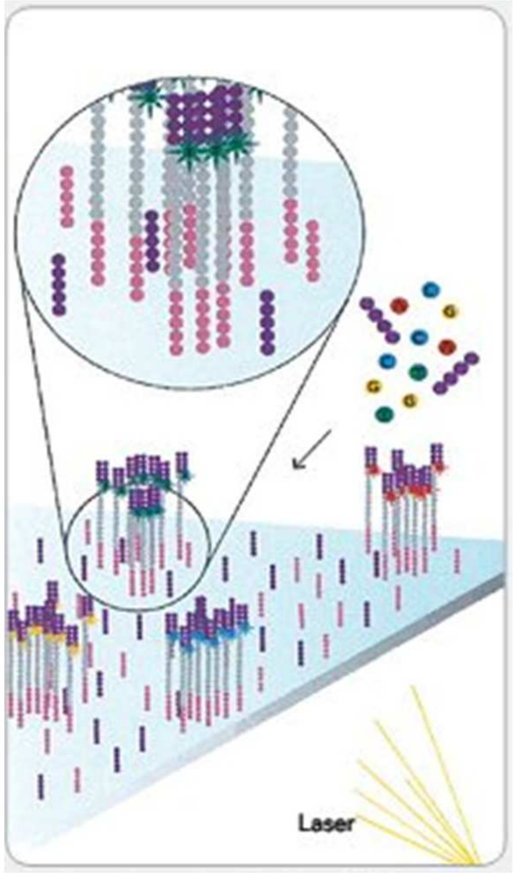
Sequencing

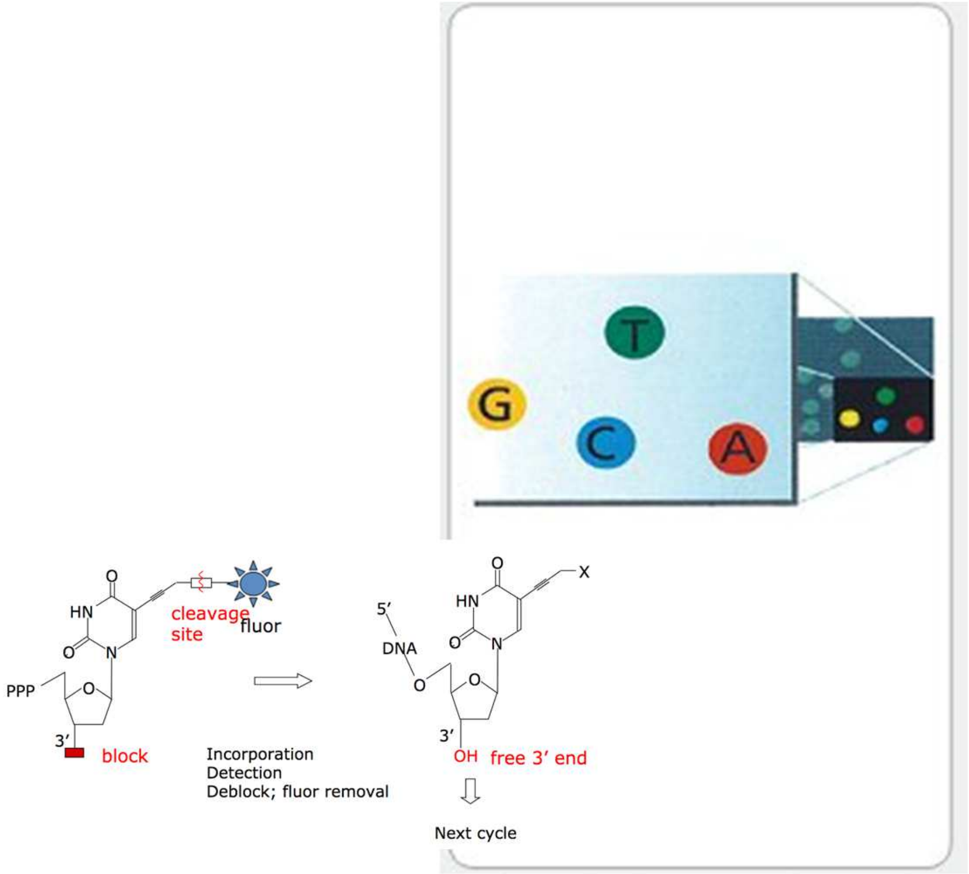
- For metagenomic analysis, rapid DNA sequence reading is needed
  - Because thousands ~ millions of sequences should be read

- High-throughput sequencing (a.k.a. next-generation sequencing) is used

- Currently-used NGS techniques
  - 454 sequencing (pyrosequencing)
  - Illumina sequencing
  - Single molecule, real-time (SMRT) sequencing
  - Nanopore sequencing
  - …

Sequencing example – Illumina seq.



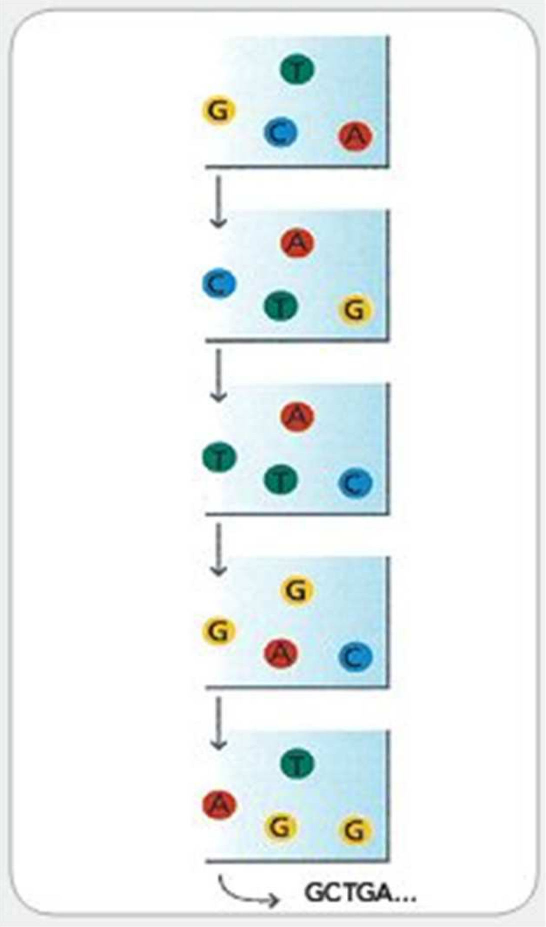


Incorporation
Detection
Deblock; fluor removal

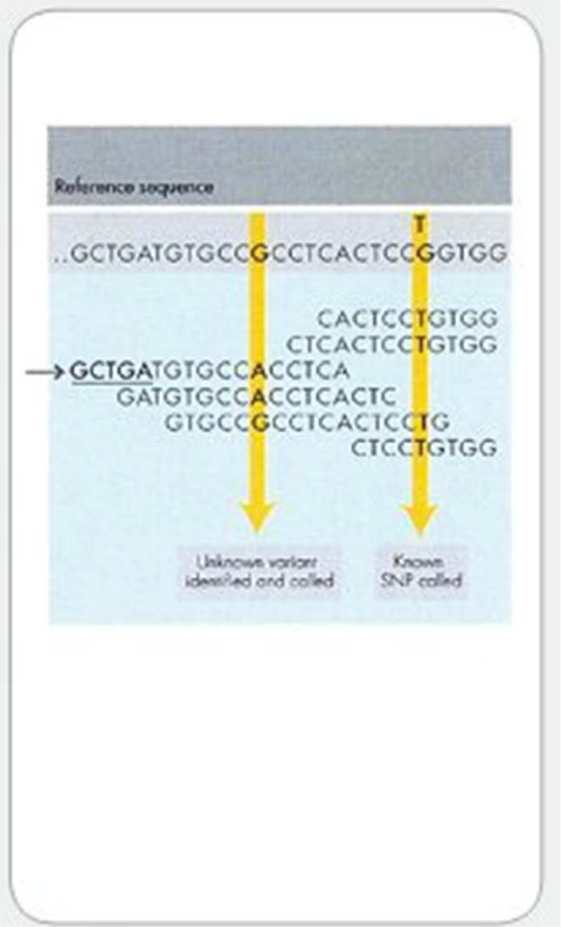
Next cycle

Labeled reversible terminators, primers and DNA polymerase are added. Polymerization stops after the first nucleotide

After laser excitation, the fluorescence from each cluster is detected



Repetition for sequencing



Analysis by aligning of the short sequences

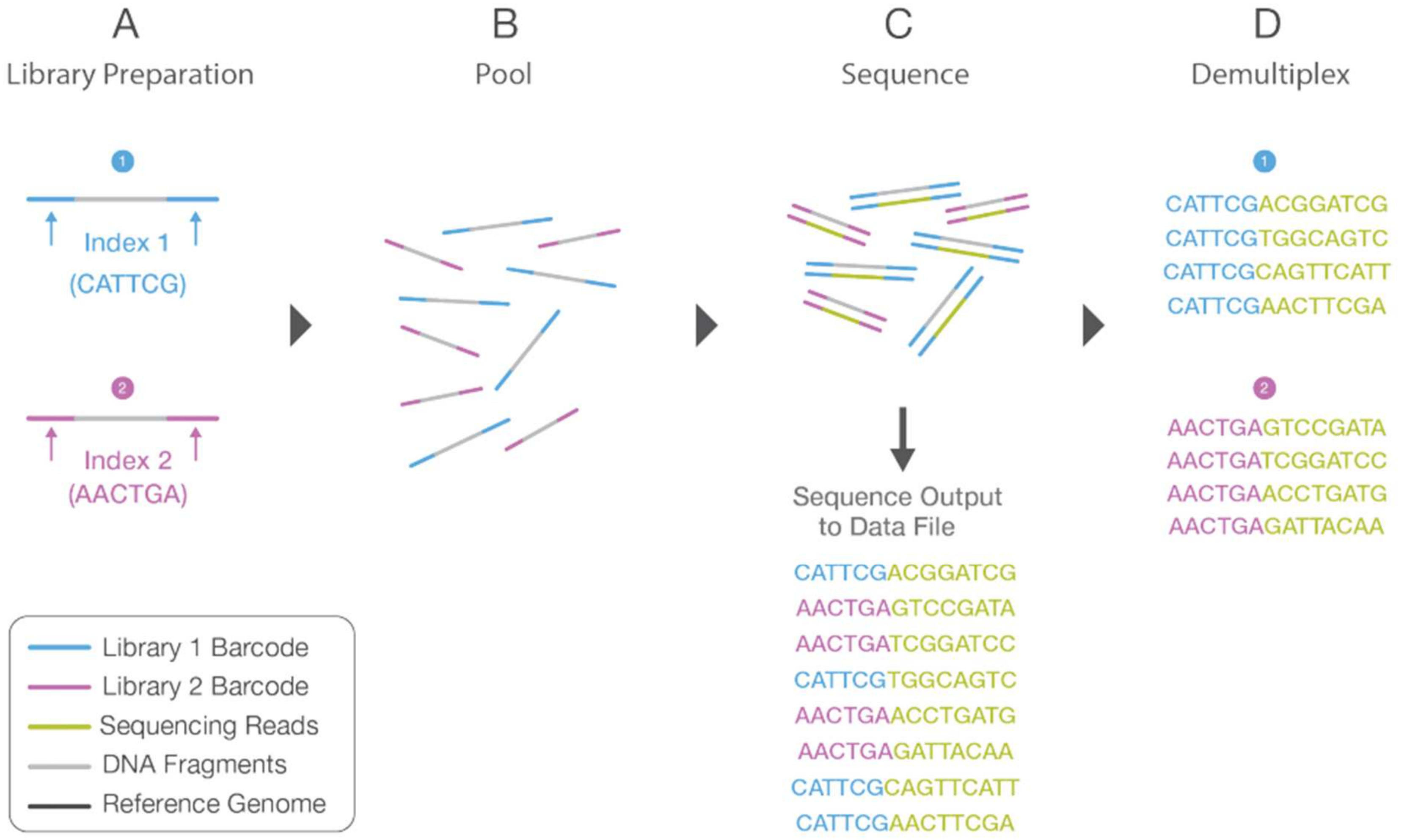
Data pretreatment

- From the raw sequence reads
  - Remove adapters and linkers
  - Exclude chimeras* and replication
  - Demultiplex barcoded samples

** artifact sequences formed by two or more sequences incorrectly joined together*

- Many software tools are available

Demultiplexing of multiplexed sample



CATTCGACGGATCG
AACTGAGTCCGATA
AACTGATCGGATCC
CATTCGTGGCAGTC
AACTGAACCTGATG
AACTGAGATTACAA
CATTCGCAGTTCATT
CATTCGAACTTCGA

Bioinformatics analysis

- First, sequences of close similarity should be grouped or be represented by one sequence because
  - 16S rRNA sequence of microorganisms that belong to the same species may not be exactly the same
  - Errors occur during DNA amplification and sequencing

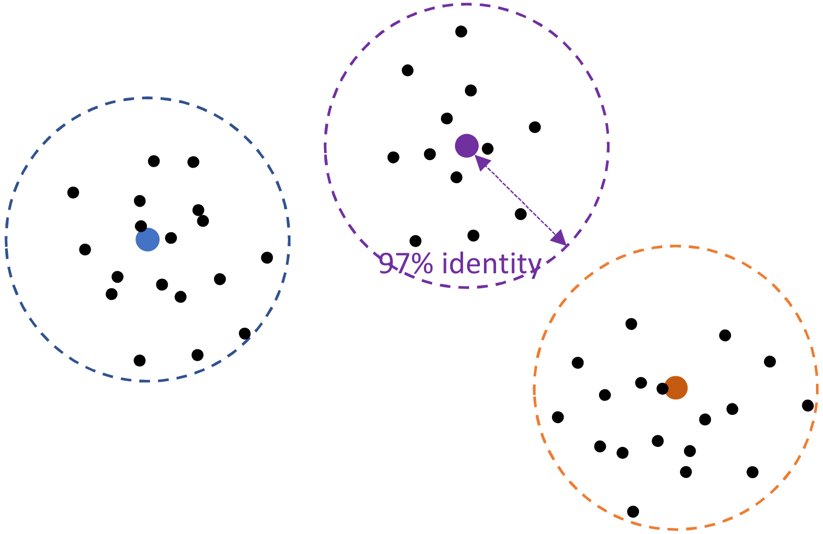
- OTU vs. ASV approach

Operational Taxonomic Units (OTU) approach

1) Group around known sequences from databases

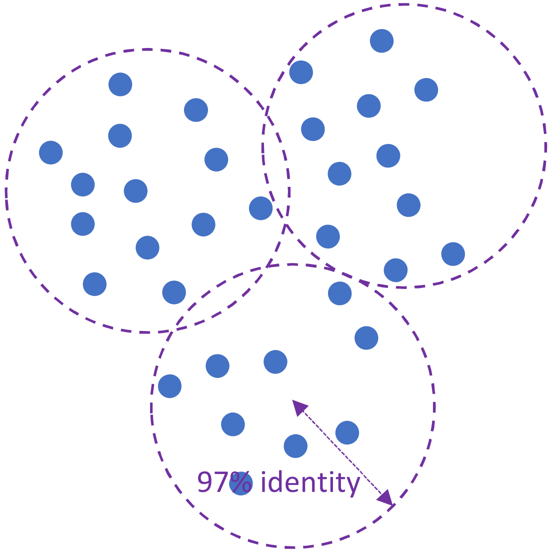
silva

high quality ribosomal RNA databases



- Each group is called as an OTU

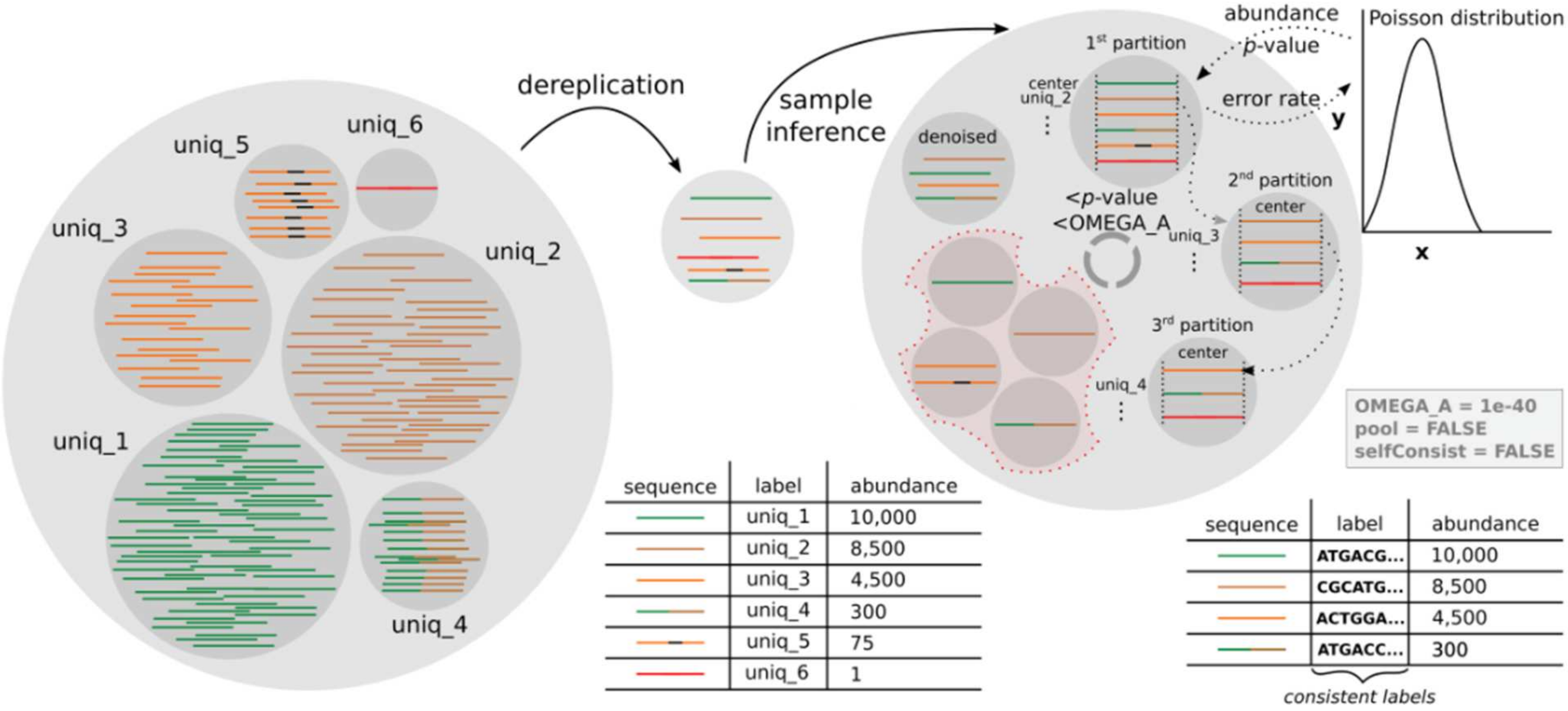
2) Left-over sequences grouped via *de novo* clustering



…..AGTGCGGTAAGCGGACTATC….
…..AGTGCGATAAGAGGACTTTC….
…..AGTGCGATAAGGGGACTATC….
…..AGTGCGATAAGCGGACTATC….
Consensus: …..AGTGCGATAAGCGGACTATC….

Amplicon sequence variants (ASV) approach

- Find the most likely single sequence of a cluster using statistics
- Each sequence is called as an ASV



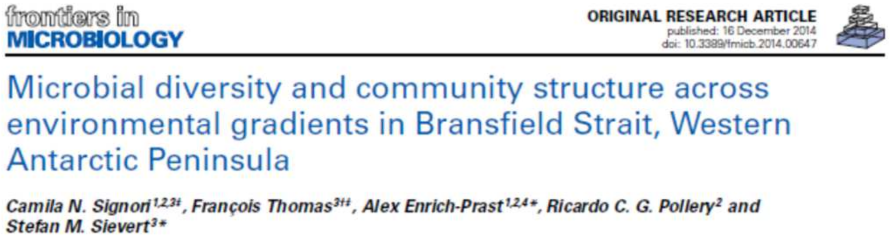
```
OMEGA_A = 1e-40
pool = FALSE
selfConsist = FALSE
```

sequence	label	abundance
🟩	uniq_1	10,000
🟫	uniq_2	8,500
🟧	uniq_3	4,500
🟩🟫	uniq_4	300
🟧⬛🟧	uniq_5	75
🟥	uniq_6	1

sequence	label	abundance
🟩	**ATGACG...**	10,000
🟫	**CGCATG...**	8,500
🟧	**ACTGGA...**	4,500
🟩🟫	**ATGACC...**	300

consistent labels

Bioinformatics analysis – example



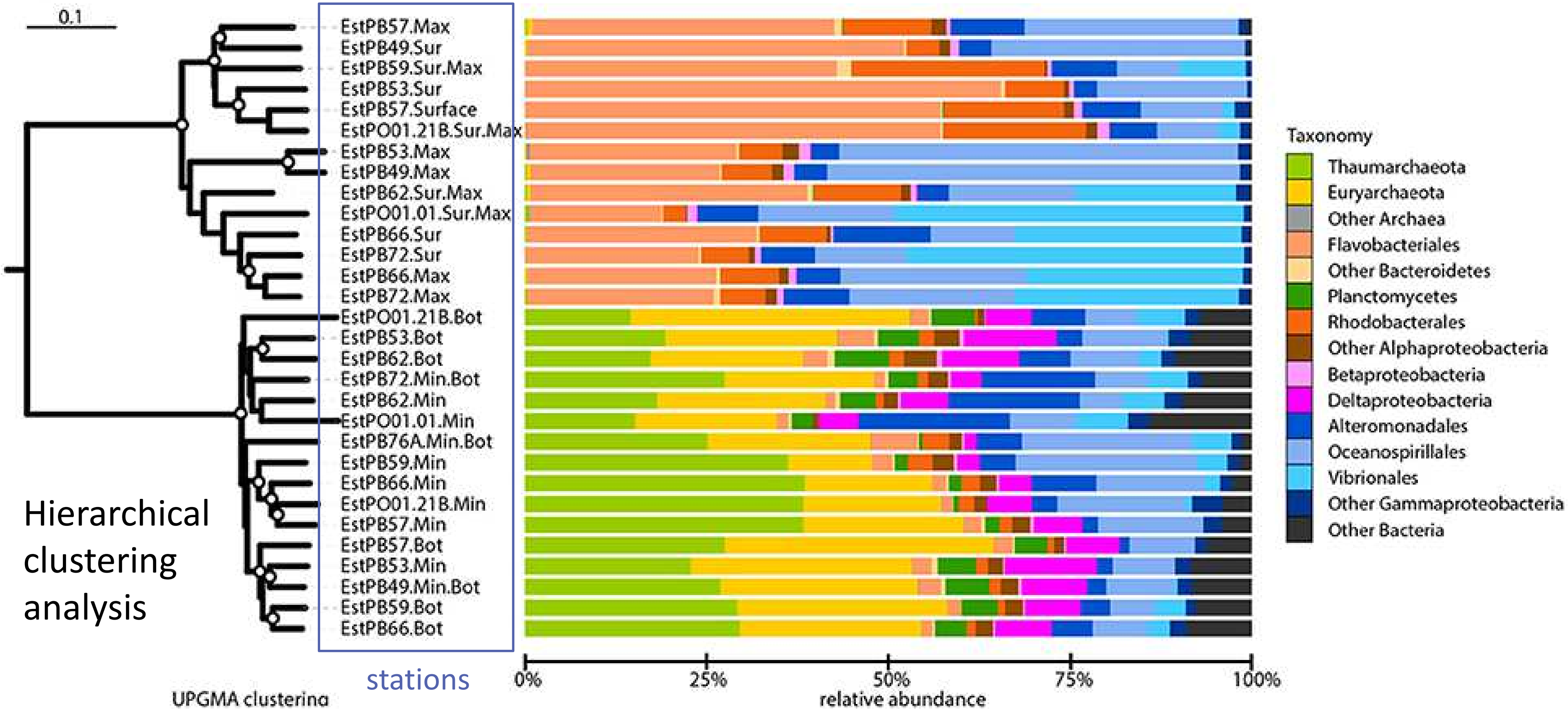
frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 16 December 2014
doi: 10.3389/fmicb.2014.00647

Microbial diversity and community structure across environmental gradients in Bransfield Strait, Western Antarctic Peninsula

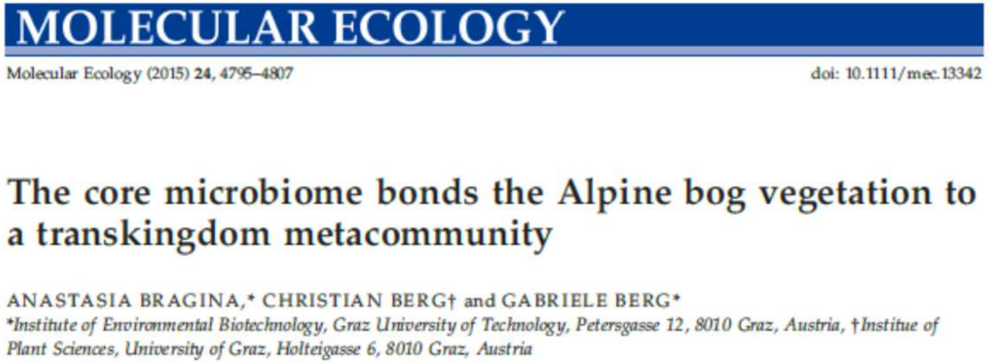
Camila N. Signori[1,2,3†], François Thomas[3††], Alex Enrich-Prast[1,2,4], Ricardo C. G. Pollery[2] and Stefan M. Sievert[3*]*

Relative abundance
(phylum/class/order/family/genus/species level)



Hierarchical clustering analysis

stations



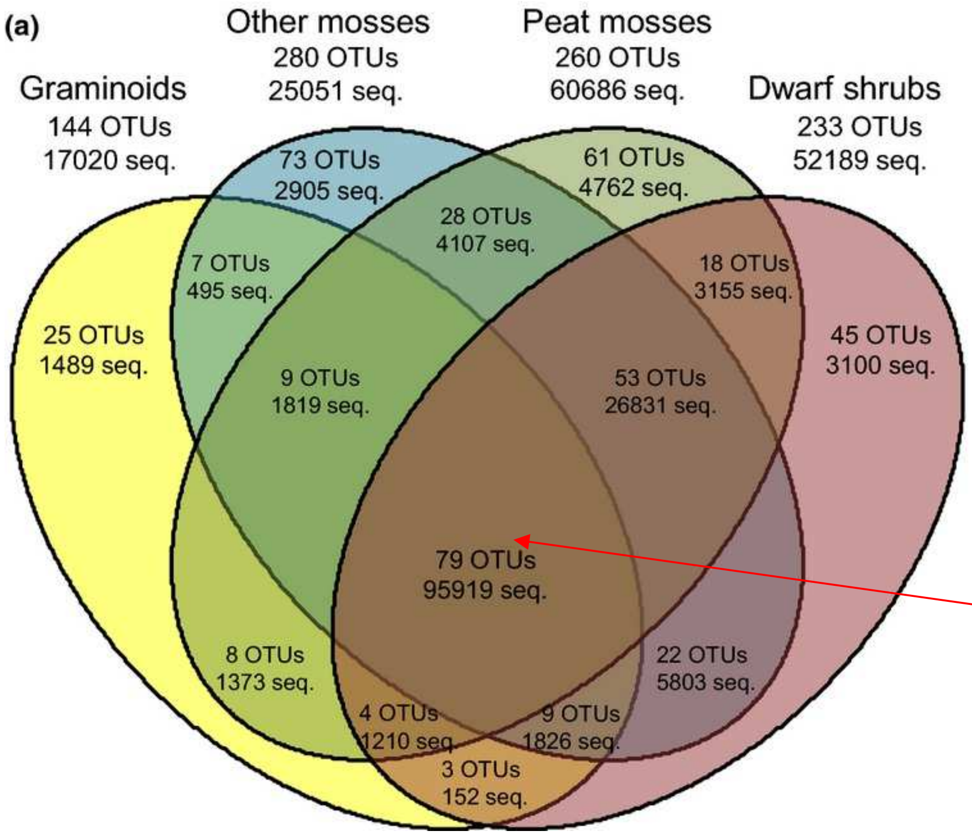
MOLECULAR ECOLOGY

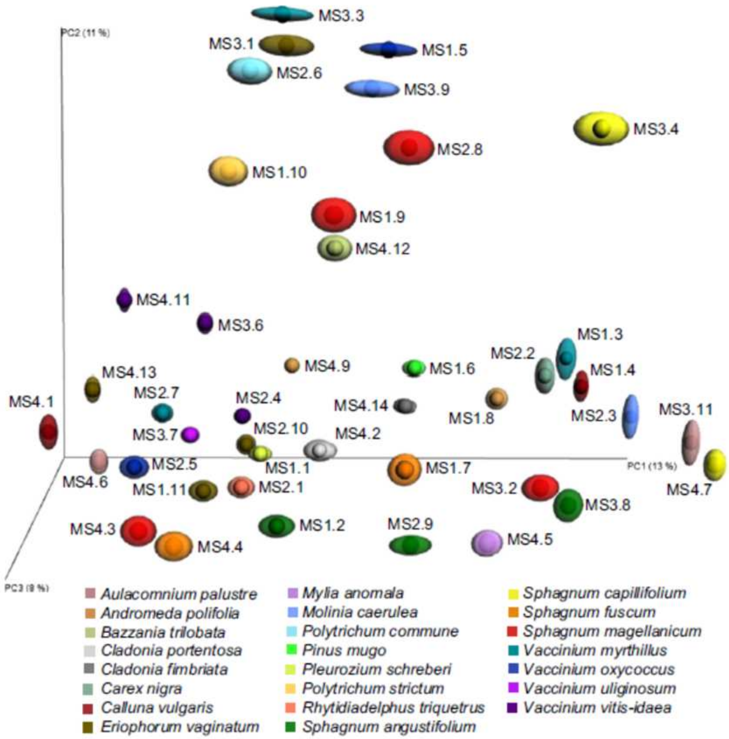
Molecular Ecology (2015) 24, 4795–4807 doi: 10.1111/mec.13342

The core microbiome bonds the Alpine bog vegetation to a transkingdom metacommunity

ANASTASIA BRAGINA,* CHRISTIAN BERG† and GABRIELE BERG*
*Institute of Environmental Biotechnology, Graz University of Technology, Petersgasse 12, 8010 Graz, Austria, †Institue of Plant Sciences, University of Graz, Holteigasse 6, 8010 Graz, Austria

Venn diagram





Principal component analysis

Group of OTUs/ASVs common to the defined subgroups of environments

Table 3
Equation and index meaning of parameters for alpha diversity analysis.

Index	Equation	Note
Chao1 index	$S_{chao1} = S_{obs} + \frac{n_1(n_1-1)}{2(n_2+1)}$	S_{chao1} and S_{obs} are the estimated and observed OUTs number, respectively; n_1 and n_2 are the number of OTUs with 1 and 2 sequences, respectively.
ACE index	$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{n_1}{C_{ACE}} \cdot \gamma_{ACE}^2 \quad (\text{for} \gamma_{ACE} < 0.8);$ $S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{n_1}{C_{ACE}} \cdot \beta_{ACE}^2 (\text{for} \gamma_{ACE} \geq 0.8).$ $C_{ACE} = 1 - \frac{n1}{N_{rare}}$ $N_{rare} = \sum_{i=1}^{abund} i \cdot n_i$ $\gamma_{ACE}^2 = \max\left[\frac{S_{rare}}{C_{ACE}} \frac{\sum_{i=1}^{abund} i(i-1)n_i}{N_{rare}(N_{rare}-1)} - 1, \quad 0\right]$ $\beta_{ACE}^2 = \max\left[\gamma_{ACE}^2 \left\{1 + \frac{N_{rare}(1-C_{ACE}) \sum_{i=1}^{abund} i(i-1)n_i}{N_{rare}(N_{rare} - C_{ACE})}\right\}, \quad 0\right]$	n_i is the number of OTUs containing *i* sequences; S_{rare} is the number of OTUs containing '*abund*' number of sequences or less; S_{abund} is the number of OTUs containing more than '*abund*' number of sequences; '*abund*' is threshold value of dominant OTU.
Shannon- Wiener index	$H_{shannon} = -\sum_{i=1}^{S_{obs}} \frac{n_i}{N} \ln \frac{n_i}{N}$	S_{obs} is actually observed OTUs number; n_i is the number of sequences in No. *i* of OTU; *N* is total number of sequences.
Simpson index	$D_{simpson} = \frac{\sum_{i=1}^{S_{obs}} n_i(n_i-1)}{N(N-1)}$	S_{obs} is actually observed OTUs number; n_i is the number of sequences in No. *i* of OTU; *N* is total number of sequences.
Good's coverage	$C = 1 - \frac{n_1}{N}$	n_1 is the number of OTUs with 1 sequence; *N* is total number of sequences.

Suppl. info (1): Types of diversity

- Alpha-diversity: mean species diversity in a site at a local scale
- Beta-diversity: ratio between regional and local species diversity
- Gamma-diversity: total species diversity in a landscape
  - Gamma-diversity = alpha-diversity + beta-diversity

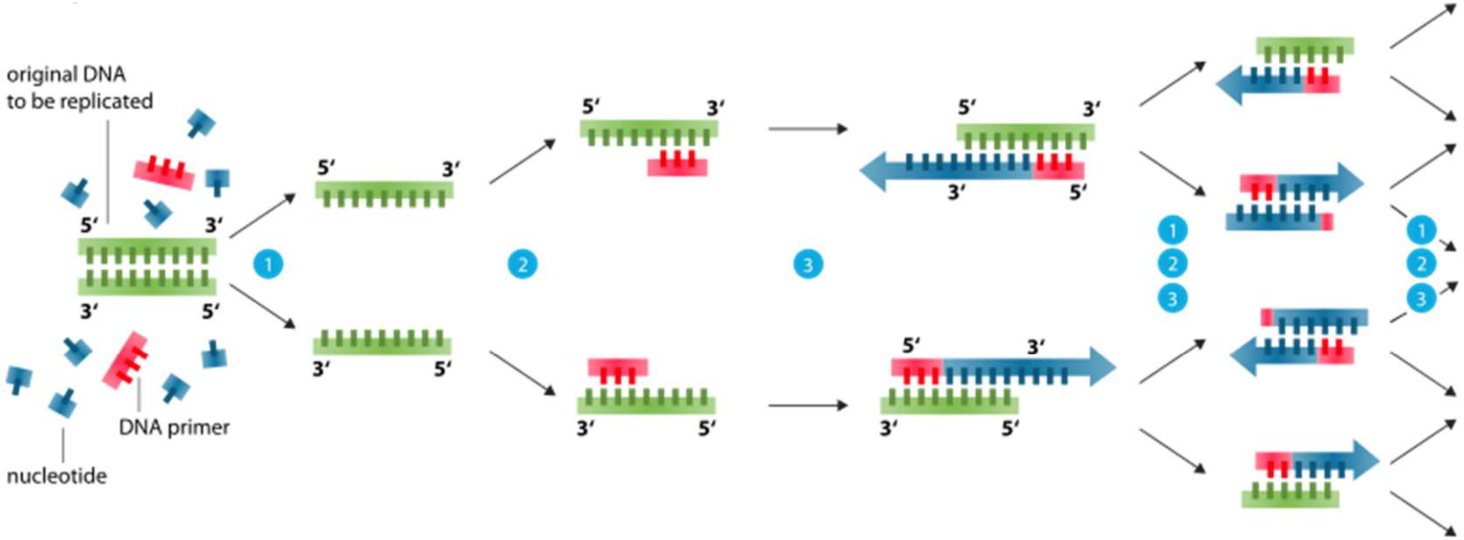
Suppl. info (2): Polymerase chain reaction (PCR)

- Small sections of the extracted DNA are amplified using naturally occurring enzymes involved in cellular DNA replication

- PCR ingredients
  - The sample DNA (template DNA)
  - PCR primers: short oligonucleotides that complement a section of the target DNA sequence
  - DNA polymerase: a naturally occurring enzyme that creates copies of DNA during cell replication
  - Mixture of nucleotides: building blocks for new DNA
  - pH buffer containing Mg^{2+}

- Typical procedure for 1 cycle
  - Heating to about 95°C to separate double stranded DNA into single strands
  - Lower the temp. to 45~65°C to allow PCR primers to anneal to the DNA template
  - Increase the temperature to about 72°C and DNA polymerase extends the copy of the template DNA

- The original DNA is substantially amplified after several tens of cycles



1 **Denaturation** at ~95°C

2 **Annealing** at 45~65°C

3 **Elongation** at ~72°C