Introduction to metabarcoding

Antonino Malacrino

EEOB, The Ohio State University

December, 16 2020

Outline

Introduction

Which is our question?

- Who is in there?
- What they can do?
- What are they doing?
- Are they really doing it?

Which is our question?

- Who is in there?
- What they can do?
- What are they doing?
- Are they really doing it?

Which is our question?

- Who is in there?
- What they can do?
- What are they doing?
- Are they really doing it?

Which is our question?

- Who is in there?
- What they can do?
- What are they doing?
- Are they really doing it?

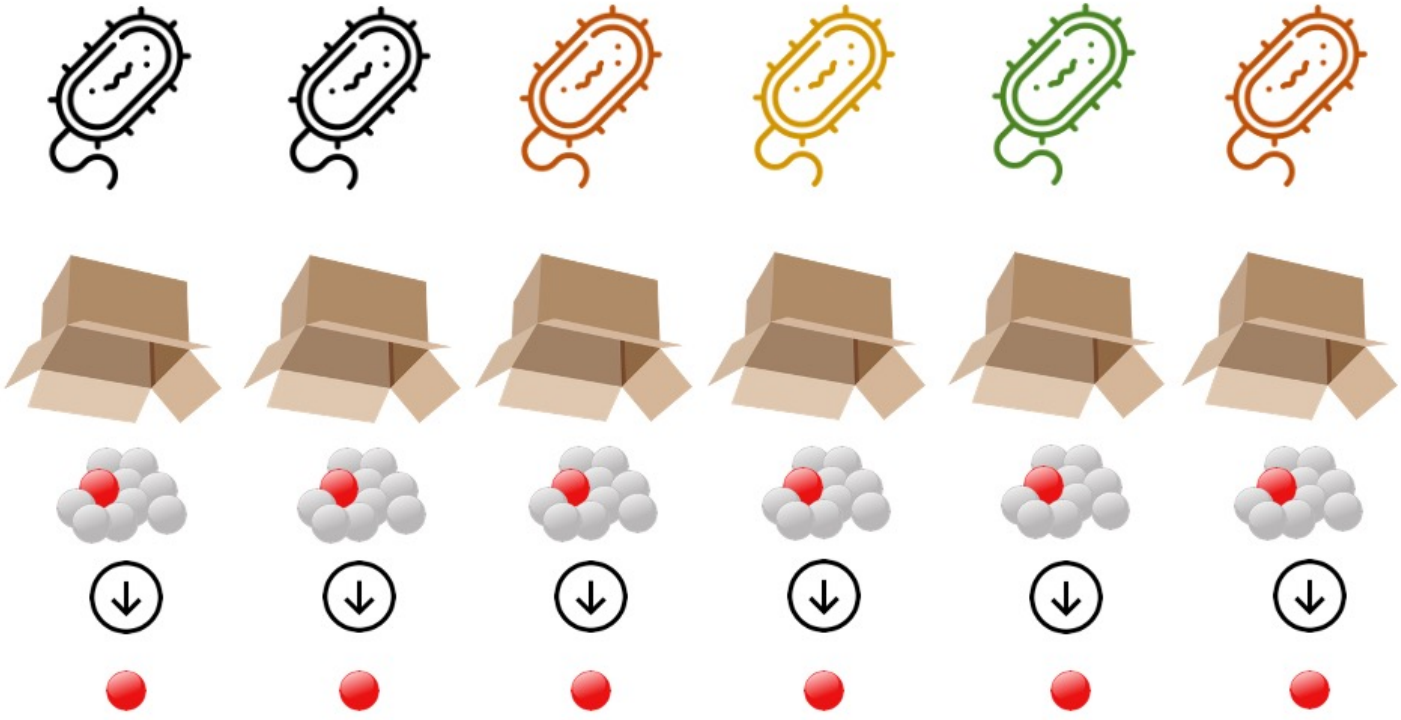
Molecular barcoding

Molecular barcoding

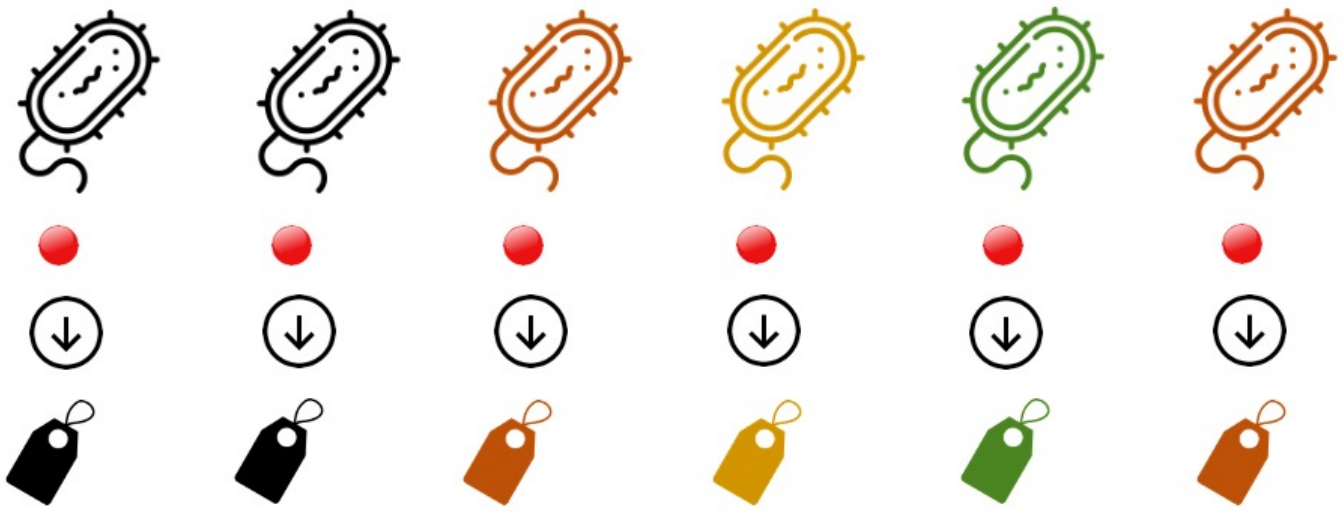
Molecular barcoding

Molecular barcoding

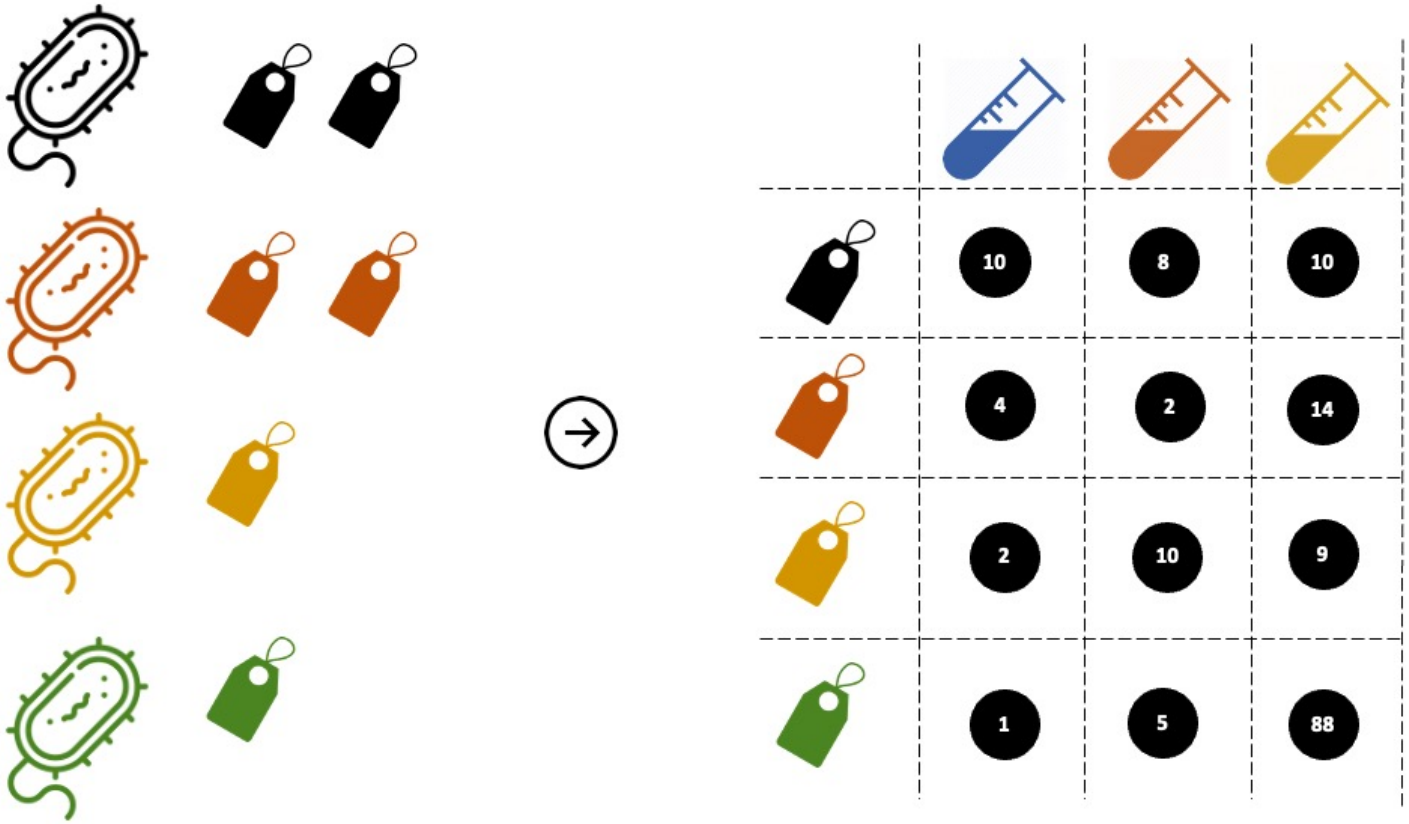
Metabarcoding



Metabarcoding



Metabarcoding



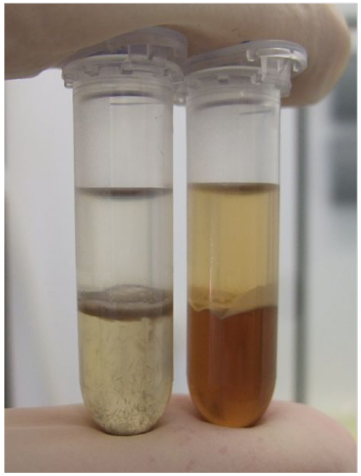
General workflow

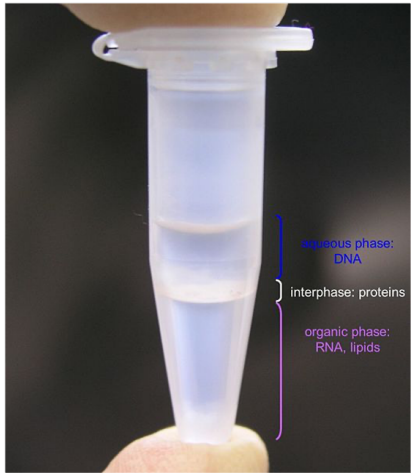
1. Sample collection
2. DNA extraction
3. PCR amplification of our target
4. Adaptor ligation
5. Sequencing
6. Data analysis

DNA extraction

Different methods

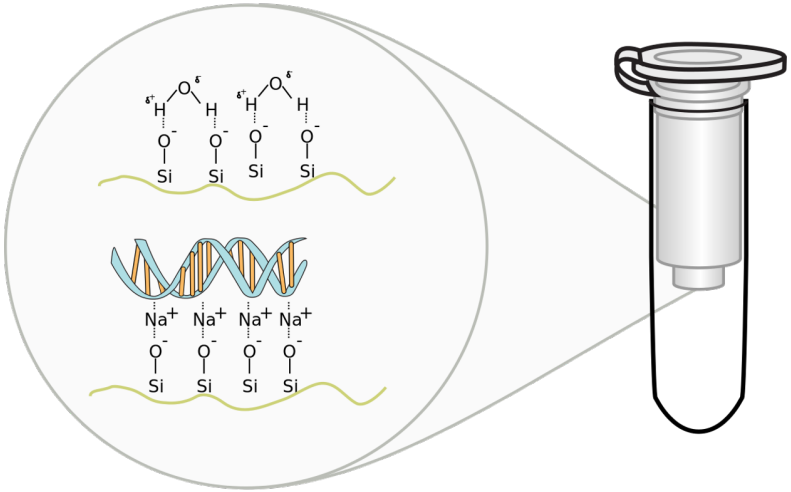
- Organic solvents
- Spin columns
- SPRI beads
- Many other (CTAB, ...)





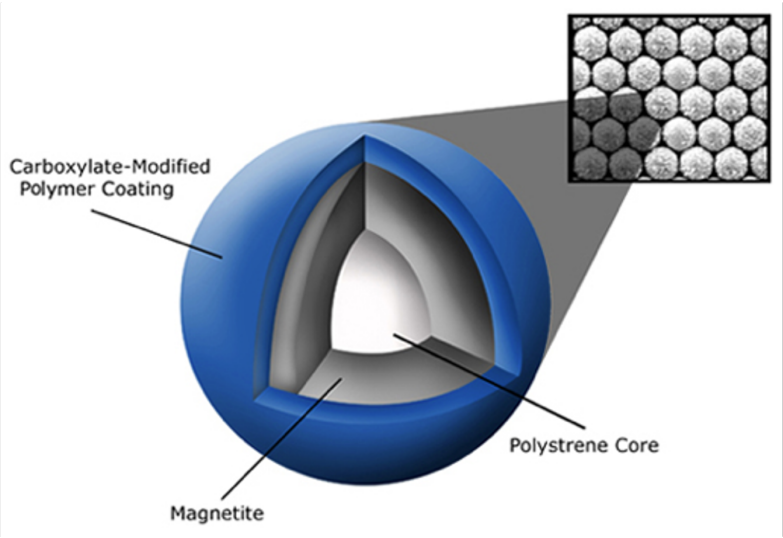
Different methods

- Organic solvents
- Spin columns
- SPRI beads
- Many other (CTAB, ...)



Different methods

- Organic solvents
- Spin columns
- SPRI beads
- Many other (CTAB, ...)

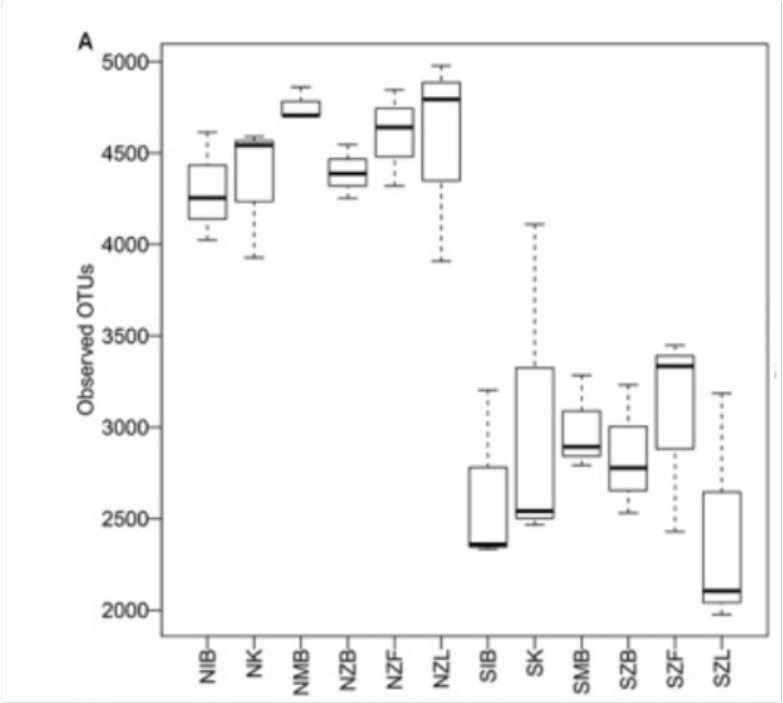


Different methods

- Organic solvents
- Spin columns
- SPRI beads
- Many other (CTAB, ...)

Technical considerations

- DNA extraction introduces a bias in the final dataset
- A recent investigation on 322 studies shows they used 72 different methods. 14 did not report such info!

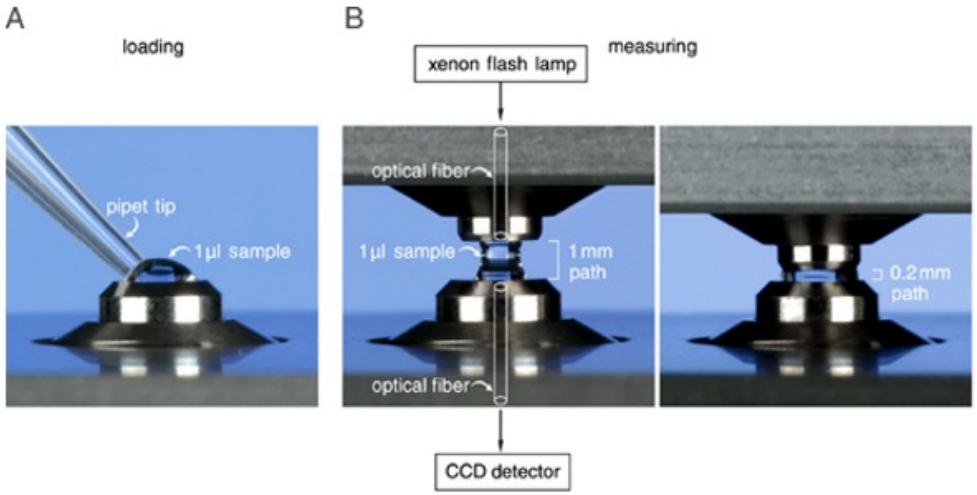


Technical considerations

- DNA extraction introduces a bias in the final dataset
- A recent investigation on 322 studies shows they used 72 different methods. 14 did not report such info!

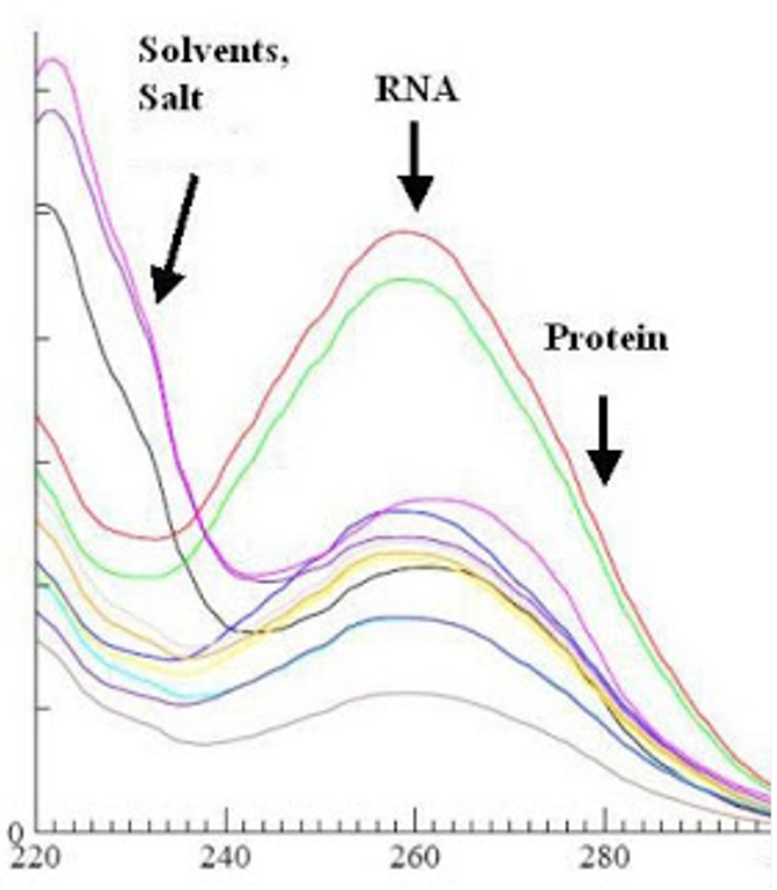
QC - Nanodrop


Thermo
SCIENTIFIC
NANODROP 2000


A
loading
B
measuring
xenon flash lamp
pipet tip
1 µl sample
optical fiber
1 µl sample
1 mm
path
0.2 mm
path
optical fiber
CCD detector

QC - Nanodrop

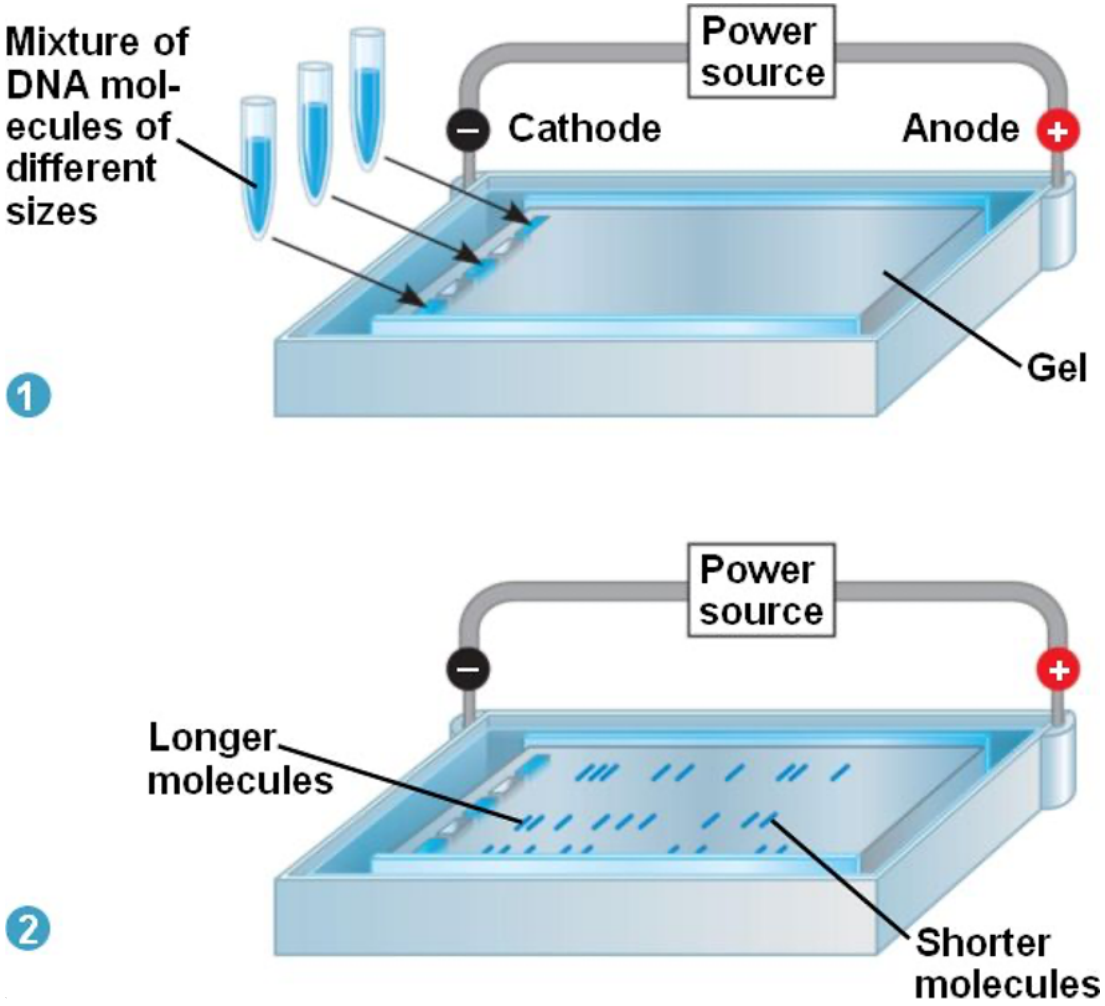
- DNA concentration ng/ul
- Ratio 260/230
- Ratio 260/280



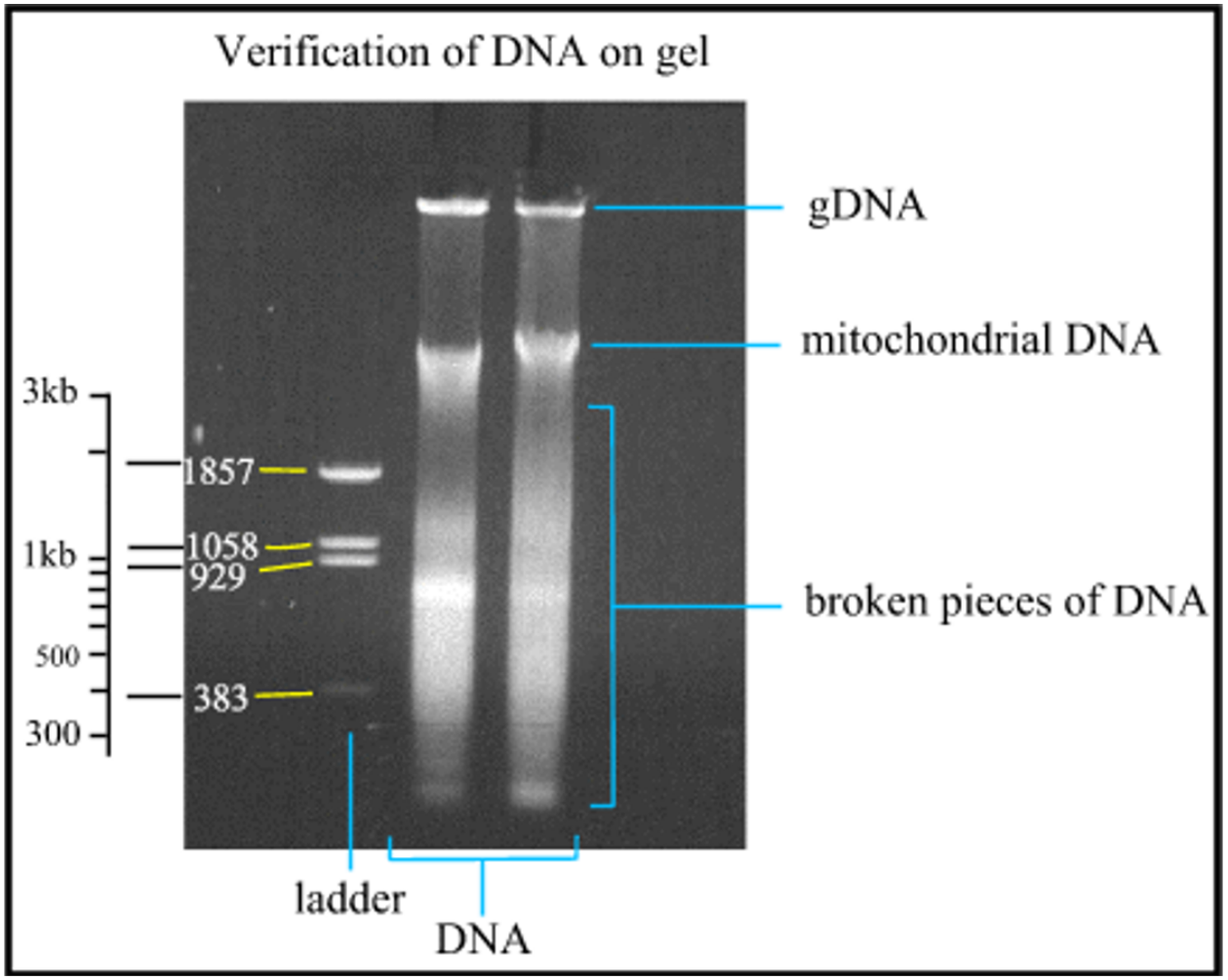
QC - Qubit


life
technologies
invitrogen
Qubit® 3.0
Fluorometer
life
technologies
Qubit 3.0
Fluorometer

QC - Gel Electrophoresis

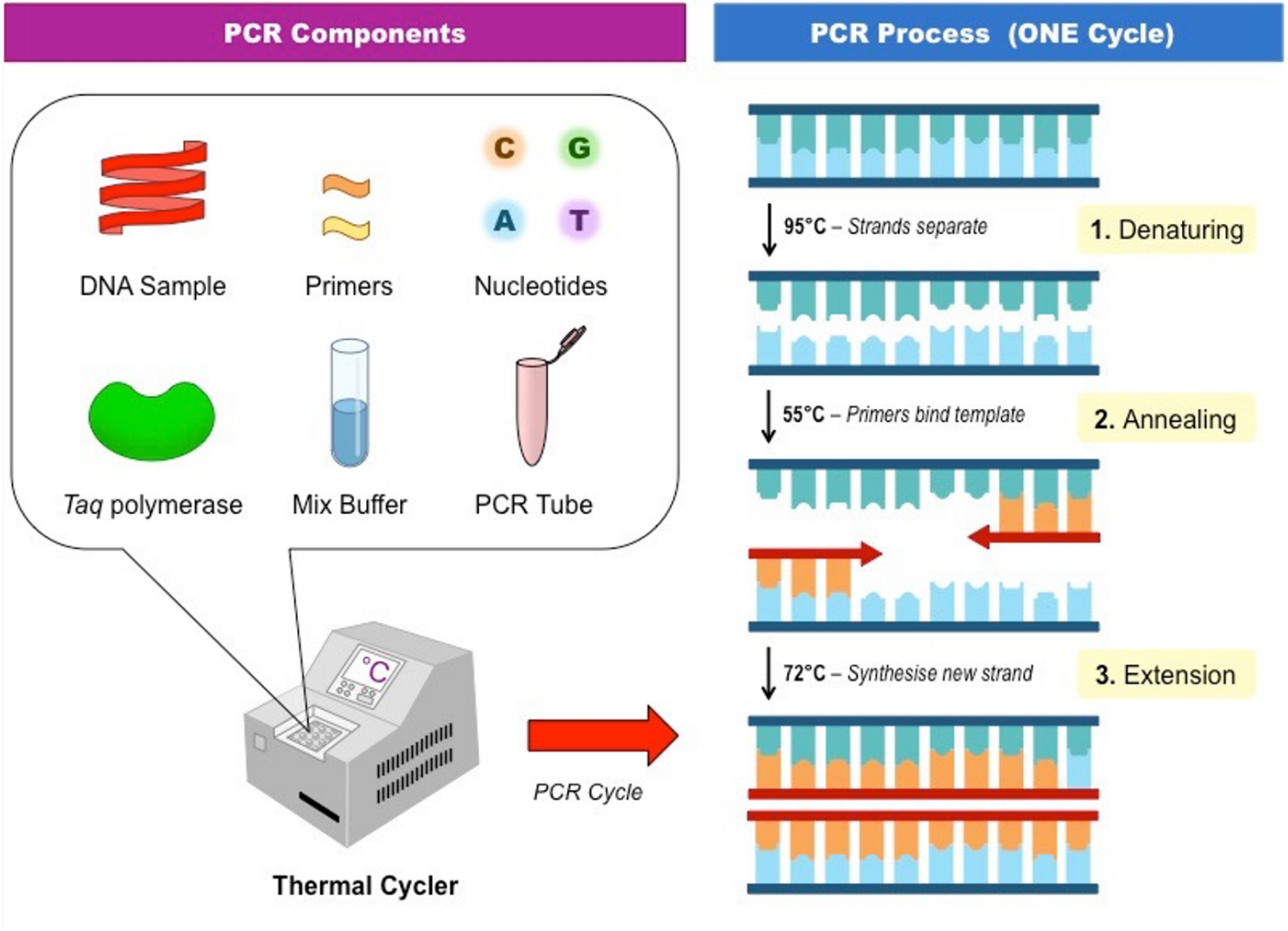


QC - Gel Electrophoresis



PCR

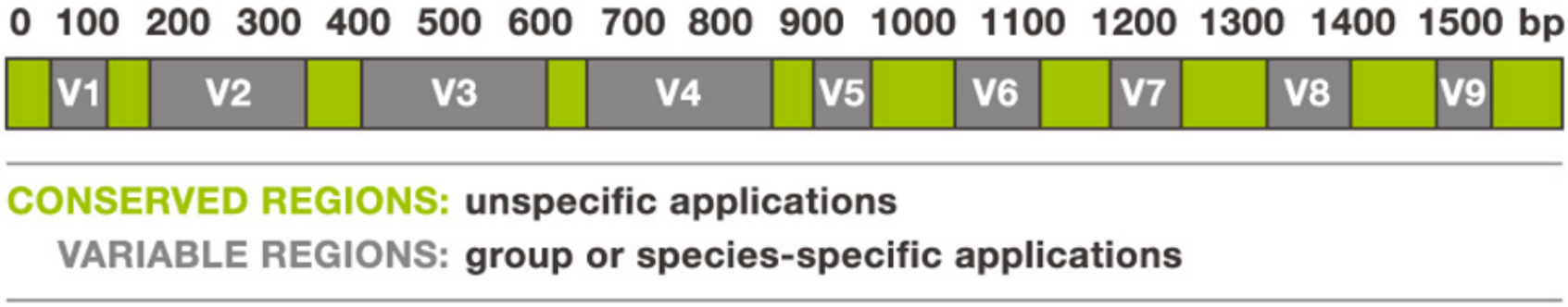
PCR


PCR Components
PCR Process (ONE Cycle)
DNA Sample
Primers
C
G
A
T
Nucleotides
Taq polymerase
Mix Buffer
PCR Tube
°C
Thermal Cycler
PCR Cycle
95°C – Strands separate
1. Denaturing
55°C – Primers bind template
2. Annealing
72°C – Synthesise new strand
3. Extension

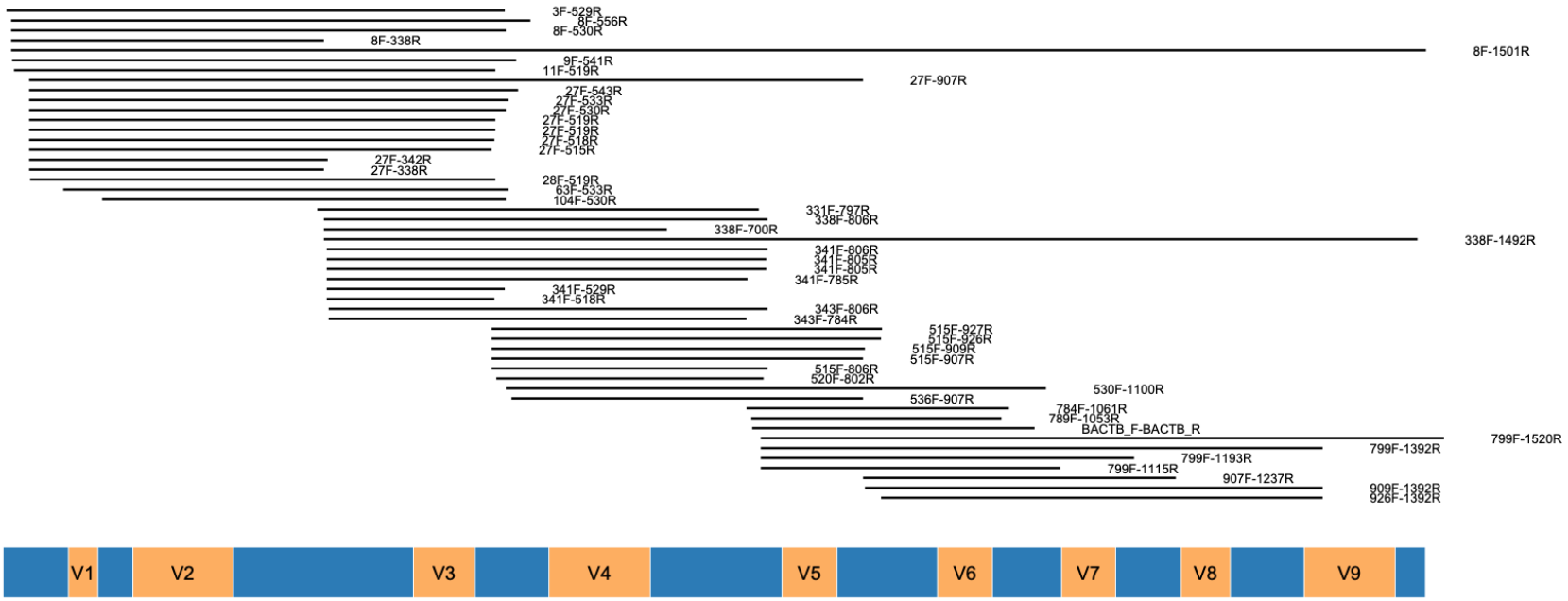
Thermocycler

Technical considerations

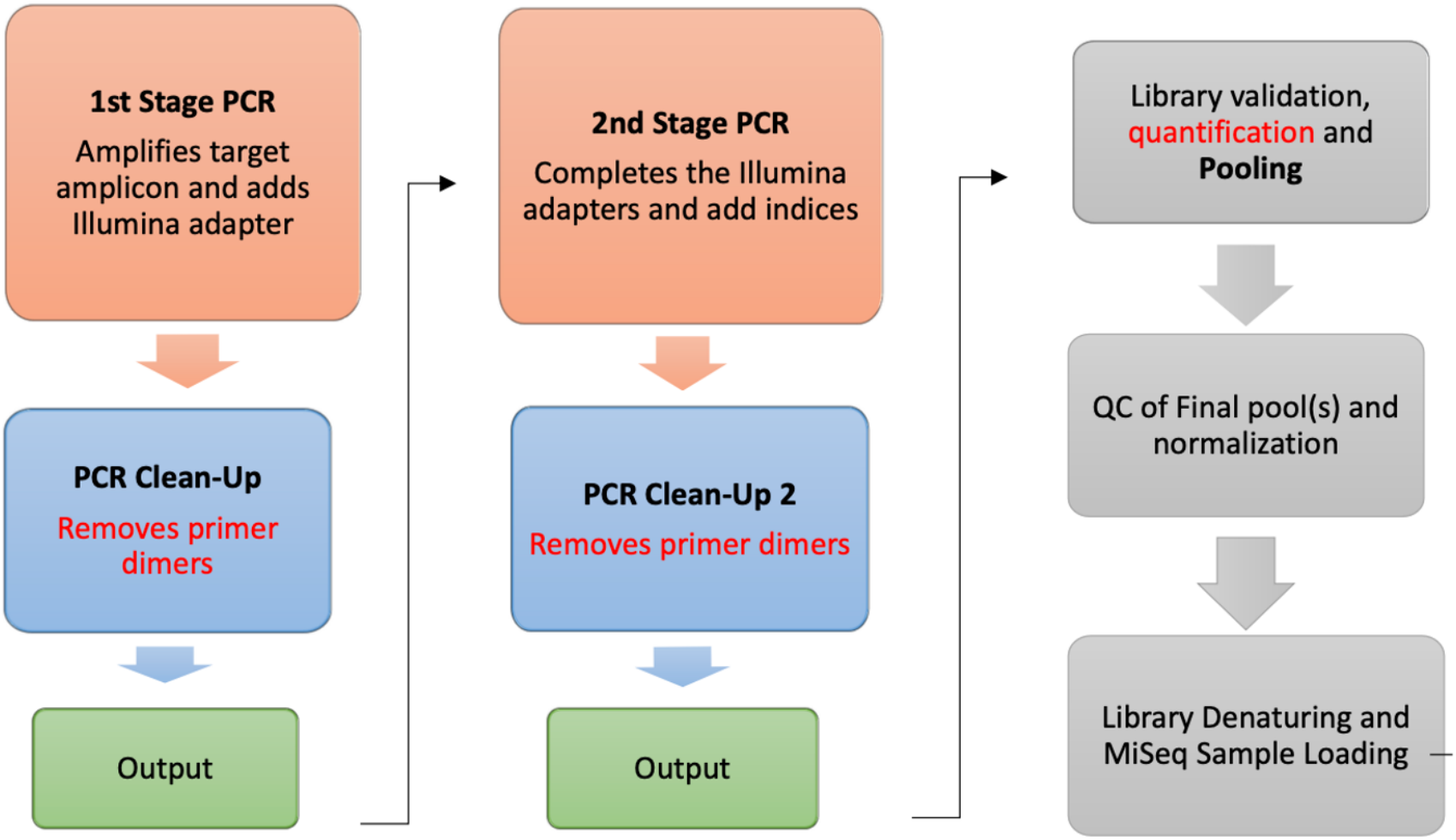
- Target gene
  - Who is our target? Bacteria, fungi, insects, fish, specific genus
  - Is the resolution optimal for our question?
  - Are PCR primers available or we have to design them?
  - Is the taxonomy database available or we have to build a custom one?
- PCR bias
  - Use a Hi-Fi polymerase
  - Use optimal annealing temperature
  - Do not exaggerate with PCR cycles
  - Run multiple PCRs on the same sample



Technical considerations



Metabarcoding workflow


1st Stage PCR
Amplifies target amplicon and adds Illumina adapter
PCR Clean-Up
Removes primer dimers
Output
2nd Stage PCR
Completes the Illumina adapters and add indices
PCR Clean-Up 2
Removes primer dimers
Output
Library validation, quantification and Pooling
QC of Final pool(s) and normalization
Library Denaturing and MiSeq Sample Loading

Control samples

- **Negative control 1.** Run molecular biology grade water throughout the pipeline. Pool it with the other samples even if you do not see amplification.
- **Negative control 2.** This is the negative control from your first PCR. If you see a band, discard the entire batch of samples and start again. If no band is observed, sequence anyway.
- **Mock community.** Pre-built or custom.

Control samples

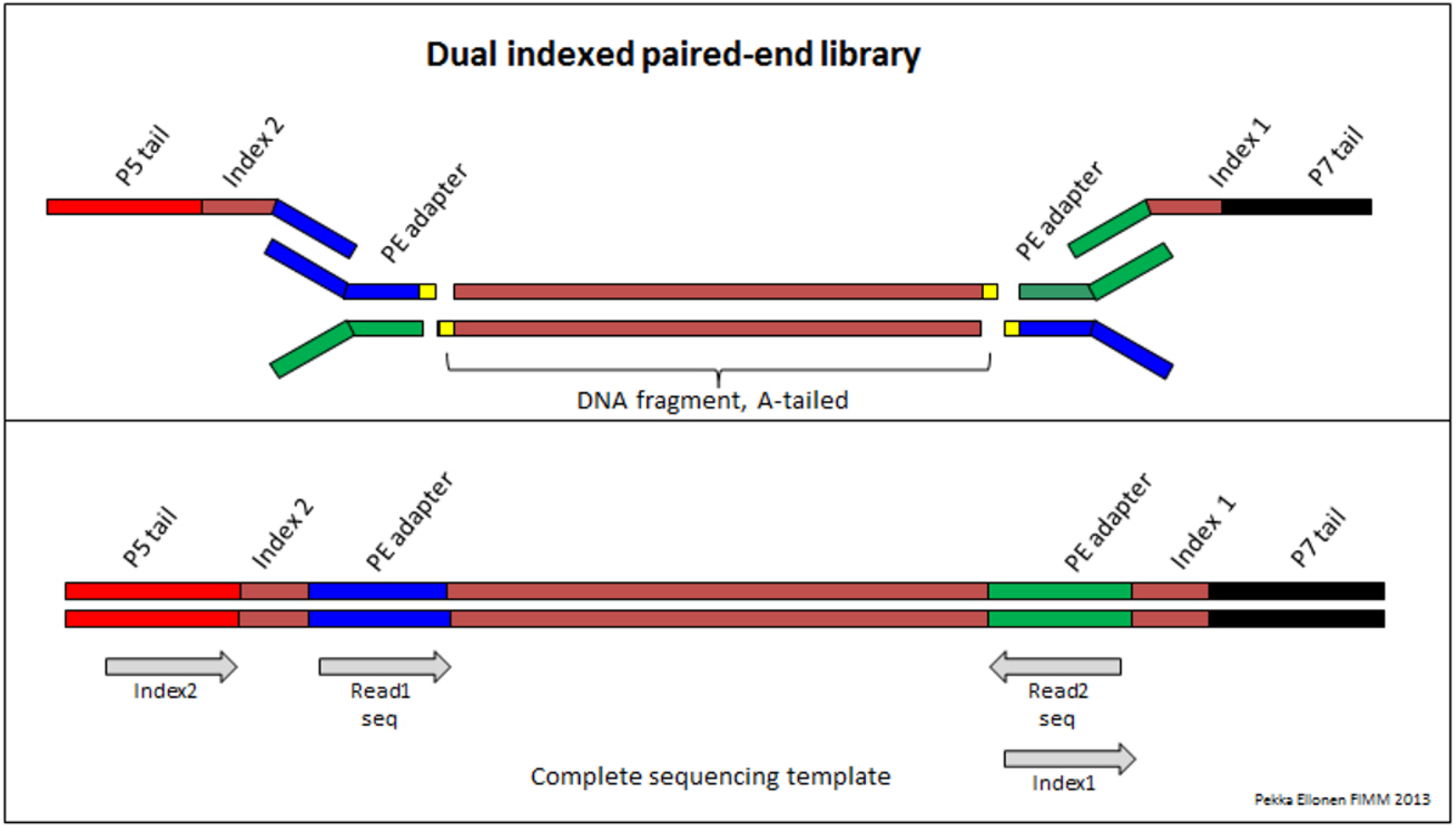
- **Negative control 1.** Run molecular biology grade water throughout the pipeline. Pool it with the other samples even if you do not see amplification.
- **Negative control 2.** This is the negative control from your first PCR. If you see a band, discard the entire batch of samples and start again. If no band is observed, sequence anyway.
- **Mock community.** Pre-built or custom.

Control samples

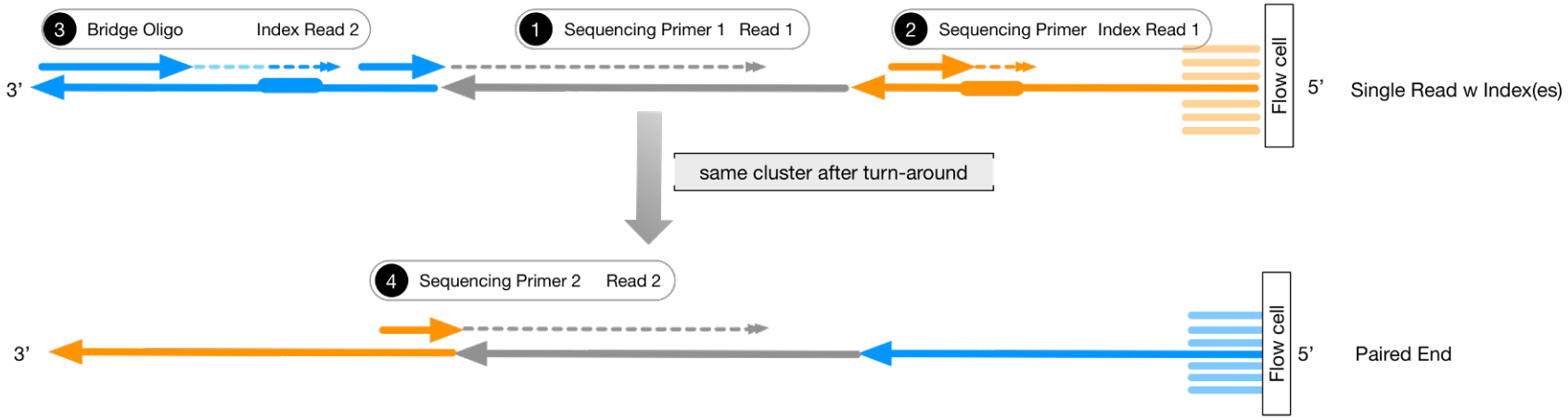
- **Negative control 1.** Run molecular biology grade water throughout the pipeline. Pool it with the other samples even if you do not see amplification.
- **Negative control 2.** This is the negative control from your first PCR. If you see a band, discard the entire batch of samples and start again. If no band is observed, sequence anyway.
- **Mock community.** Pre-built or custom.

Library preparation

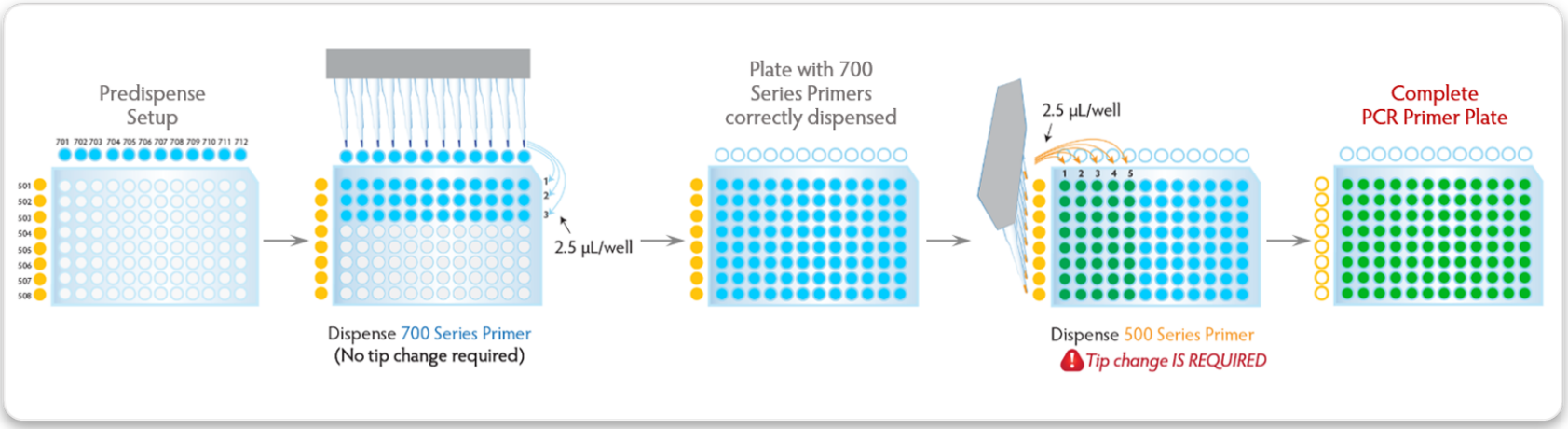
Multiplexing



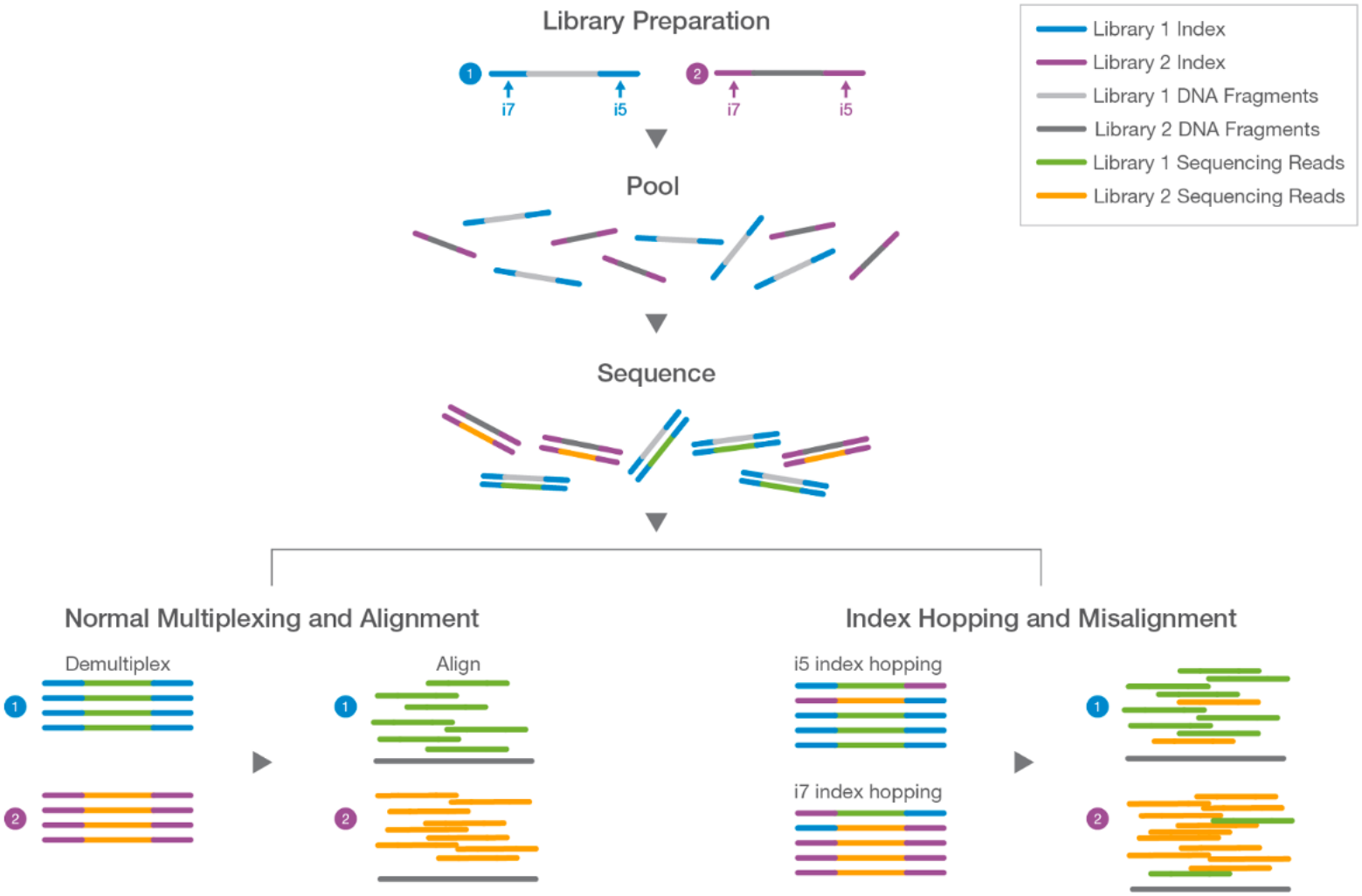
Multiplexing



Multiplexing



Index hopping



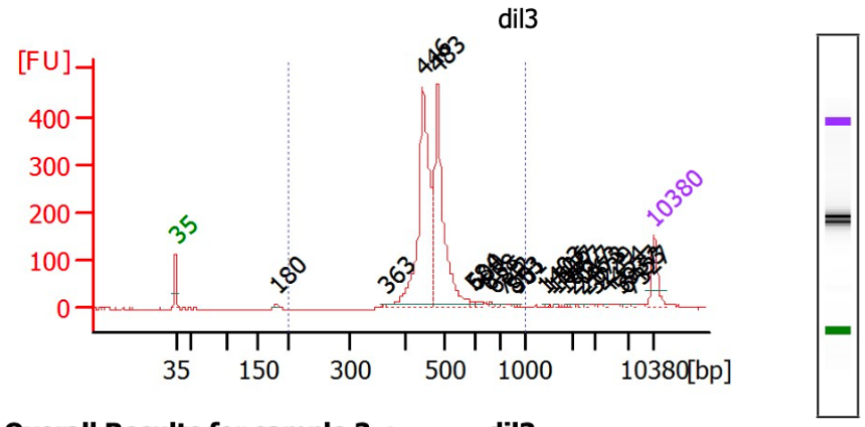
Equimolar pooling

Goal: guarantee that all samples are sequenced at the same depth

1. Qubit
2. Calculate nM
3. Pool according to sample concentration

Final quality control

- Bioanalyzer / Tapestation
- qPCR
- Qubit



Overall Results for sample 3 : dil3

Number of peaks found:	23
Noise:	1.2
Corr. Area 1:	2,193.1

Region table for sample 3 : dil3

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	2,193.1	94	477	12.8	1,496.61	4,823.2	

Illumina sequencing technology

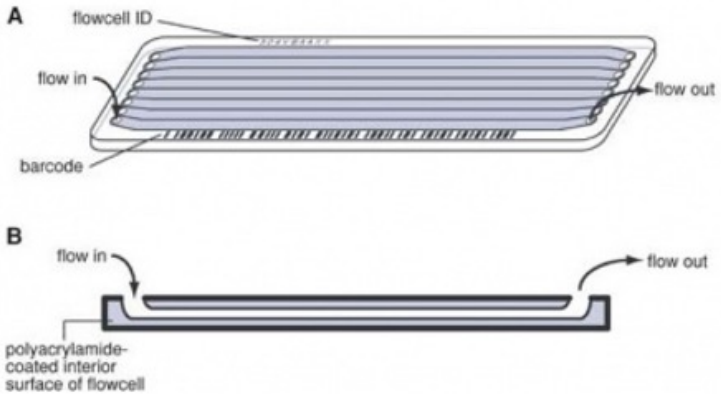
Illumina sequencing platforms

From genome-wide discovery to targeted validation and screening

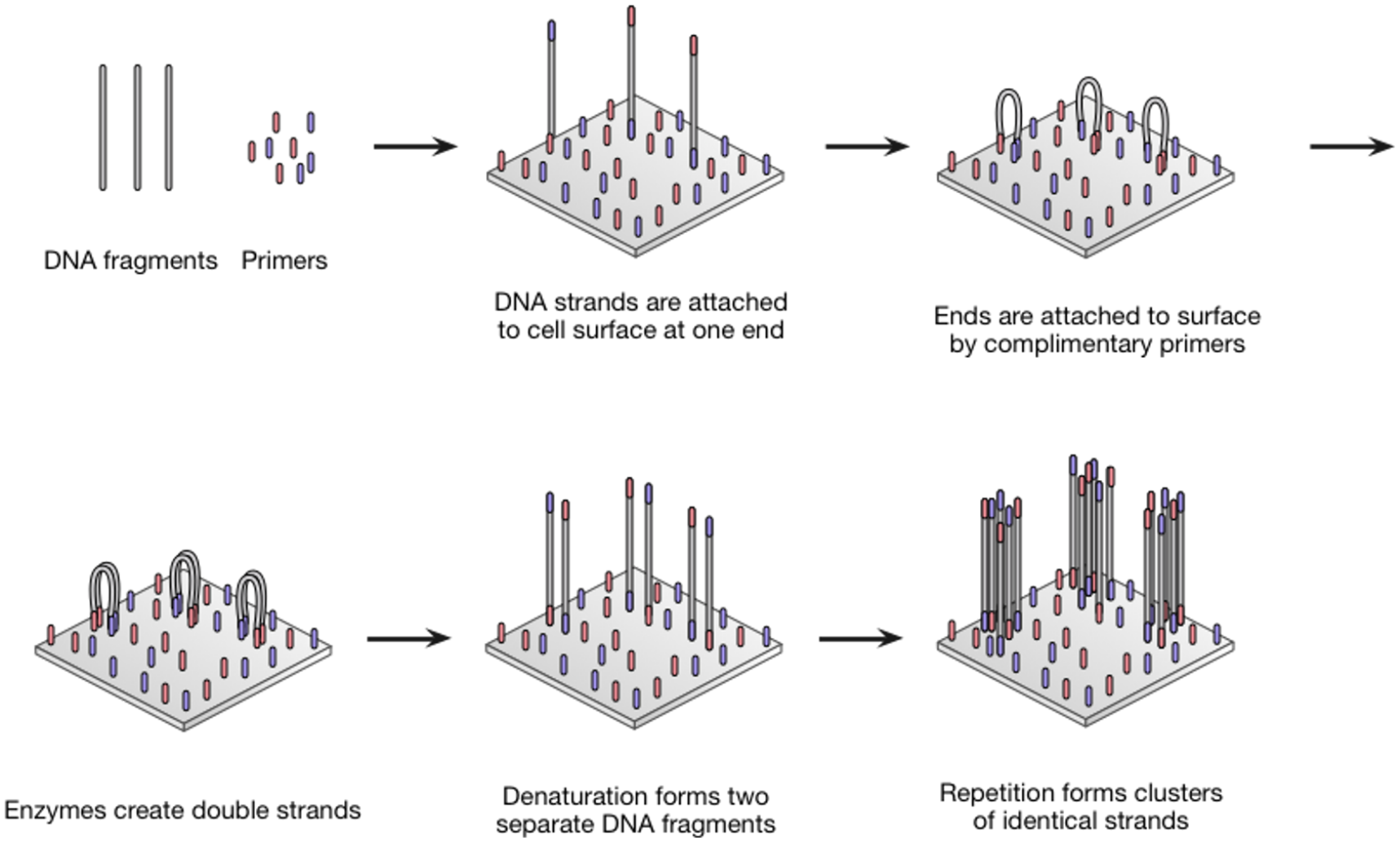
	Sequencing						Sequencing & Arrays	Arrays
Instrument	NovaSeq™ 6000 System	HiSeq X™ Ten* System	HiSeq™ 4000 System	MiSeq™ and MiSeqDx™ ‡ Systems	MiniSeq™ System	iSeq™ 100 System	NextSeq™ 550 and NextSeq™ 550Dx‡ Systems	iScan™ System

Flow cell

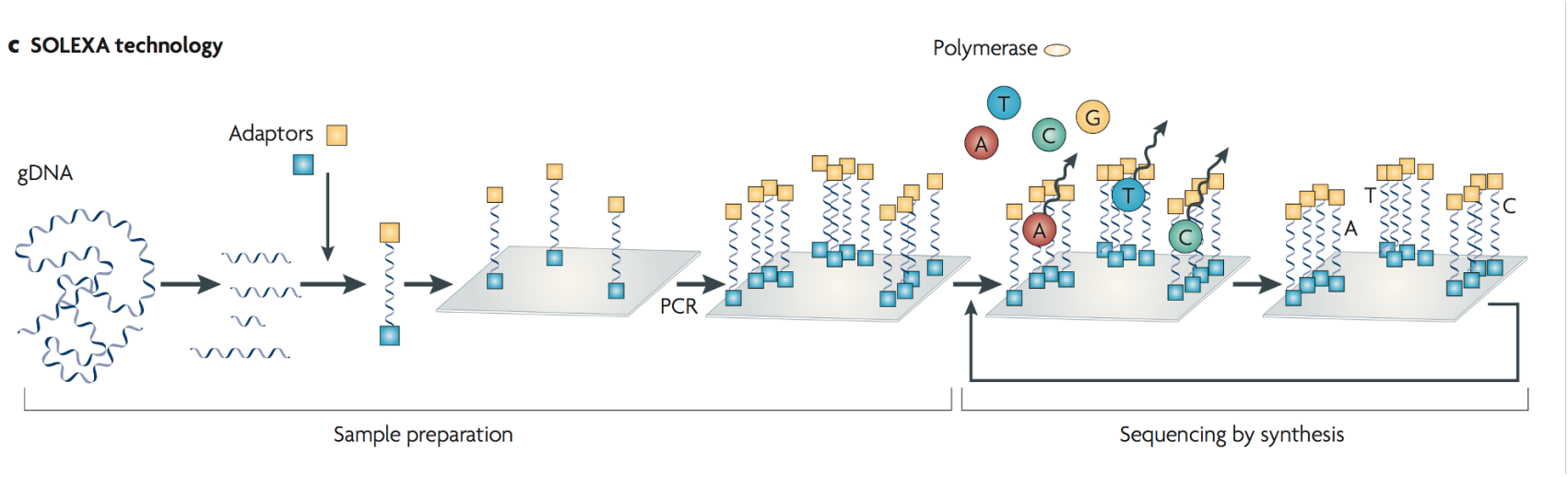

illumina
HL53CBCXX


A
flowcell ID
flow in
flow out
barcode
B
flow in
flow out
polyacrylamide-
coated interior
surface of flowcell

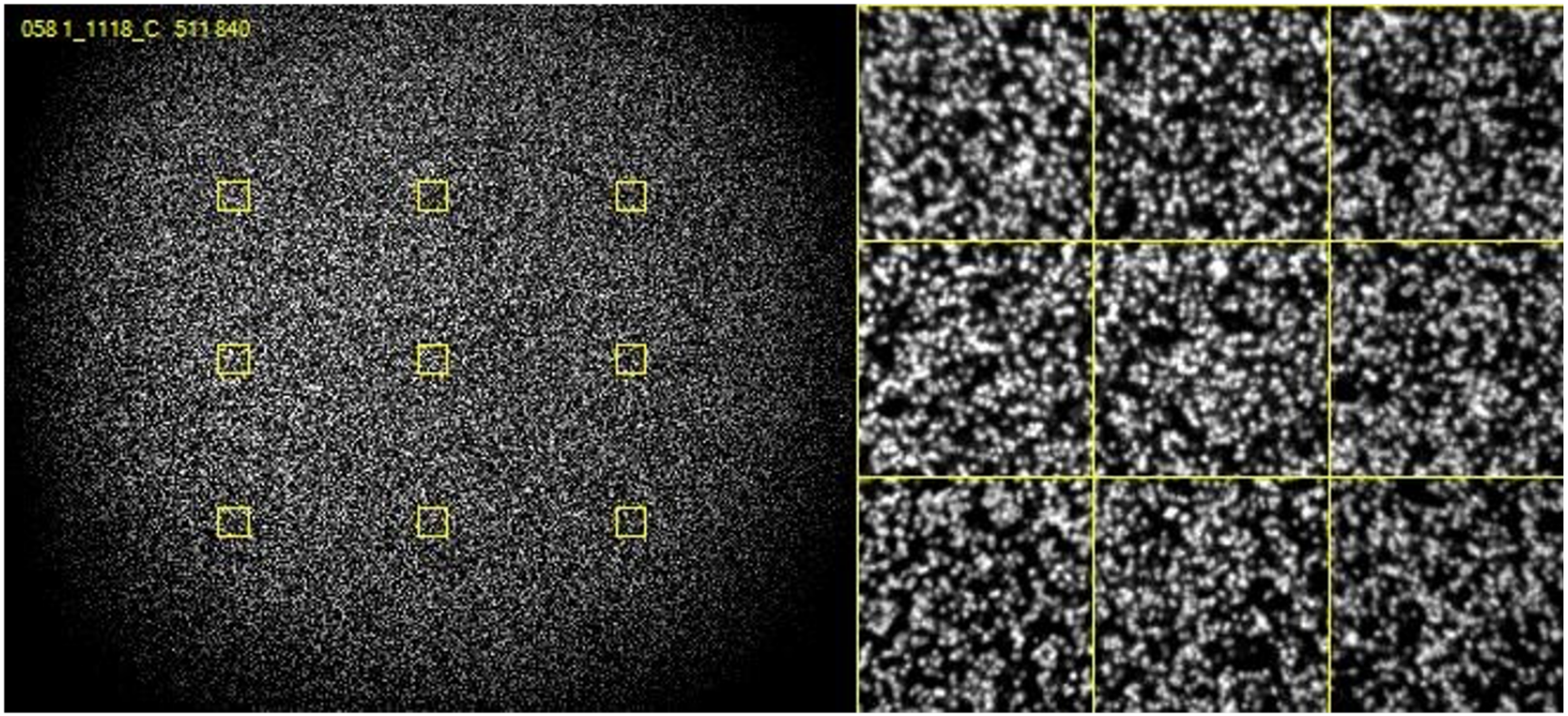
Illumina sequencing technology



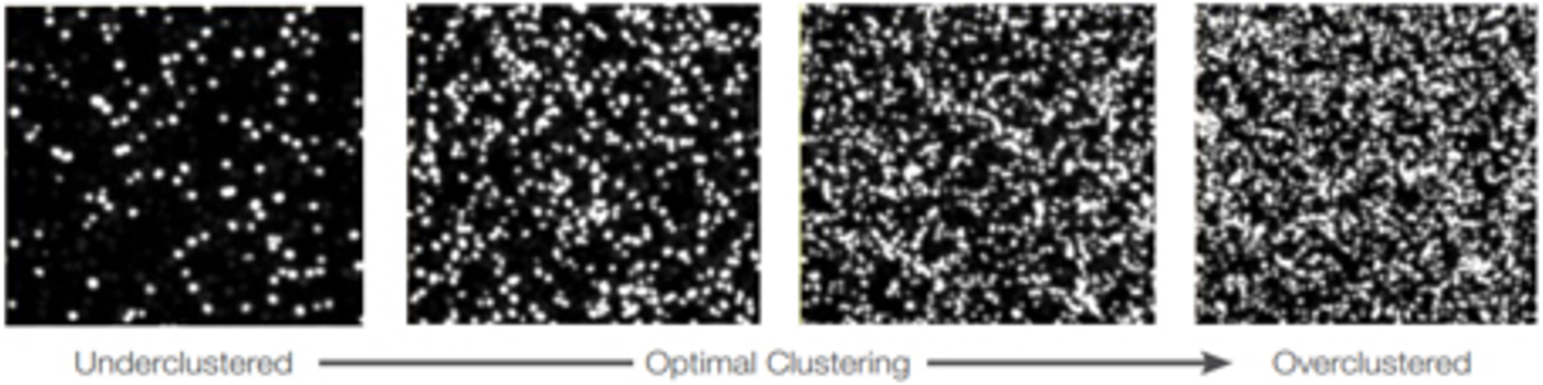
Illumina sequencing technology



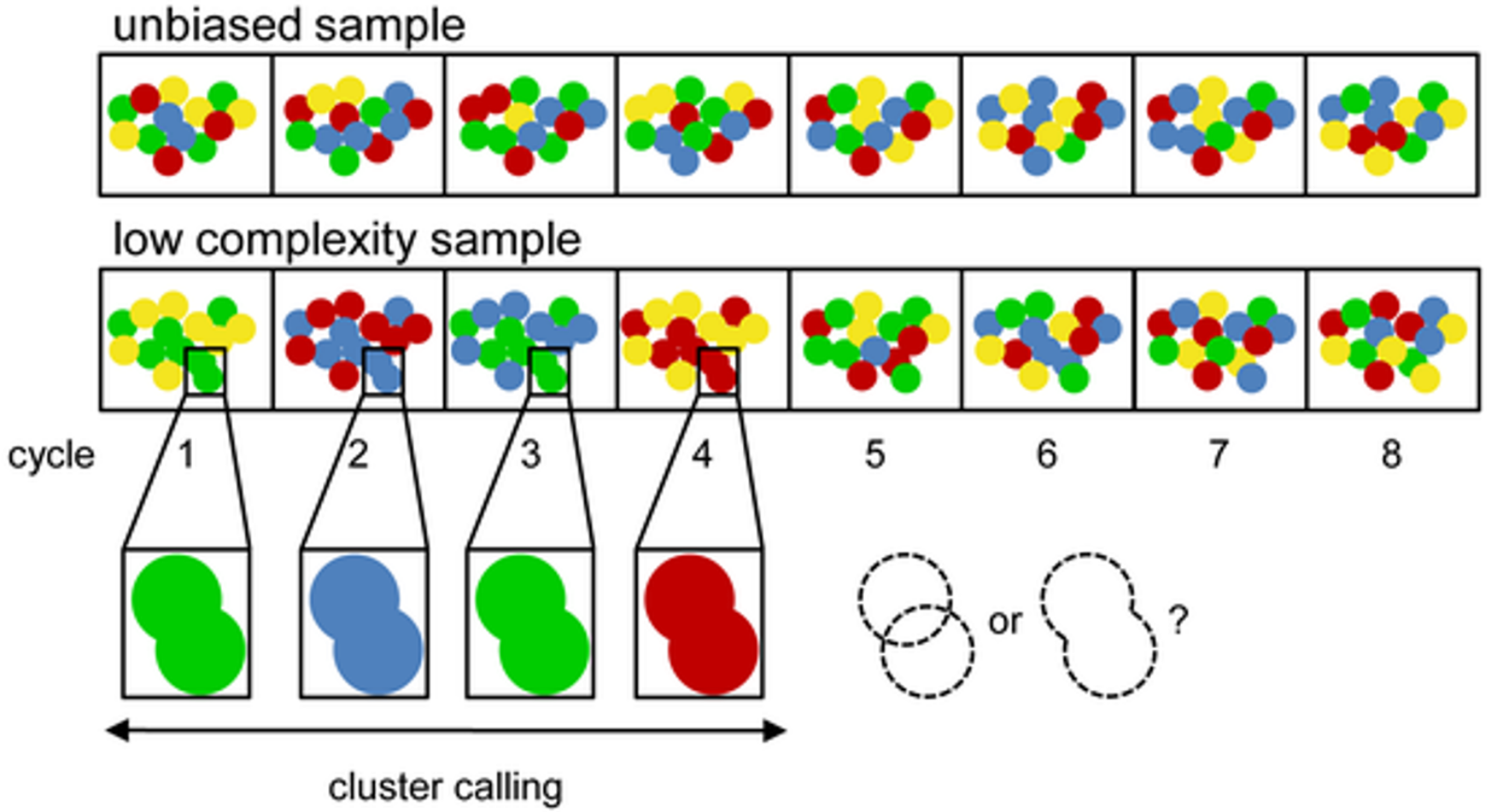
Illumina sequencing technology



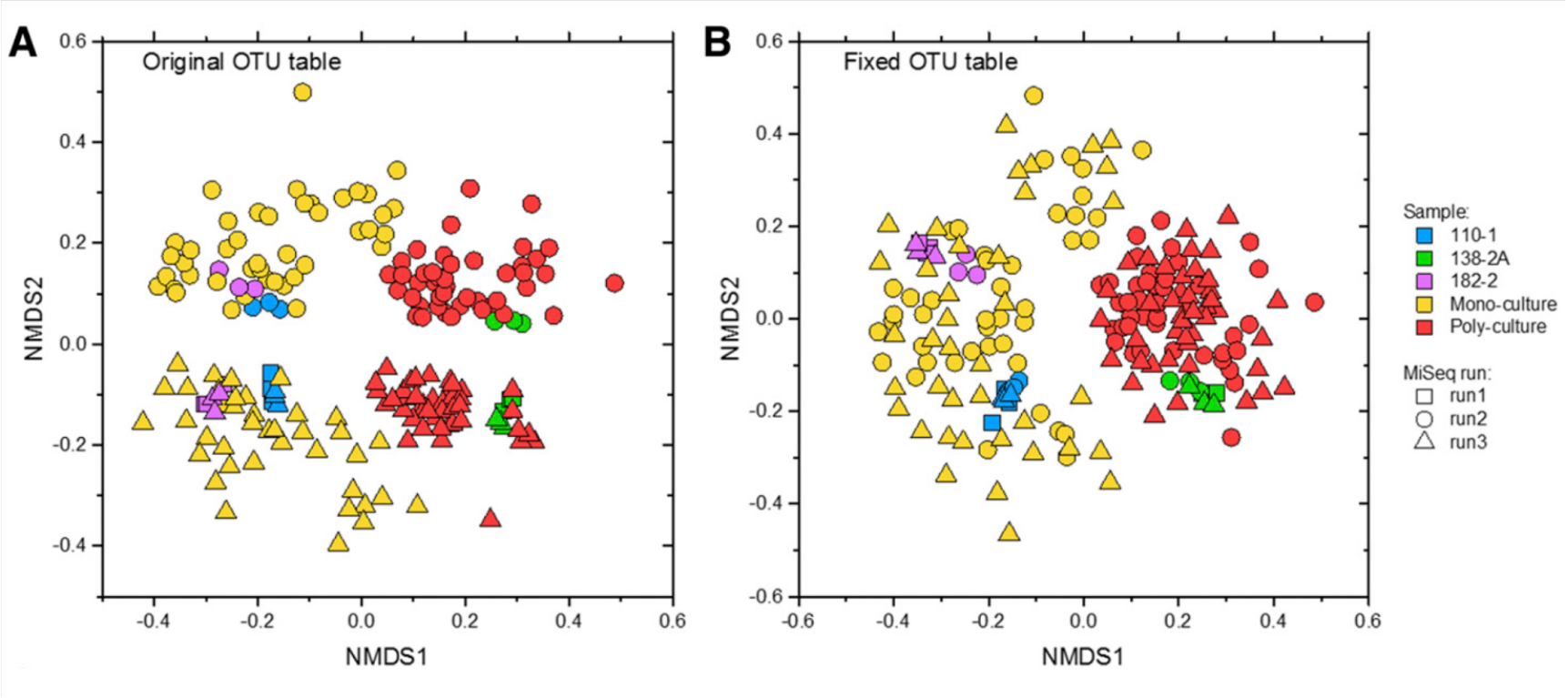
Illumina sequencing technology



Illumina sequencing technology



Technical considerations



Sequencing run


Sequencing Analysis Viewer
Run Folder: Q:\170214_K00150_0166_AHHJJHBBXX
Browse
Refresh
Analysis
Imaging
Summary
Indexing
Status
Extracted: 166
Called: 166
Scored: 166
Flow Cell Chart
Intensity
Surface:All
Cycle:1
Channel:a
Fix Scale
HHJJHBBXX Intensity
All Surfaces Cycle 1 a Channel
Data By Cycle
Intensity
Lane:All
Surface:All
Channel:All
Fix Scale
HHJJHBBXX All Lanes All Channels
R1
R:R:R4
Intensity
Cycle
QScore Distribution
Lane:All
Surface:All
Read:All
Cycle:All
Fix Scale
HHJJHBBXX All Lanes
>=Q30
481.0G
94.8%
Total (million)
Q Score
Data By Lane
Density
Surface:All
HHJJHBBXX Read 1
Density
PF
Cluster Density (K/mm2)
Lane
Qscore Heatmap
Lane:All
Surface:All
HHJJHBBXX All Lanes
% of max
Q Score
Cycle

www.hpc.oicr.on.ca/archive/h801/130328_SN801_0102_BD1W0DACXX/Data/Status_Files/Summary.htm

Total: 209 Extracted: 209 Called: 209 Scored: 209 Copied: 209

130328_SN801_0102_BD1W0DACXX

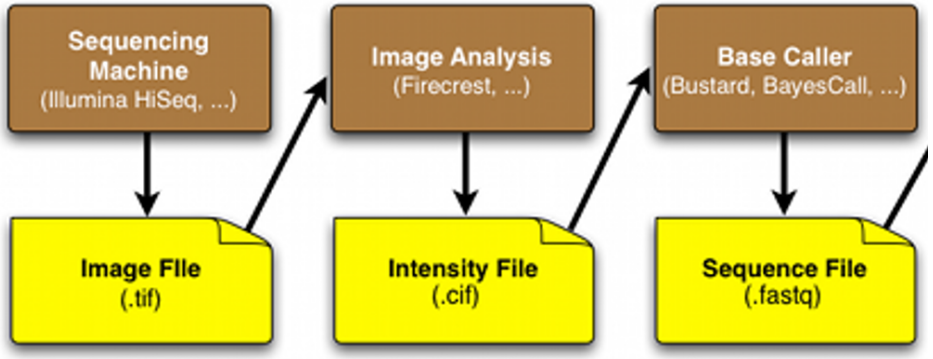
Run Info | Tile Status | Charts | Summary | Plots: Cluster Density | Data By Cycle

Read #1

Lane	Tiles	Clu.Dens. (#/mm^2)	% PF Clusters	Clusters PF (#/mm^2)	% Phas./Preph.	Cycles Err Rated	% Aligned	% Error Rate	% Error Rate 35 cycle	% Error Rate 75 cycle	% Error Rate 100 cycle	1st Cycle Int	% Intensity Cycle 20
1	96	651K +/- 73.2K	92.4 +/- 2.24	601.2K +/- 60.57K	0.135 / 0.220	100	0.63 +/- 0.021	0.22 +/- 0.043	0.10 +/- 0.018	0.16 +/- 0.025	0.22 +/- 0.043	2154 +/- 171.1	208.8 +/- 13.01
2	96	596K +/- 70.2K	92.1 +/- 1.72	548.5K +/- 59.51K	0.135 / 0.217	100	0.75 +/- 0.026	0.21 +/- 0.043	0.10 +/- 0.030	0.16 +/- 0.043	0.21 +/- 0.043	3322 +/- 469.7	140.5 +/- 21.52
3	96	817K +/- 89.5K	88.6 +/- 3.48	721.8K +/- 62.12K	0.131 / 0.214	100	0.41 +/- 0.029	0.25 +/- 0.040	0.11 +/- 0.044	0.19 +/- 0.037	0.25 +/- 0.040	5464 +/- 344.6	78.9 +/- 1.42
4	96	560K +/- 78.3K	94.2 +/- 1.74	526.5K +/- 67.52K	0.127 / 0.219	100	0.78 +/- 0.046	0.21 +/- 0.048	0.11 +/- 0.070	0.16 +/- 0.053	0.21 +/- 0.048	5686 +/- 350.0	79.5 +/- 1.45
5	96	459K +/- 68.2K	95.6 +/- 1.27	438.3K +/- 61.51K	0.137 / 0.228	100	1.00 +/- 0.054	0.19 +/- 0.034	0.09 +/- 0.026	0.14 +/- 0.023	0.19 +/- 0.034	5867 +/- 342.8	79.0 +/- 1.38
6	96	492K +/- 72.4K	95.3 +/- 1.52	468.1K +/- 64.24K	0.138 / 0.225	100	0.97 +/- 0.050	0.20 +/- 0.045	0.10 +/- 0.028	0.15 +/- 0.027	0.20 +/- 0.045	5795 +/- 331.7	79.1 +/- 1.22
7	96	754K +/- 92.8K	90.2 +/- 3.11	678.4K +/- 69.17K	0.142 / 0.215	100	0.47 +/- 0.032	0.24 +/- 0.072	0.11 +/- 0.029	0.19 +/- 0.057	0.24 +/- 0.072	5622 +/- 350.6	76.4 +/- 1.43
8	96	657K +/- 86.0K	92.3 +/- 2.44	605.2K +/- 69.96K	0.145 / 0.219	100	0.60 +/- 0.027	0.22 +/- 0.048	0.10 +/- 0.026	0.17 +/- 0.038	0.22 +/- 0.048	5773 +/- 309.3	77.2 +/- 1.29

Primary data analysis

- Primary analysis
  - From image to base calling
  - Cluster detection (4th cycle)
  - Cluster intensity correction
  - Base calling
  - Clusters are filtered (CPF, 25th cycle)
  - Q scores are assigned to each base
- CASAVA
  - Demultiplex
  - Create fastq files



Primary data analysis

- Primary analysis
  - From image to base calling
  - Cluster detection (4th cycle)
  - Cluster intensity correction
  - Base calling
  - Clusters are filtered (CPF, 25th cycle)
  - Q scores are assigned to each base
- CASAVA
  - Demultiplex
  - Create fastq files

Secondary data processing

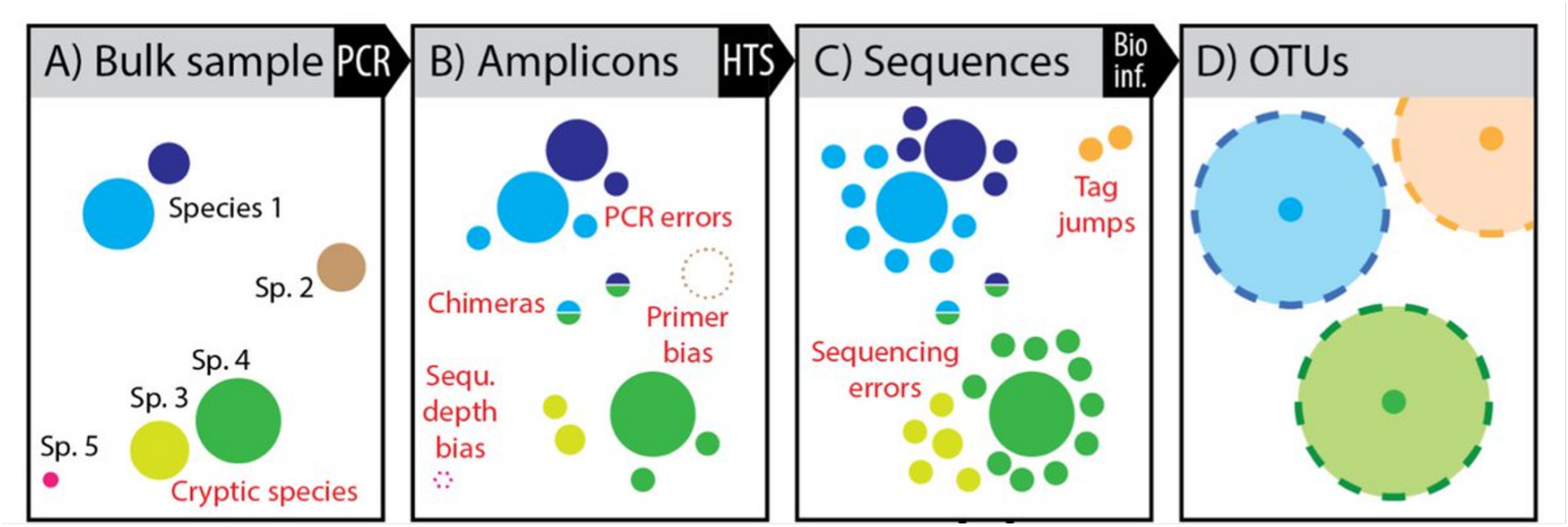
Workflow

- De-multiplexing
- Reads pre-processing
- Dereplication
- Clustering of variants
- Filtering of artefacts
- Alignment to references
- Taxonomy annotation
- Downstream analysis

First choice!

- OTUs (Operational Taxonomic Unit)
- AVSs (Amplicon Sequence Variant)

Remember: bias is everywhere!



Which is our goal?

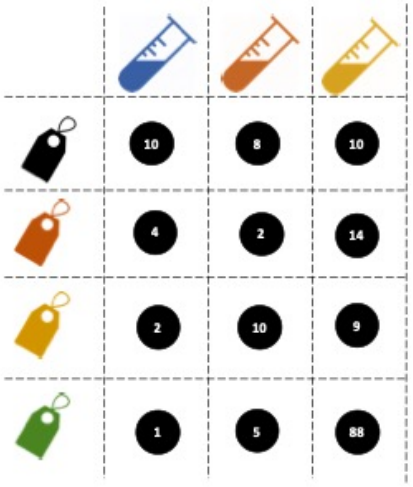
Metadata

1	sampleid	Treatment	Years
2	AM-16S-1	maize-mono	1999
3	AM-16S-2	maize-mono	1999
4	AM-16S-3	maize-mono	1999
5	AM-16S-4	push-pull	1999
6	AM-16S-5	push-pull	1999

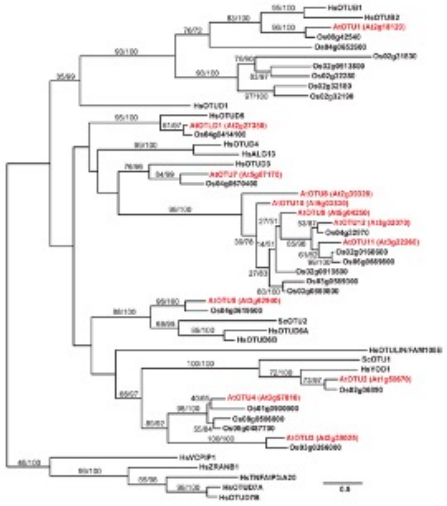
Taxonomy

ASV148428	Bacteria(100);Proteobacteria(100);Gammaproteobacte...
ASV212114	Bacteria(100);Cyanobacteria(100);Cyanobacteria(100);...
ASV9620	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...
ASV147186	Bacteria(100);Proteobacteria(100);Betaproteobacteria(...
ASV89359	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...
ASV1061	Bacteria(100);Proteobacteria(100);Gammaproteobacte...
ASV328581	Bacteria(100);Bacteroidetes(100);Bacteroidia(100);Bac...
ASV86104	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...

OTU table

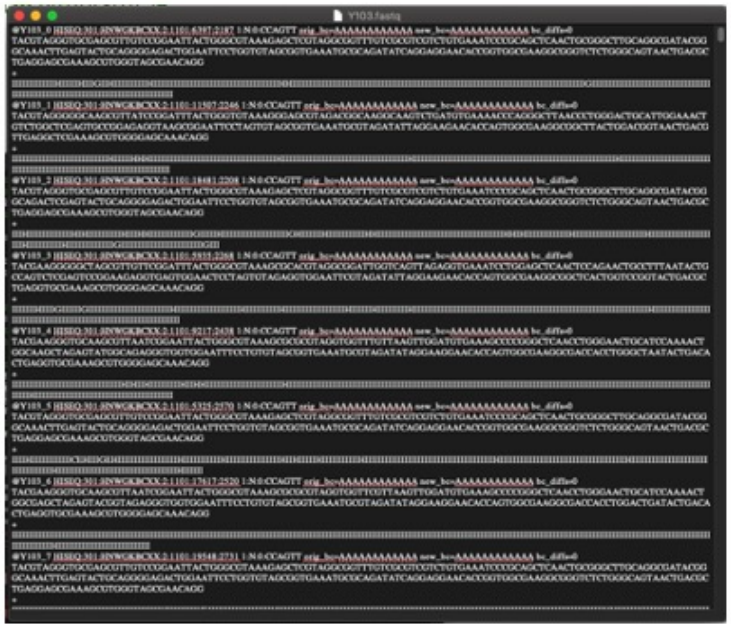


Phylogenetic tree



What do we have?

Raw data



1

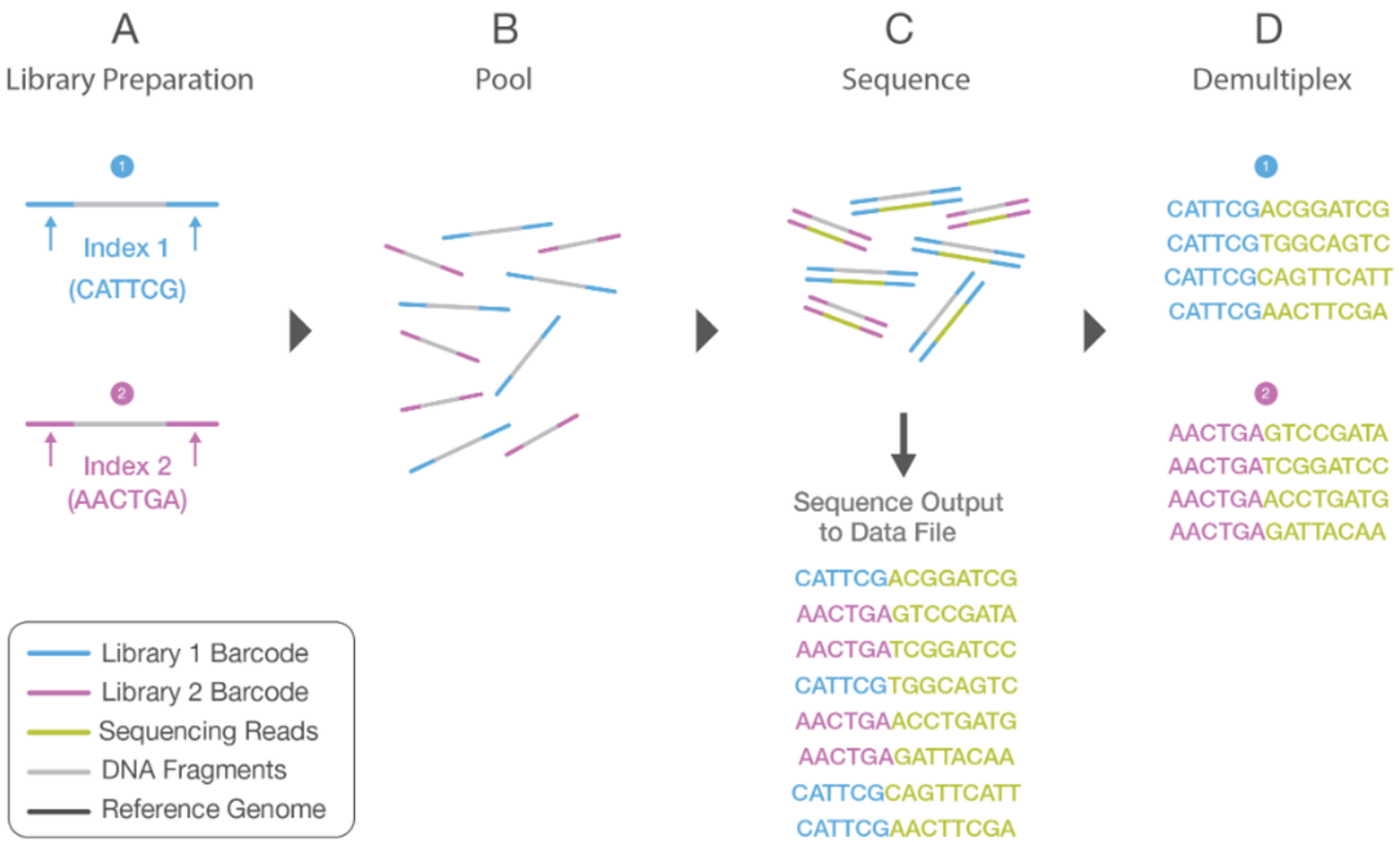
Metadata

1	sampleid	Treatment	Years
2	AM-16S-1	maize-mono	1999
3	AM-16S-2	maize-mono	1999
4	AM-16S-3	maize-mono	1999
5	AM-16S-4	push-pull	1999
6	AM-16S-5	push-pull	1999

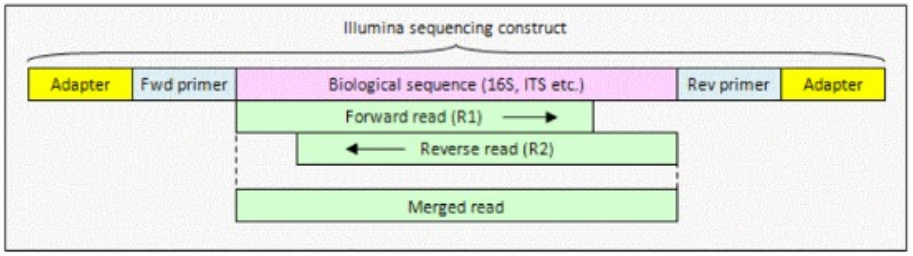
*.fastq files

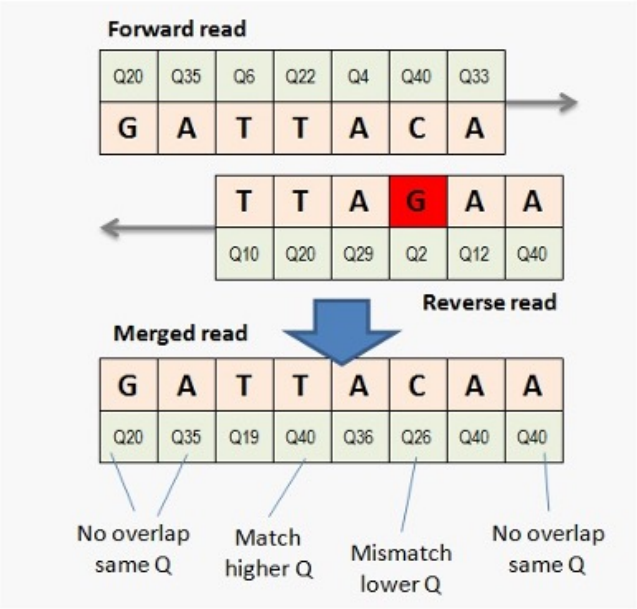
Identifier	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	+
Quality scores	hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	+
Quality scores	hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

Demultiplexing



Merge PE reads





Remove suspicious reads

```
>GQY1XT001A6MUA
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATACAGTTTCCAATG
>GQY1XT001BTRWS
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCACAGTTTCCAAAGCAGTTCCGGGGTTGGG
>GQY1XT001AK4J0
TCTAGCCGCACAGTTTCAAAAGCACTCCCAGGGTT
>GQY1XT001BBPBR
AATGGTACCCGTCAATTCATTTGACGTTGCCCCCCGTTTACTGTGCGGACTACCAGTCGCACTCAAGGCCCCCAGTTTCAACGG
>GQY1XT001BDDE9
AATGGTACCCGTCAATTCCTTTAATCTTGCGGGTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTTACACAGTTTCCAGAG
>GQY1XT001CIUF3
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCTTTACGGCGTGGACTACCAGGCGCCCTCCAGCCCGGCAGTTTCCAGTGCAGTCCCGGGGTT
>GQY1XT001BKRP5
AATGGTACCCGTCAATTCATTTAATCTCTCCCCCTTTCCCCCCCCCCCCCTTTCCCCCCCCCCCCCTTTCCCCCCCCCC
>GQY1XT001B44ZE
AATGGTACCCGTCAATTCATTTAACCTTGCGGGGTTTTACCGCGTGGACTACCAGGCGCCCTCAAGAAGAACAGTTTTGAACGCAGCTATGGGTT
>GQY1XT001CIW3P
AATGGTACCCGTCAATTCATTTGACGTTGCCTCTCGTTTACTGCGTGGACTACCAGTCGCACTCAAGGCCCCCA
>GQY1XT001A731D
AATGGTACCCGTCAATTCATTTAACGTTGCCCCCGTTACTGCGTGGACTACCAGGGGCAATCAAGACTGCCA
```

Trimming same length

```
>GQY1XT001A6MUA
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATACAGTTTCCAATG
>GQY1XT001BTRWS
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCACAGTTTCCAAAGCAGTTCCGGGGTTGGG
>GQY1XT001AK4J0
TCTAGCCGCACAGTTTCAAAAGCACTCCCAGGGTT
>GQY1XT001BBPBR
AATGGTACCCGTCAATTCATTTGACGTTGCCCCCCGTTTACTGTGCGGACTACCAGTCGCACTCAAGGCCCCCAGTTTCAACGG
>GQY1XT001BDDE9
AATGGTACCCGTCAATTCCTTTAATCTTGCGGGTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTTACACAGTTTCCAGAG
>GQY1XT001CIUF3
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCTTTACGGCGTGGACTACCAGGCGCCCTCCAGCCCGGCAGTTTCCAGTGCAGTCCCGGGGTT
>GQY1XT001BKRP5
AATGGTACCCGTCAATTCATTTAATCTCTCCCCCTTTCCCCCCCCCCCCCTTTCCCCCCCCCCCCCTTTCCCCCCCCCC
>GQY1XT001B44ZE
AATGGTACCCGTCAATTCATTTAACCTTGCGGGGTTTTACCGCGTGGACTACCAGGCGCCCTCAAGAAGAACAGTTTTGAACGCAGCTATGGGTT
>GQY1XT001CIW3P
AATGGTACCCGTCAATTCATTTGACGTTGCCTCTCGTTTACTGCGTGGACTACCAGTCGCACTCAAGGCCCCCA
>GQY1XT001A731D
AATGGTACCCGTCAATTCATTTAACGTTGCCCCCGTTACTGCGTGGACTACCAGGGGCAATCAAGACTGCCA
```

Dereplication

```
>GQY1XT001A6MUA
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001BTRWS
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
>GQY1XT001BBPBR
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001BDDE9
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001CIUF3
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
>GQY1XT001B44ZE
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001CIW3P
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
>GQY1XT001A731D
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
```

Dereplication

```
>GQY1XT001A6MUA  DEPTH = 5
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001BTRWS  DEPTH = 3
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
```

```
>GQY1XT001BBPBR
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001BDDE9
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001CIUF3
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
>GQY1XT001B44ZE
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001CIW3P
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
>GQY1XT001A731D
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
```

Cluster variants

```
>*S16-0000006
TACGTTTATCGCGTT-AGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGTTTAGGGTGTGGACTAA
>#S16-0000046
TACGTTTATCGCGTTTAGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGTTTAGGGTGTGGACTAA
>#S16-0000241
TACGTTTATCGCGTT-AGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGT-TAGGGTGTGGACTAA
>#S16-0000375
TACGTTTATCGCATT-AGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGTTTAGG-TGTGGACTAA
>*S16-0000001
GGCACTTAAAGCGTTAGCTACGGCGCAGAAACCACGGGTGG-CCCCCACACCTAGTGCCCAACGTTTACAGCGTGGT
>#S16-0000209
GGCACTTAAAGCGTTAGCTACGGCGCAGAAACCACGGGTGGTCCCCCACACCTAGTGCCCAACGTTTACAGCGTGGG
>#S16-0000667
GGCACTTAAAGCGTTAGCTACGGCGCAGAAACCACGGGTGG-CCCCCACACCTAGTGC-CAACGTTTACAGCGTGGT
>*S16-0000004
TCGACTTAACGCGTTAGCTCCGGAAGCCACGCCTCAAGG-GCACAACCTCCAAGTCGACATCGTTTACGGCGTGGAT
>#S16-0000625
TCGACTTAACGCGTTAGCTCCGGAAGCCACGCCTCAAGG-GCACAACCTCCAAGTCGACATCGT-TACGGCGTGGAT
>#S16-0000673
TCGACTTAACGCGTTAGCTCCGGAAGCCACGCCTCAAGGGGCACAACCTCCAAGTCGACATCGTTTACGGCGTGGAT
```

Cluster variants

```
>*S16-0000006  DEPTH + 3
TACGTTTATCGCGTT-AGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGTTTAGGGTGTGGACTAA
>*S16-0000001  DEPTH + 2
GGCACTTAAAGCGTTAGCTACGGCGCAGAAACCACGGGTGG-CCCCCACACCTAGTGCCCAACGTTTACAGCGTGGT
>*S16-0000004  DEPTH + 2
TCGACTTAACGCGTTAGCTCCGGAAGCCACGCCTCAAGG-GCACAACCTCCAAGTCGACATCGTTTACGGCGTGGAT
```

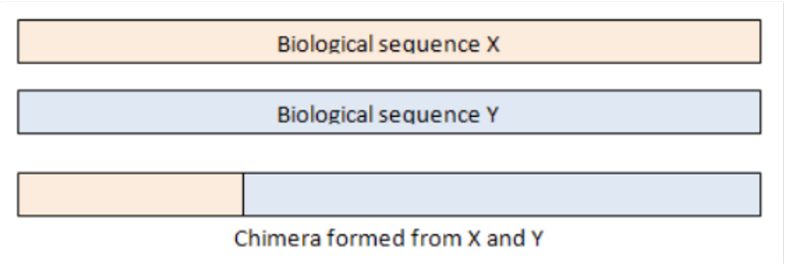
```
>#S16-0000046
TACGTTTATCGCGTTTAGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGTTTAGGGTGTGGACTAA
>#S16-0000241
TACGTTTATCGCGTT-AGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGT-TAGGGTGTGGACTAA
>#S16-0000375
TACGTTTATCGCATT-AGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCGTTTAGG-TGTGGACTAA
>#S16-0000209
GGCACTTAAAGCGTTAGCTACGGCGCAGAAACCACGGGTGGTCCCCCACACCTAGTGCCCAACGTTTACAGCGTGGG
>#S16-0000667
GGCACTTAAAGCGTTAGCTACGGCGCAGAAACCACGGGTGG-CCCCCACACCTAGTGC-CAACGTTTACAGCGTGGT
>#S16-0000625
TCGACTTAACGCGTTAGCTCCGGAAGCCACGCCTCAAGG-GCACAACCTCCAAGTCGACATCGT-TACGGCGTGGAT
>#S16-0000673
TCGACTTAACGCGTTAGCTCCGGAAGCCACGCCTCAAGGGGCACAACCTCCAAGTCGACATCGTTTACGGCGTGGAT
```

Filtering artefacts

- PCR errors
  - Most *Taq* polymerases introduce point mutations (error) at a rate of 1 every 1000 bases
  - **Solution**: use Hi-Fi polymerases with lower error rates ($$$)
- Chimeras
  - Chimeras are sequences formed by two or more biological sequences joined together
  - **Solution**: reduce number of PCR cycles and increase annealing temperature

Filtering artefacts

- PCR errors
  - Most *Taq* polymerases introduce point mutations (error) at a rate of 1 every 1000 bases
  - **Solution**: use Hi-Fi polymerases with lower error rates ($$$)
- Chimeras
  - Chimeras are sequences formed by two or more biological sequences joined together
  - **Solution**: reduce number of PCR cycles and increase annealing temperature



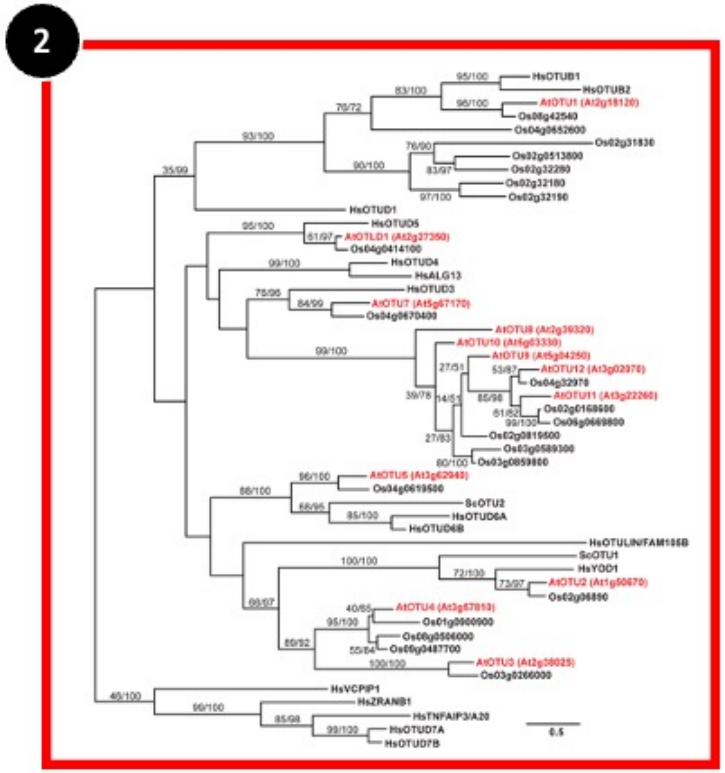
Filtering artefacts

```
>*16S-0000011 | depth=44 | freq=2.42
TTCAGTCGCTCCCCTAGCTTTCGCACTTCAGCGTCAGTTGCCGTCCAGTGAACTATCTTCATCATCGGCATT
CCTGCACATATCTACGAATTTCACCTCTACTCGTGCAGTTCCGTCCACCTCTCCAGCACTCTAGCCAAACAG
>*16S-0000076 | depth=33 | freq=1.82
TTCAATGTTTGCTCCCCACGCTTTCGAGCCTCAGCGTCAGTTACAAGCCAGAGAGCCGCTTTCGCCACCGGT
GTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTCTTGCACTCAAGTTAAA
>*16S-0000052 | depth=32 | freq=1.76
TTCACGATACCCGCACCTTCGAGCTTAAGCGTCAGTGGCGCTCCCGTCAGCTGCCTTCGCAATCGGAGTTCT
TCGTCATATCTAAGCATTTCACCGCTACACGACGAATTCCGCCAACGTTGTGCGTACTCAAGGAAACCAGTA
>*16S-0000141 | depth=15 | freq=0.83
TTCAACGTTCGCTCCCCTGGCTTTCGCGCCTCAGCGTCAGTTTTCGTCCAGAAAGTCGCCTTCGCCACTGGT
GTTCTTCCTAATATCTACGCATTTCACCGCTACACTAGGAATTCCACTTTCCTCTCCGATACTC
>#16S-0000058 | depth=12 | freq=0.66
TTCAGTCGCTCCCCTAGCTTTCGCACTTCAGCGTCAGTTGCCGTAAGCCAGAGAGCCGCTTTCGCCACCGGT
GTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTCTTGCACTCAAGTTAAA
>*16S-0000098 | depth=10 | freq=0.55
TTTAGTCCTGTTCGCTCCCCACGCTTTCGCTCCTCAGCGTCAGTAACGGCCCAGAGACCCGCCTTCGCCACC
GGTGTTCTTCCTGATATCTGCGCATTCCACCGCTACACCAGGAGTTCCAGCCTCCC
>#16S-0000295 | depth=2 | freq=0.11
TTCACGATACCCACGCTTTCGAGCATCAGCGTCAGTTGCGCTACAGTAAGCTGCCTTCGCAATCGGAGTTCT
TCGTGATATCTAAGCATTTCACCGCTACACCACGAATTCCGCCTACTTTCGGCGCACTCAAGCCCCCCAGTT
>#16S-0000021 | depth=1 | freq=0.06
TTCAACGTTCGCTCCCCTGGCTTTCGCGCCTCAGCGTCAGTTTTCGTCCAGAAAGTCGCCTTCGCCACTGGT
```

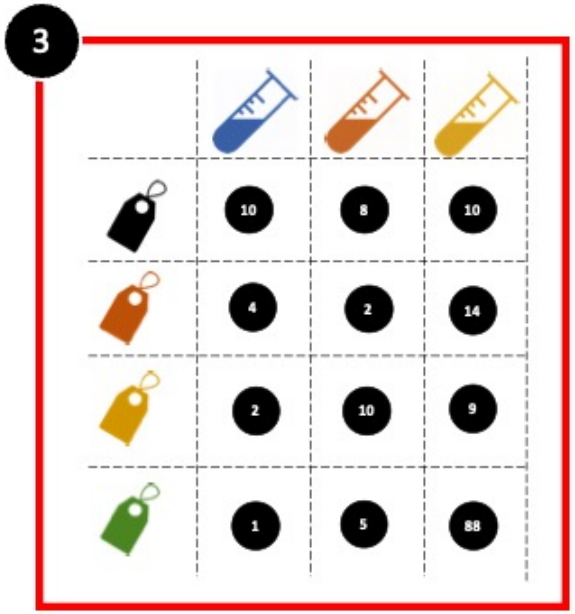
Chimera

Contaminations

Phylogenetic tree!



OTU table!



Align to reference

```
>*16S-0000002 | depth=42 | freq=2.31
TTCAACCTTGCGGTCGTACTCCCCAGGCGGAGTGCTTAATGCGTTAGCTGCGGCACTAAACCCCGGAAAGGGTCTAACACCTAGCACTCATCGTT
TACGGCGTGGACTACCAGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGAGCCTCAGCGTCAGTTACAAGCCAGAGAGCCGCTTTCGCCACCG
GTGTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTCTTGCACTCAAGTTAAACAGTTTCCAAAGCGTACTATG
GTTAAGCCACAGCCTTTAACTTCAGACTTATCT
>*16S-0000019 | depth=12 | freq=0.66
TTCAGCCTTGCGGCCGTACTCCCCAGGCGGATTACTTATCGCATTCGCTTCGGCACAGACAGTCTTCCTGCCCACACCCAGTAATCATCGTTTAC
GGCCGGGACTACCAGGGTATCTAATCCTGTTCGCTCCCCCGGCTTTCGCACTTCAGCGTCAGTTACCGTCCAGTGAACTATCTTCATCATCGGCA
TTCCTGCACATATCTACGAATTTCACCTCTACTCGTGCAGTTCCGTCCACCTCTCCGGTACTCCAGCCTATCAGTTTCAAAGGCAGGCCTGCGGT
TGAGCCGCAGGTTTTCACCCCTGACTTGAAAGG
```

VS.

green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

Nucleotide BLAST
nucleotide ▸ nucleotide

Assign taxonomy

AY053482.1;tax=k:Bacteria,p:Firmicutes,c:Bacilli,o:Lactobacillales,f:Streptococcaceae,
g:Streptococcus,s:pseudopneumoniae

Sequence ID: lcl|Query_210570 **Length:** 1429 **Number of Matches:** 1

Range 1: 565 to 882 Graphics

Score	Expect	Identities	Gaps	Strand
588 bits(318)	7e-172	318/318(100%)	0/318(0%)	Plus/Minus

Taxonomy!

4

ASV148428	Bacteria(100);Proteobacteria(100);Gammaproteobacte...
ASV212114	Bacteria(100);Cyanobacteria(100);Cyanobacteria(100);...
ASV9620	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...
ASV147186	Bacteria(100);Proteobacteria(100);Betaproteobacteria(...
ASV89359	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...
ASV1061	Bacteria(100);Proteobacteria(100);Gammaproteobacte...
ASV328581	Bacteria(100);Bacteroidetes(100);Bacteroidia(100);Bac...
ASV86104	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...

Data analysis

What we will use?

What do we need to start?

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 43879 taxa and 289 samples ]
sample_data() Sample Data:       [ 289 samples by 9 sample variables ]
tax_table()   Taxonomy Table:    [ 43879 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 43879 tips and 43878 internal nodes ]
```

otu_table()

	16S.SOI.41	16S.SOI.48	16S.SOI.18	16S.SOI.46	16S.SOI.59	16S.SOI.34	16S.ROO.57	16S.ROO.43	16S.ROO.58
denovo7709	1	1	0	0	0	0	0	0	0
denovo7708	0	0	1	1	1	0	0	0	0
denovo22216	0	0	0	0	0	1	0	0	0

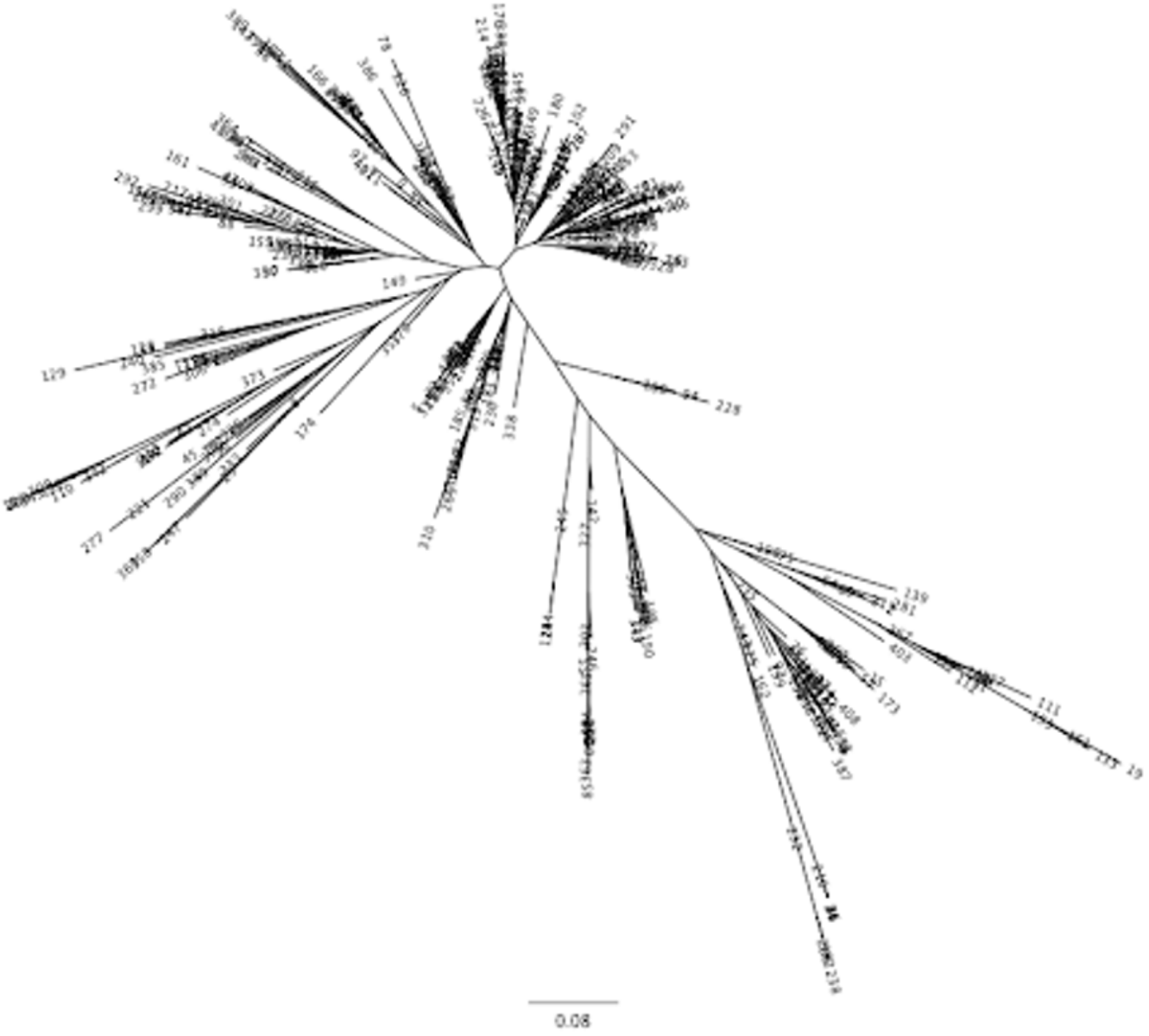
sample_data()

	A	B	C	D	E	F	G	H
1	SampleID	Community	Category	Sample_type	Genotype	Soil	Aphid	H_defensa
2	16S.APH.1	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Absent
3	16S.APH.2	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Absent
4	16S.APH.3	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Absent
5	16S.APH.4	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Absent
6	16S.APH.5	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Absent
7	16S.APH.6	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Present
8	16S.APH.7	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Present
9	16S.APH.8	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Present
10	16S.APH.9	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Present
11	16S.APH.10	Bacterial	Experimenta	Aphid	TBR	WHS	Present	Present
12	16S.APH.11	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Absent
13	16S.APH.12	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Absent
14	16S.APH.13	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Absent
15	16S.APH.14	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Absent
16	16S.APH.15	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Absent
17	16S.APH.16	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Present
18	16S.APH.17	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Present
19	16S.APH.18	Bacterial	Experimenta	Aphid	TBR	MICROB	Present	Present

tax_table()

```
            Rank1            Rank2
denovo7709  "D_0__Bacteria" "D_1__Proteobacteria"
denovo7708  "D_0__Bacteria" "D_1__Proteobacteria"
denovo22216 "D_0__Bacteria" "D_1__Bacteroidetes"
denovo11322 "D_0__Bacteria" "D_1__Bacteroidetes"
denovo44859 "D_0__Bacteria" "D_1__Chloroflexi"
```

phy_tree()



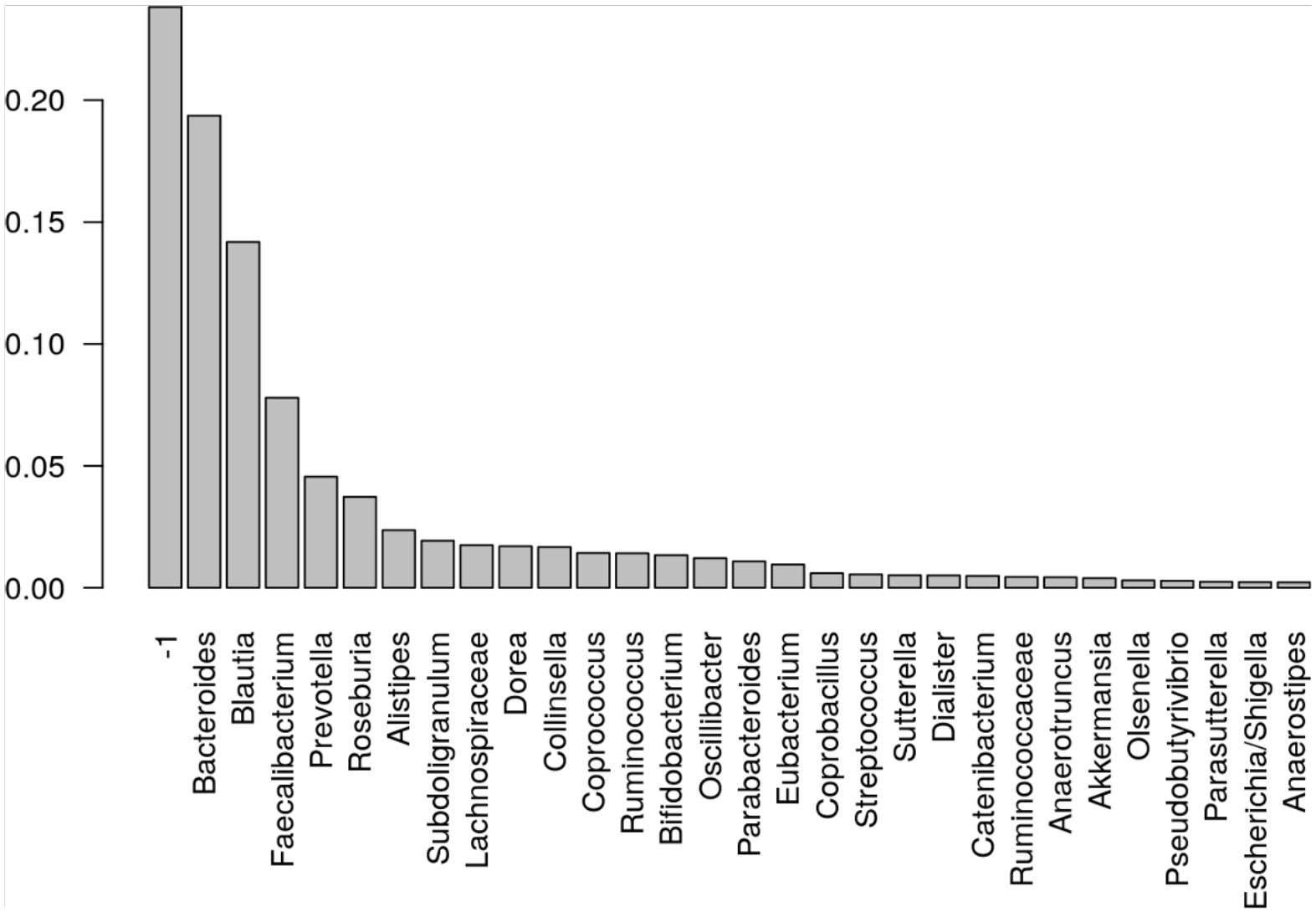
Terminology

- Quantitative versus qualitative metrics
  - qualitative metrics only account for whether an organism is present or absent
  - quantitative metrics account for abundance
- Phylogenetic versus non-phylogenetic metrics
  - non-phylogenetic metrics treat all OTUs as being equally related
  - phylogenetic metrics incorporate evolutionary relationships between the OTUs

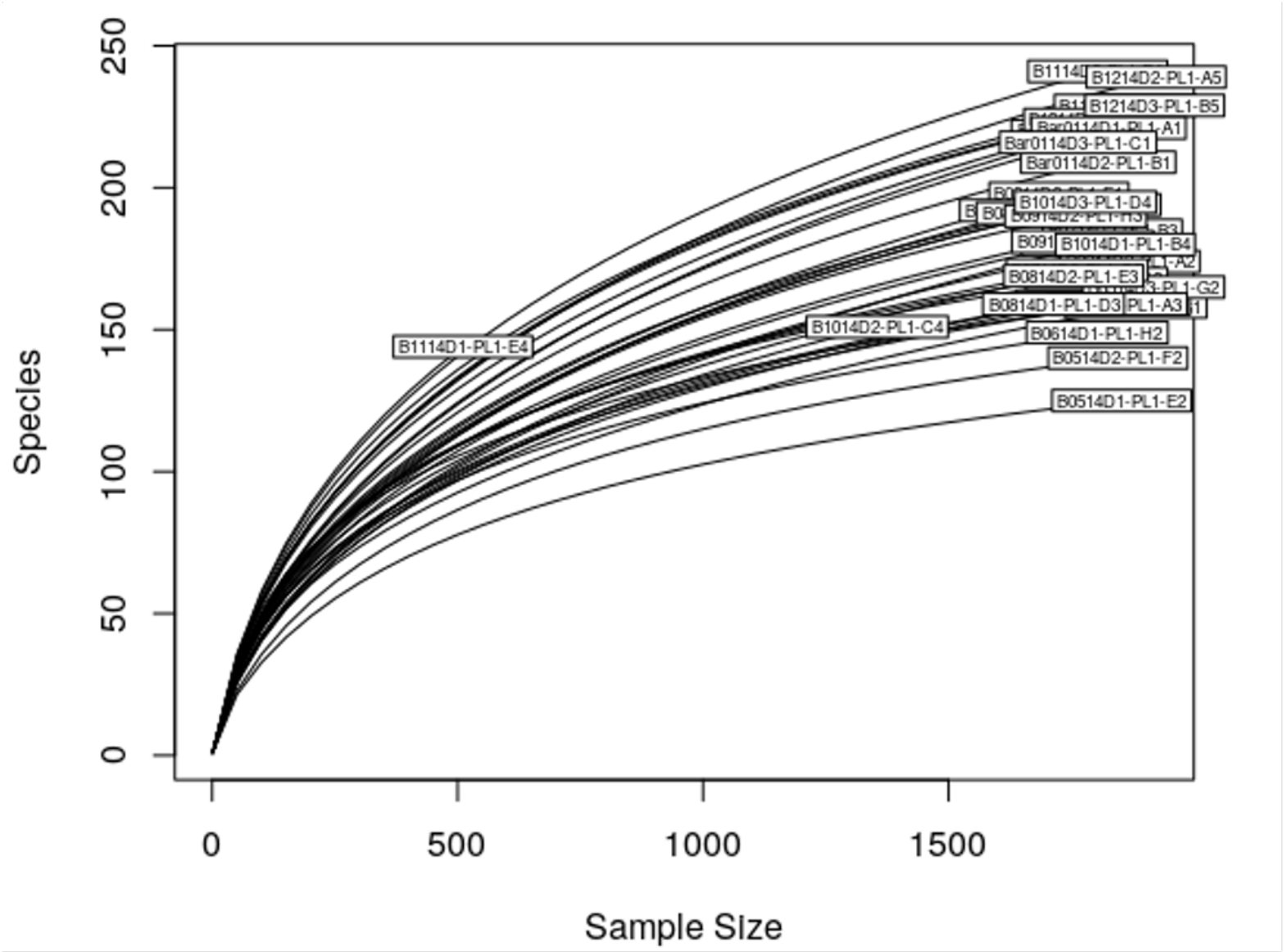
Terminology

- Quantitative versus qualitative metrics
  - qualitative metrics only account for whether an organism is present or absent
  - quantitative metrics account for abundance
- Phylogenetic versus non-phylogenetic metrics
  - non-phylogenetic metrics treat all OTUs as being equally related
  - phylogenetic metrics incorporate evolutionary relationships between the OTUs

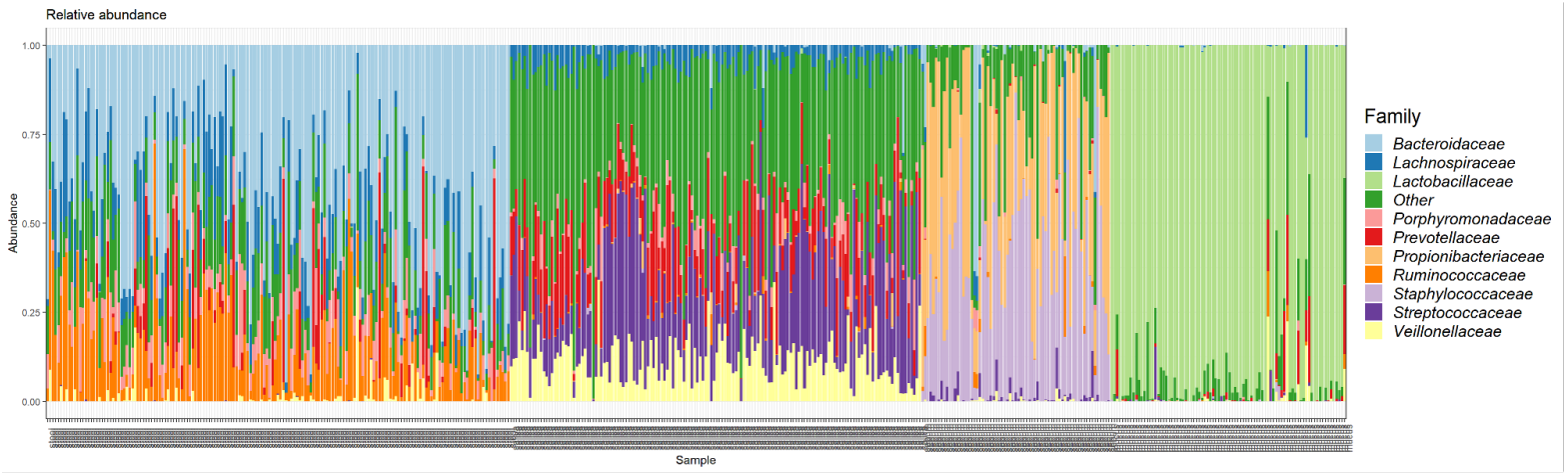
Exploratory plots



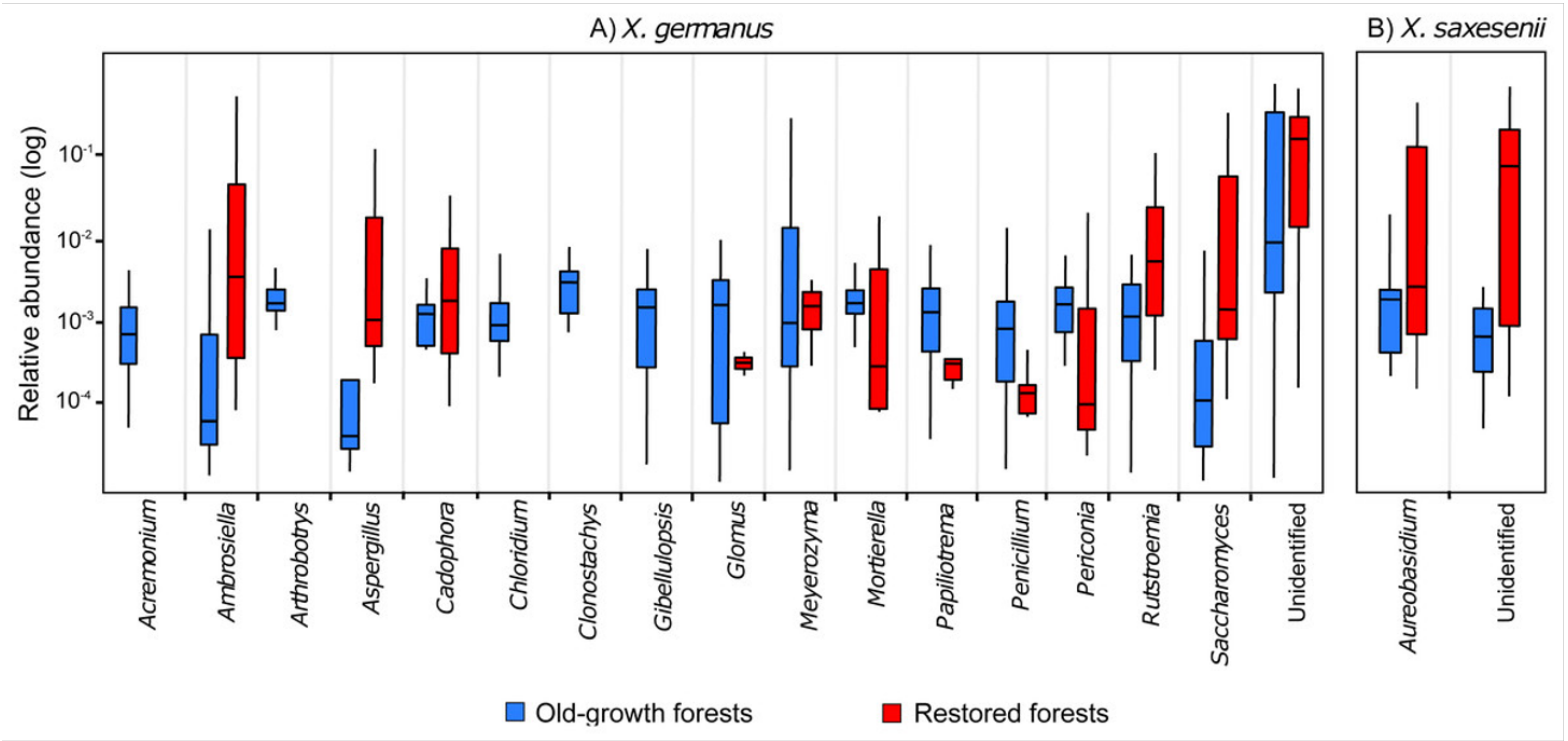
Exploratory plots



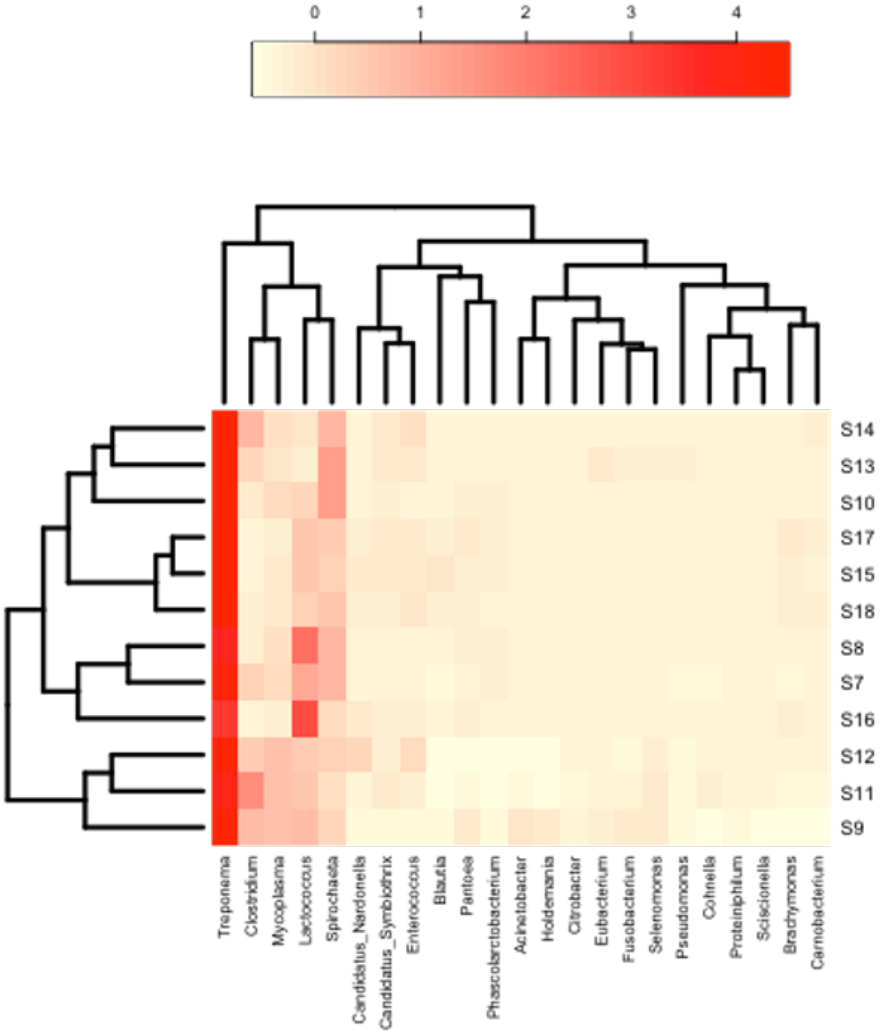
Who is in there?



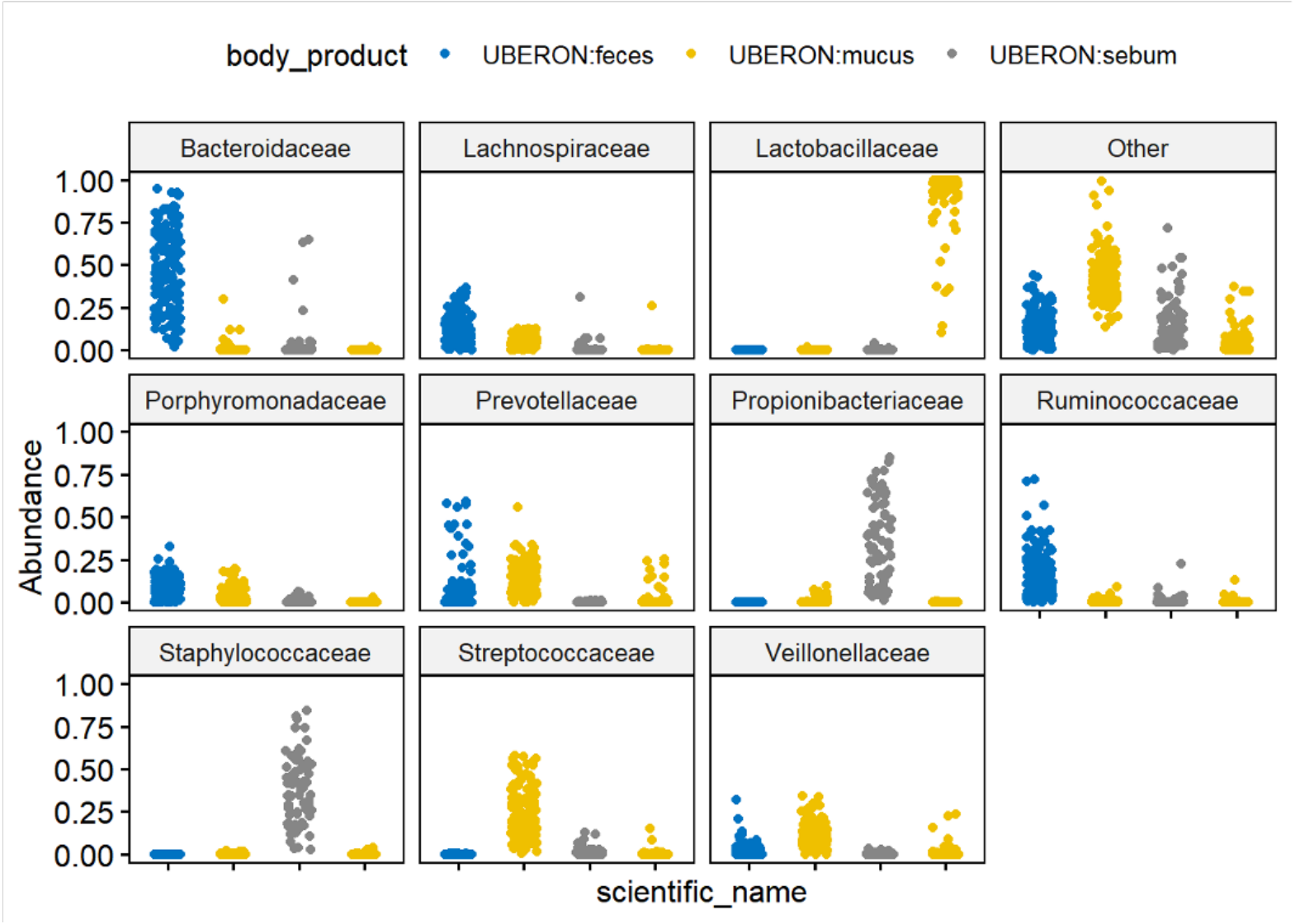
Who is in there?



Who is in there?



Who is in there?



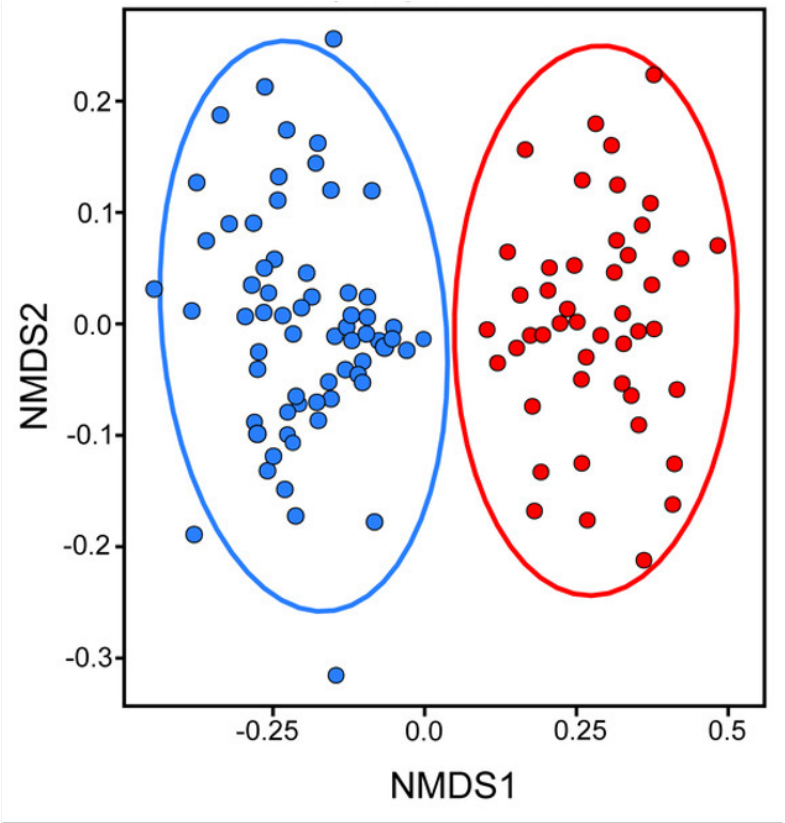
Does community structure vary between treatments?

- Multivariate statistics
  - Dissimilarity matrices (Bray-Curtis, Jaccard, (w)unifrac)
  - Ordination plots (PCA, PCoa, NMDS, ...)
  - Statistical tests (PERMANOVA, ANOSIM, ...)

$$\begin{bmatrix} 0 & & & & & & & \\ 4.15 & 0 & & & & & & \\ 11.02 & 15.01 & 0 & & & & & \\ 7.16 & 3.03 & 18.02 & 0 & & & & \\ 43.72 & 47.49 & 32.80 & 50.41 & 0 & & & \\ 54.37 & 58.23 & 43.36 & 61.19 & 11.12 & 0 & & \\ 46.34 & 50.20 & 35.34 & 53.16 & 3.78 & 8.03 & 0 & \\ 55.42 & 59.27 & 44.42 & 62.23 & 12.05 & 1.12 & 9.08 & 0 \end{bmatrix}.$$

Does community structure vary between treatments?

- Multivariate statistics
  - Dissimilarity matrices (Bray-Curtis, Jaccard, (w)unifrac)
  - Ordination plots (PCA, PCoa, NMDS, ...)
  - Statistical tests (PERMANOVA, ANOSIM, ...)



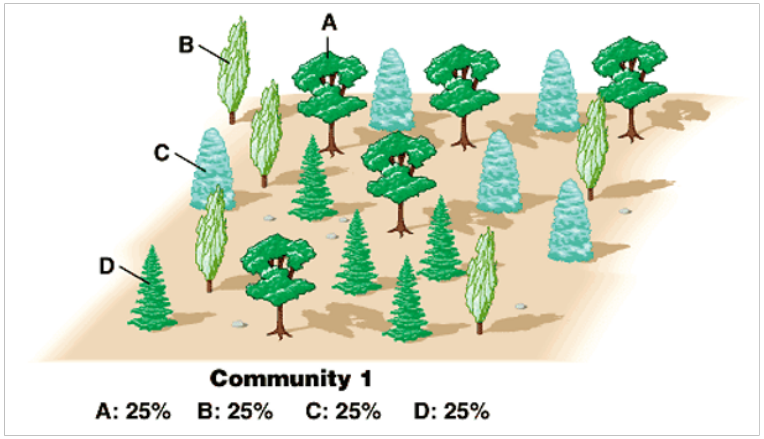
Does community structure vary between treatments?

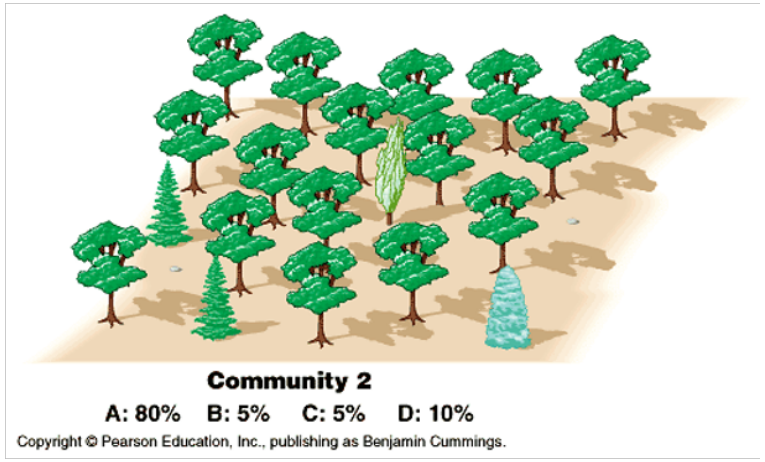
- Multivariate statistics
  - Dissimilarity matrices (Bray-Curtis, Jaccard, (w)unifrac)
  - Ordination plots (PCA, PCoa, NMDS, ...)
  - Statistical tests (PERMANOVA, ANOSIM, ...)

```
##
## Call:
## adonis(formula = erie_bray ~ Station, data = sampledf)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## Station    2    0.6754 0.33772  2.7916 0.09531  0.003 **
## Residuals 53    6.4118 0.12098         0.90469
## Total     55    7.0872                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Does the community diversity vary between treatments?

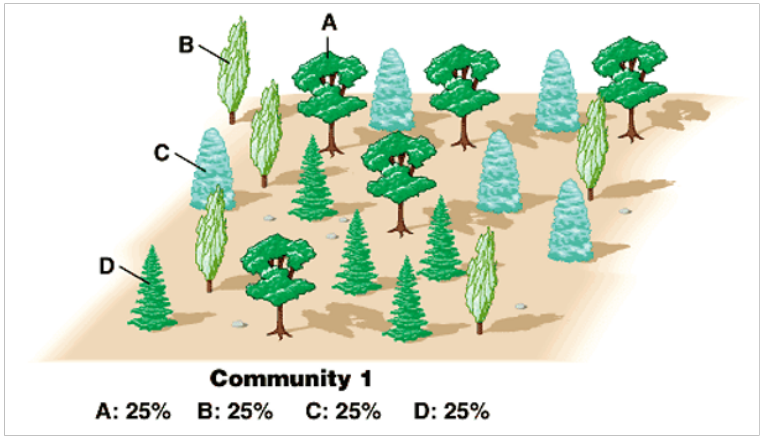
- **Richness** refers to how many different types of organisms are present in a sample
- **Evenness** tells us how even or uneven the distribution of species abundances are in a given environment
- **Diversity** is a measurement of species richness combined with evenness

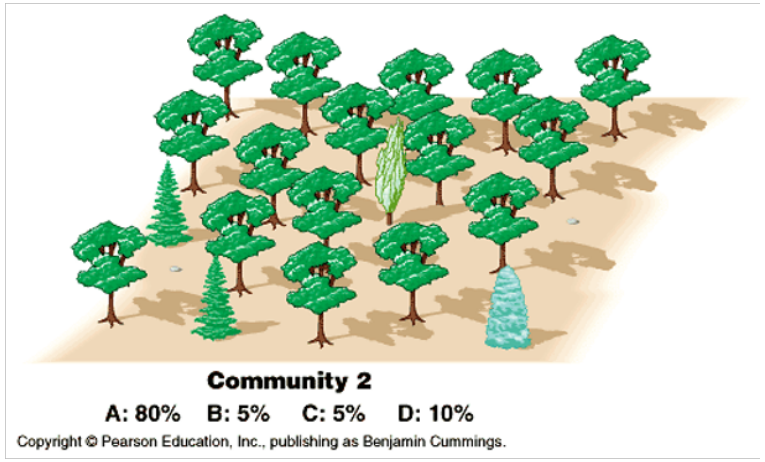




Does the community diversity vary between treatments?

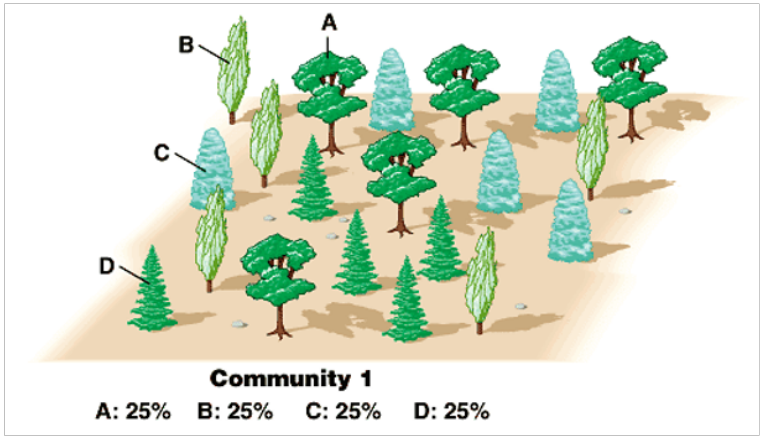
- **Richness** refers to how many different types of organisms are present in a sample
- **Evenness** tells us how even or uneven the distribution of species abundances are in a given environment
- **Diversity** is a measurement of species richness combined with evenness

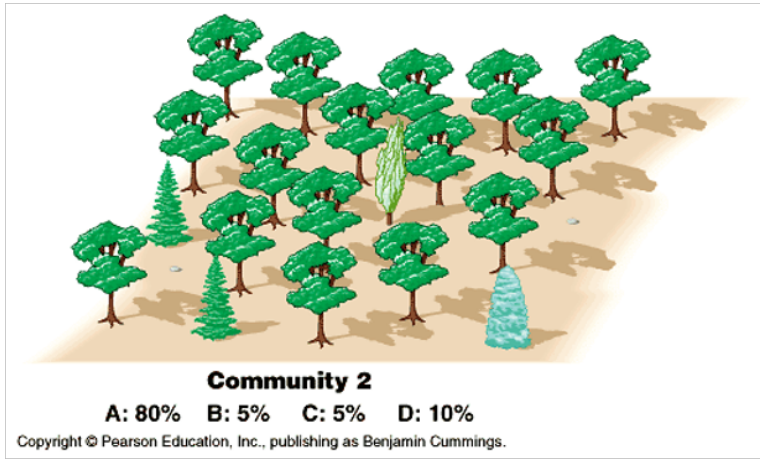




Does the community diversity vary between treatments?

- **Richness** refers to how many different types of organisms are present in a sample
- **Evenness** tells us how even or uneven the distribution of species abundances are in a given environment
- **Diversity** is a measurement of species richness combined with evenness

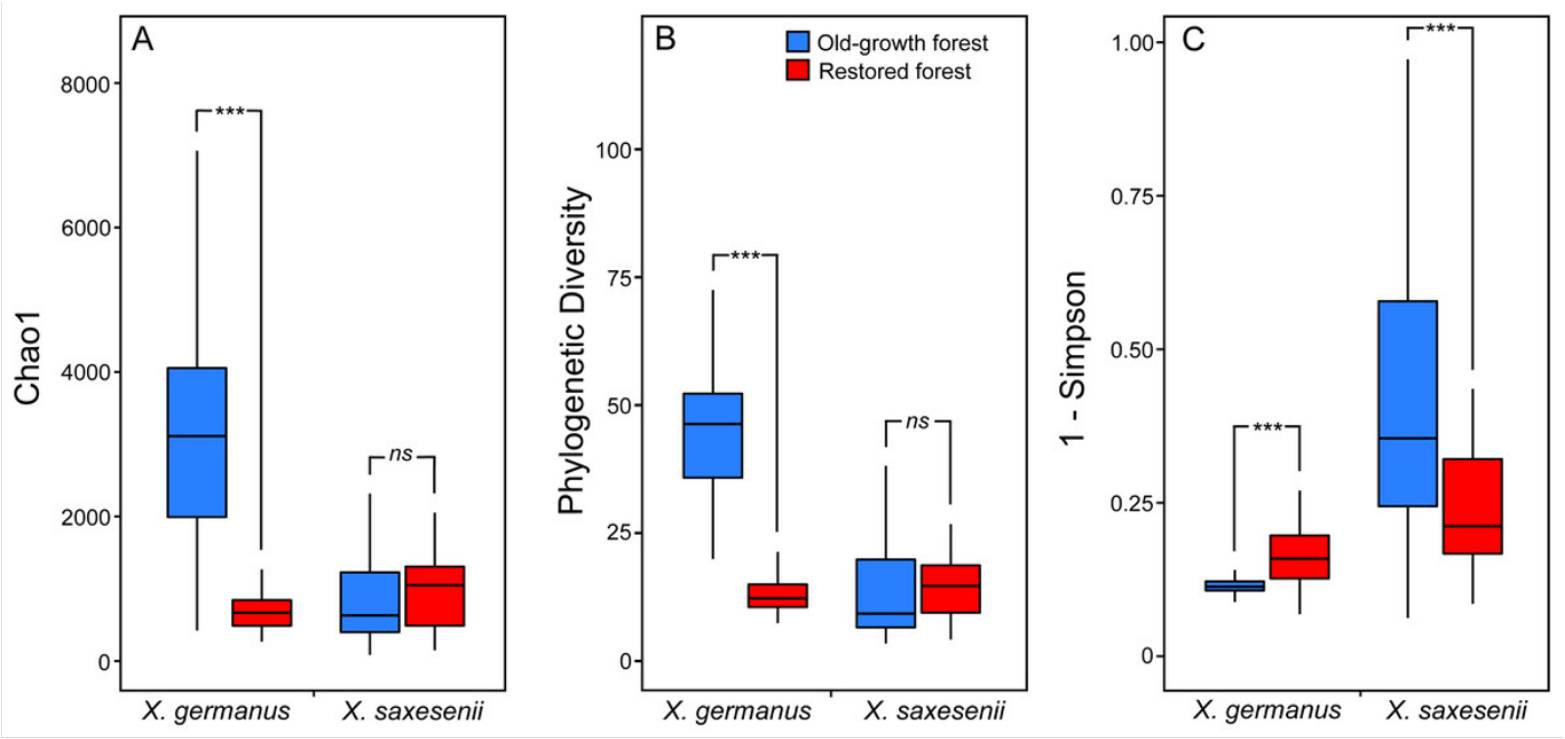




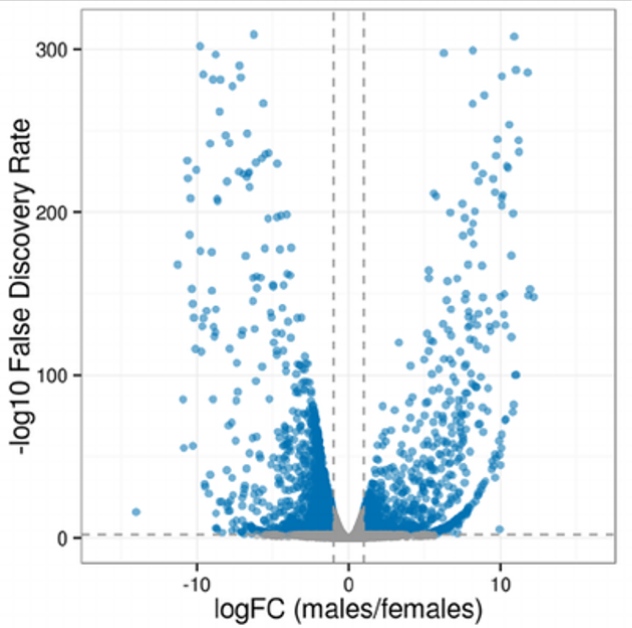
Lots of metrics!

- **Observed OTUs**: we simply count the OTUs that are observed in a given sample
- **Shannon index**assumes all species are represented in a sample and that they are randomly sampled
- **Simpson** index is a dominance index because it gives more weight to common or dominant species.
- **Chao1** index gives more weight to the low abundance species, only the singletons and doubletons are used to estimate the number of missing species
- **PD** is computed simply as the sum of the branch length in a phylogenetic tree that is "covered" or represented in a given sample.

Does the community diversity vary between treatments?



Who differs between treatments?



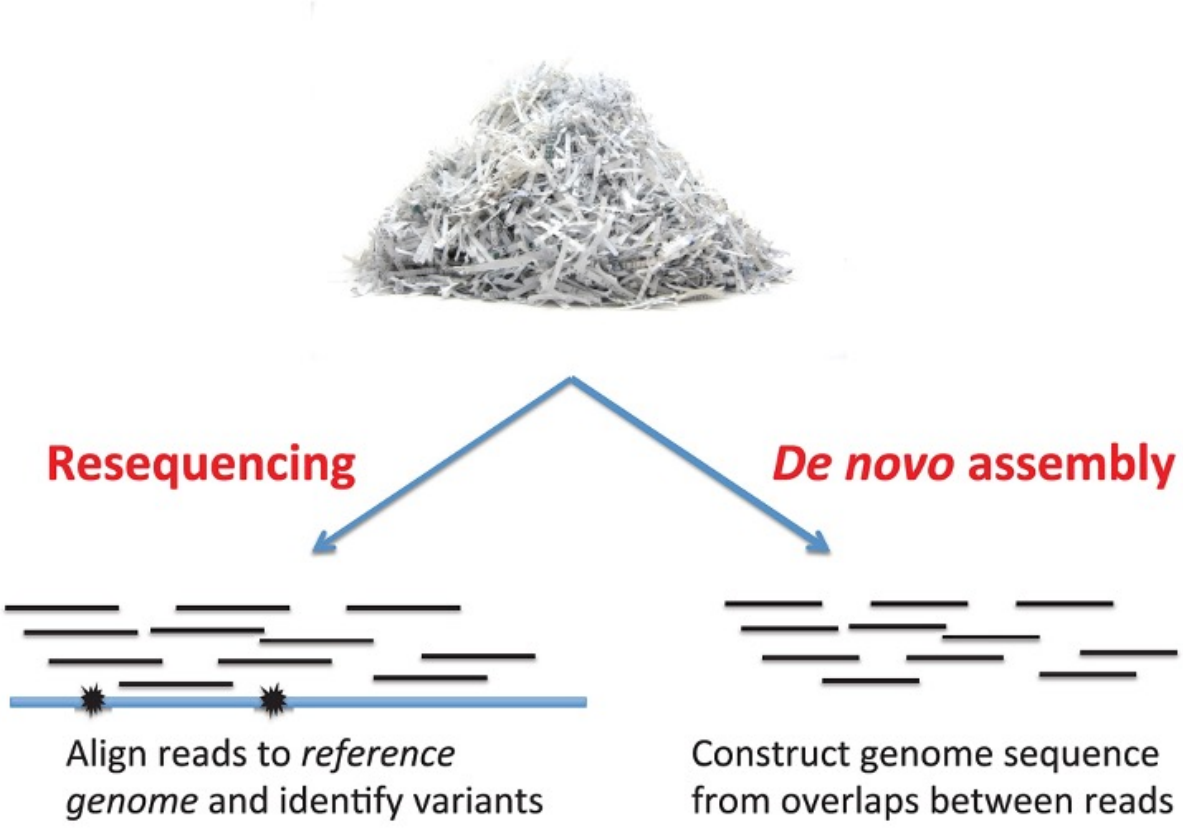
baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	taxon
29.20535	1.91205	0.13432	14.23457	0.00000	0.00000	Clostridium difficile et rel.
51.65152	3.04116	0.28687	10.60107	0.00000	0.00000	Mitsuokella multiacida et rel.
12.39749	1.83825	0.18531	9.91994	0.00000	0.00000	Klebisiella pneumoniae et rel.
44.16494	1.78333	0.23072	7.72937	0.00000	0.00000	Megasphaera elsdenii et rel.
66.93783	1.68345	0.25330	6.64609	0.00000	0.00000	Escherichia coli et rel.
3.63459	1.53142	0.23140	6.61792	0.00000	0.00000	Weissella et rel.
5.74035	3.07334	0.47848	6.42308	0.00000	0.00000	Serratia
0.42171	1.70079	0.47147	3.60743	0.00031	0.00075	Moraxellaceae

More tools

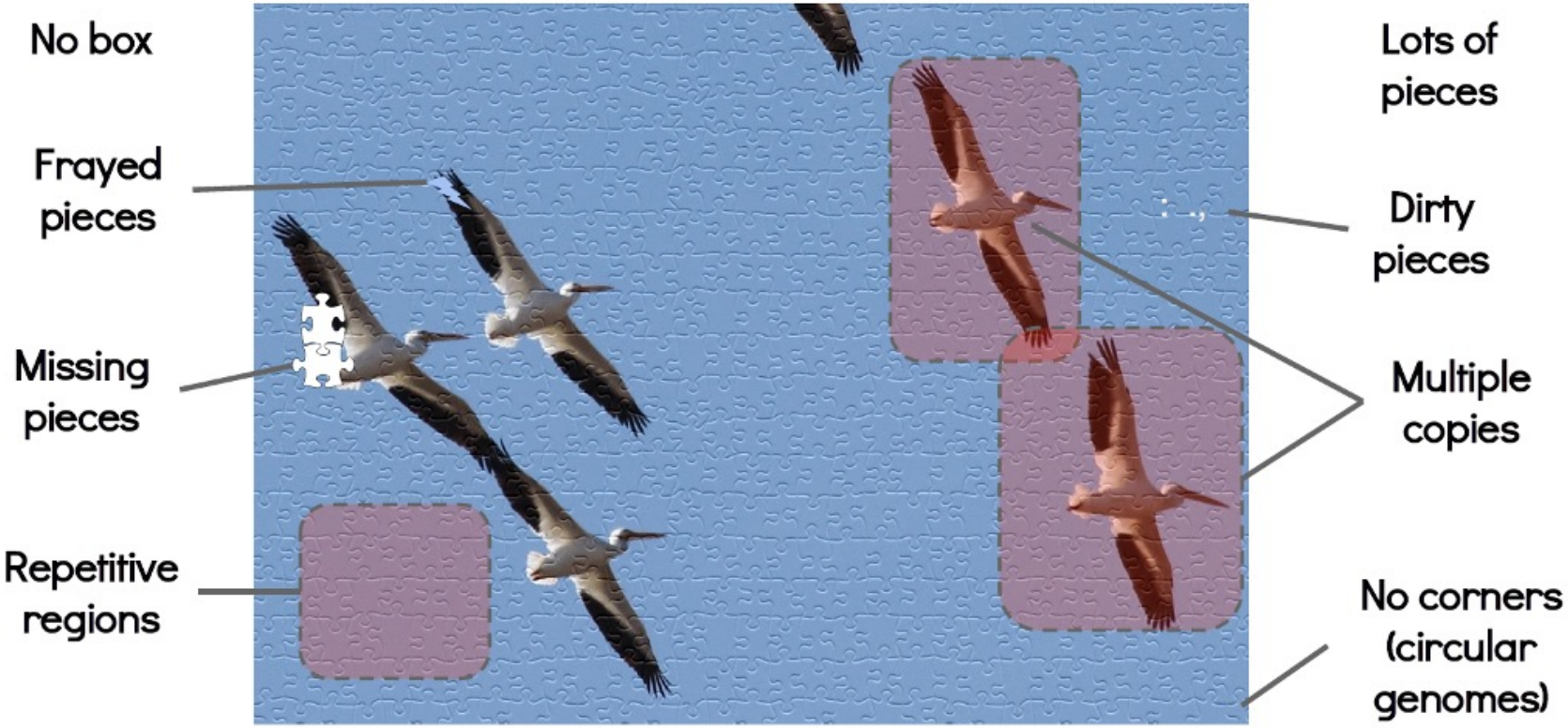
Metagenomics

Metagenomics

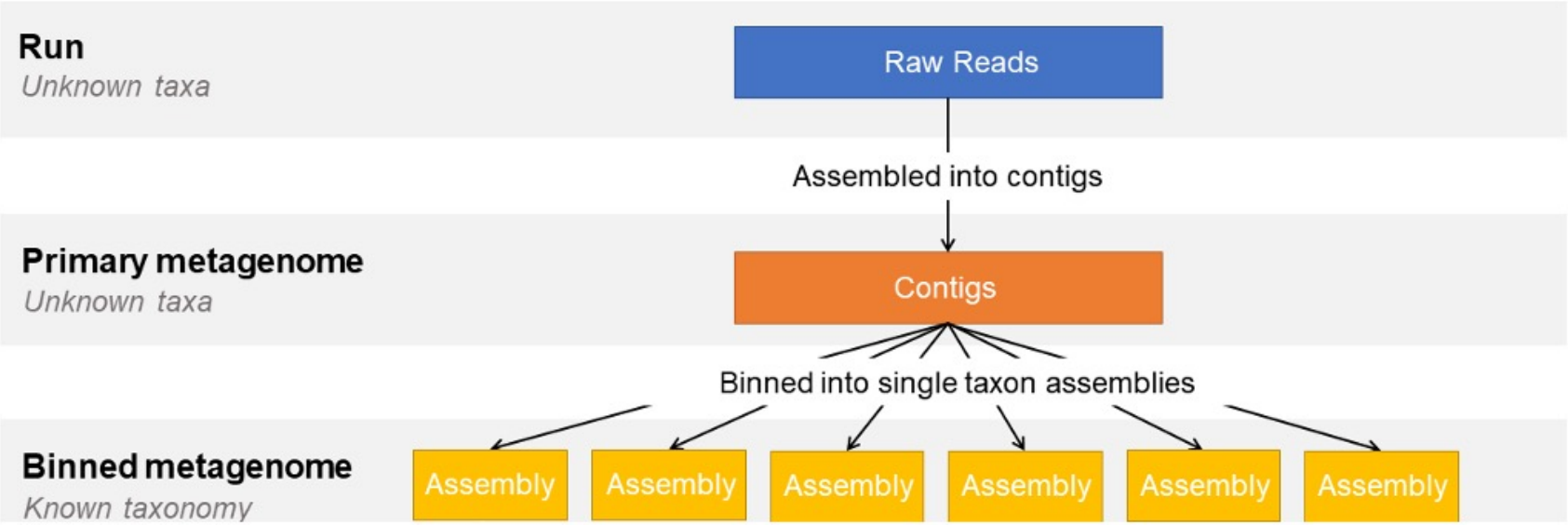
Let's take a step back: genomics



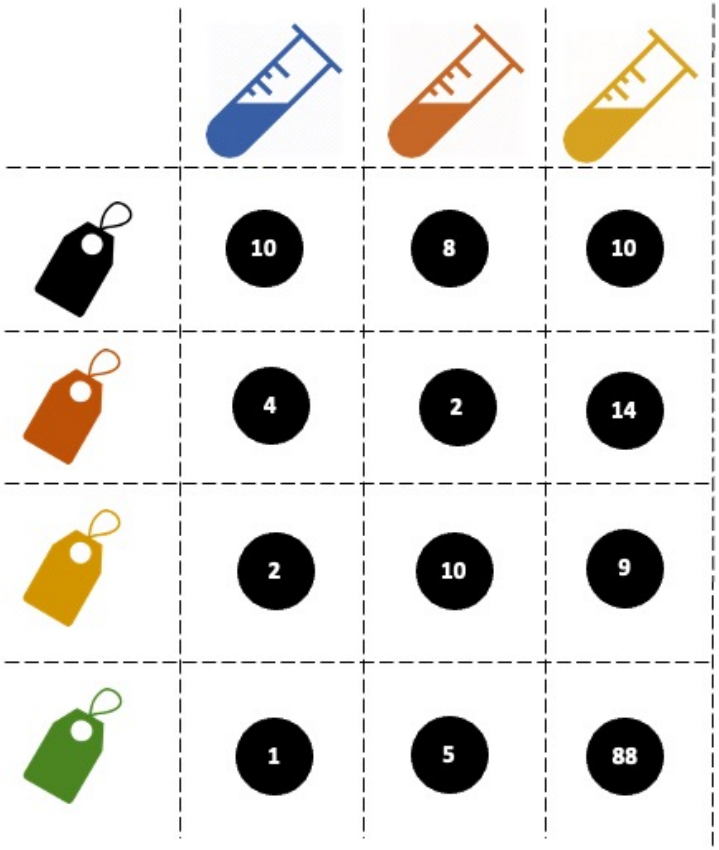
Let's take a step back: genomics



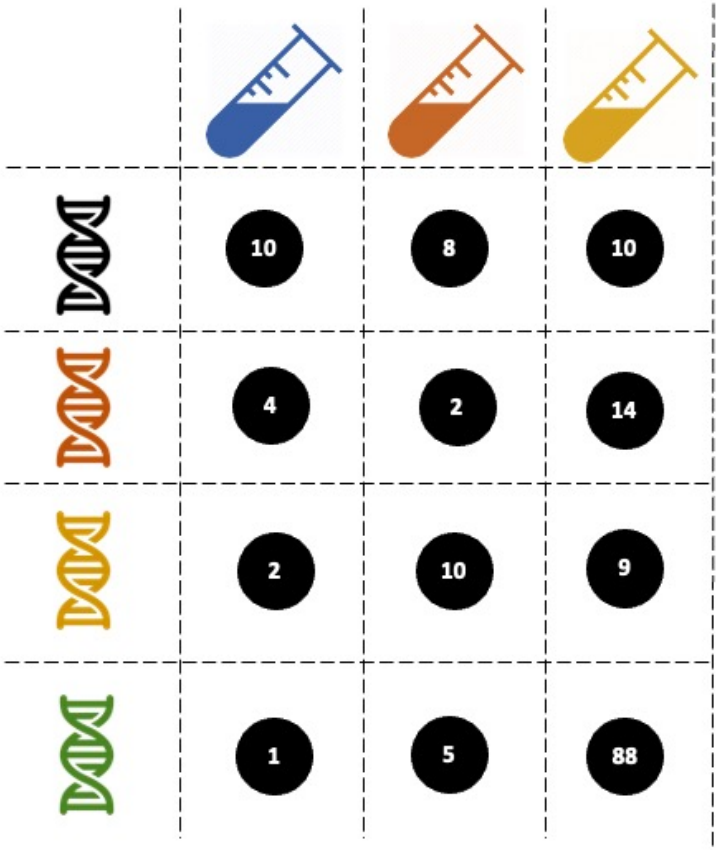
Metagenomics



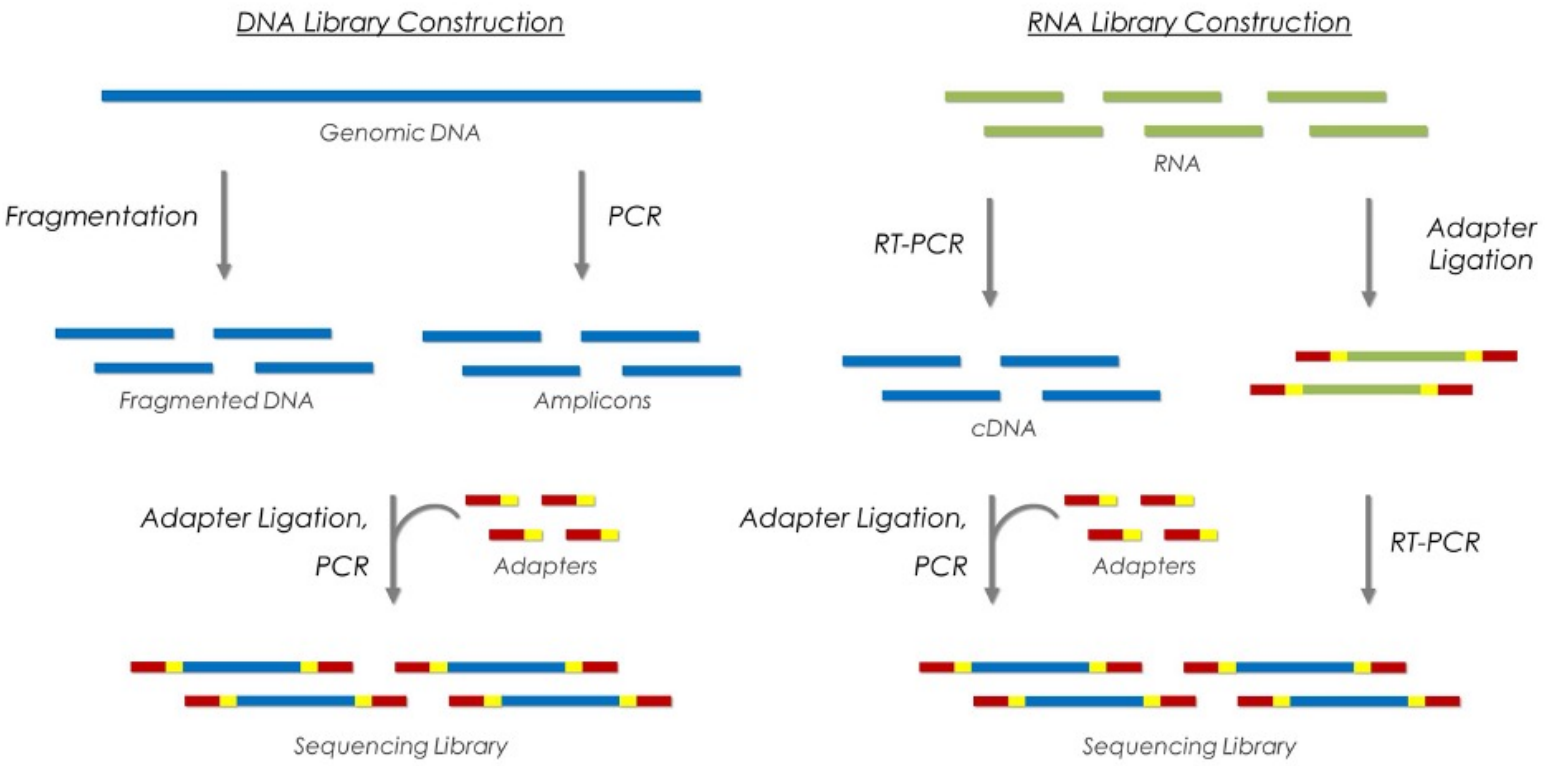
Metagenomics



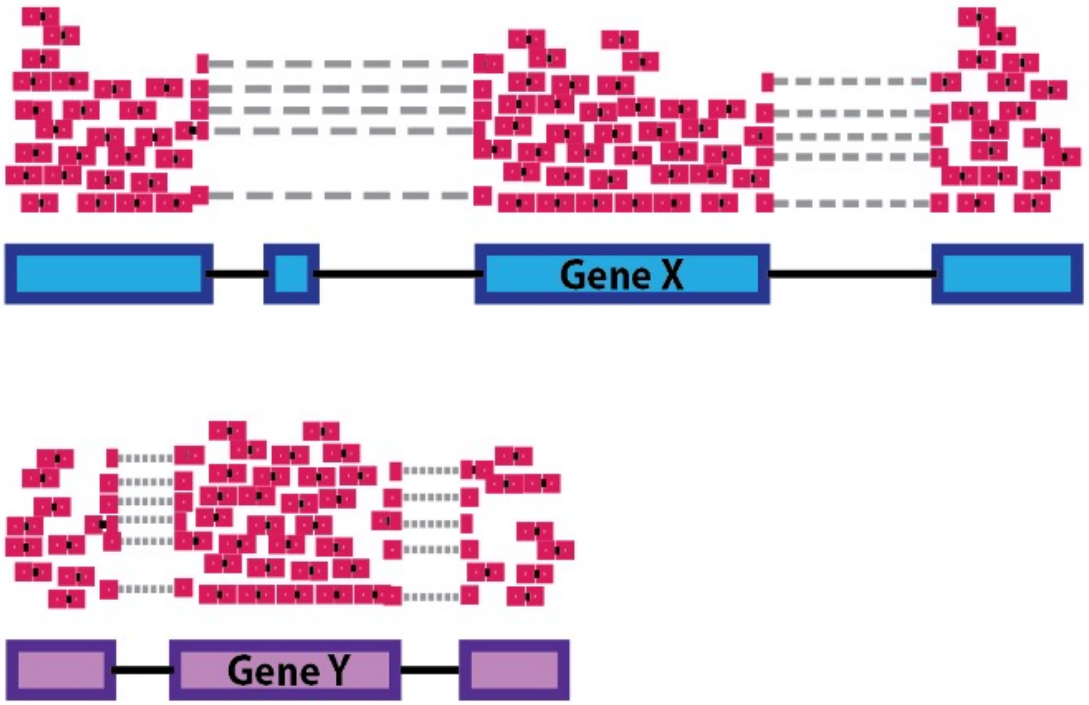
Metagenomics



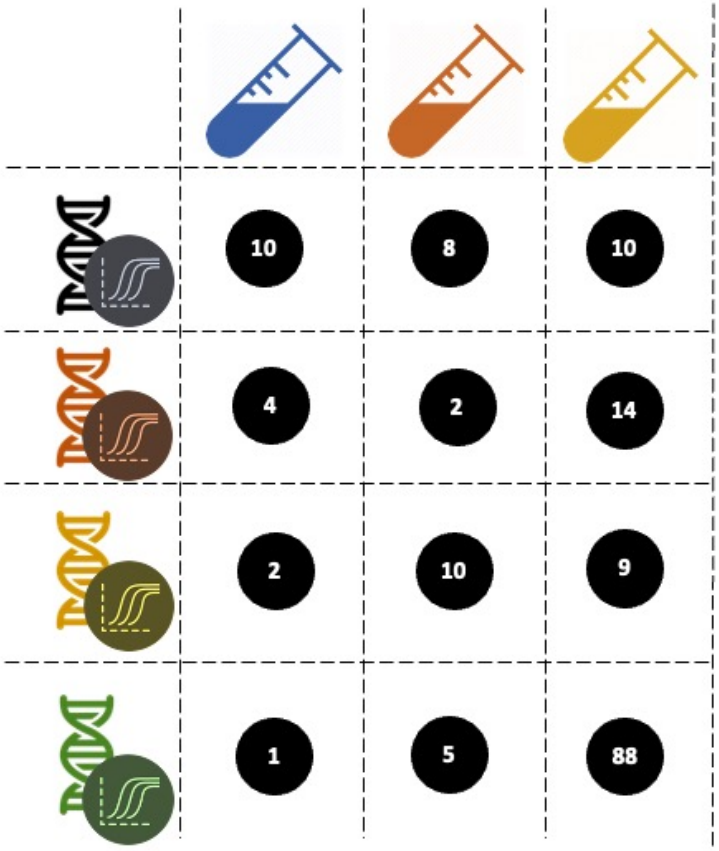
Metatranscriptomics



Transcriptomics



Metatranscriptomics



Questions?