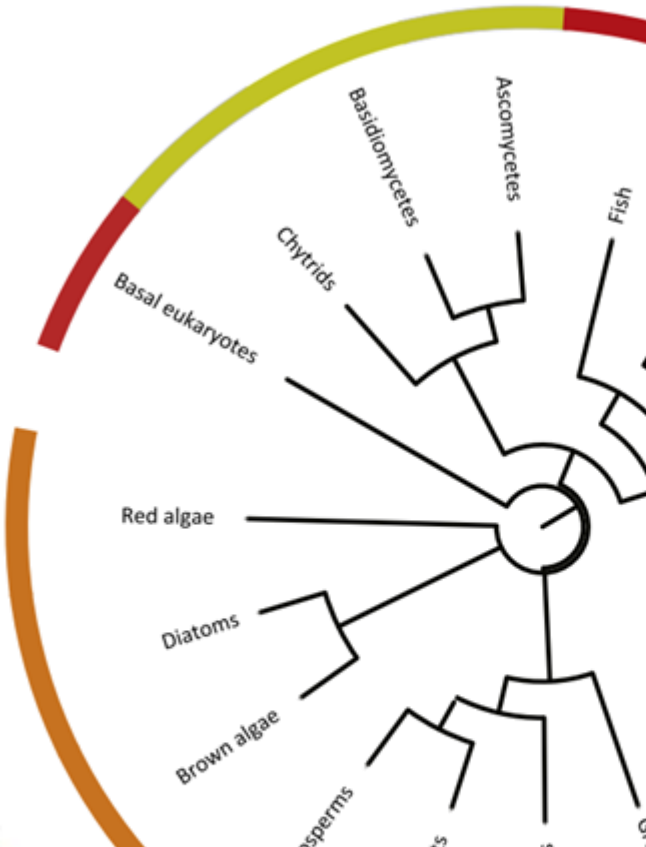
Introduction to Bioinformatics analysis of Metabarcoding data

Theoretical part

Alvaro Sebastián Yagüe

www.sixthresearcher.com

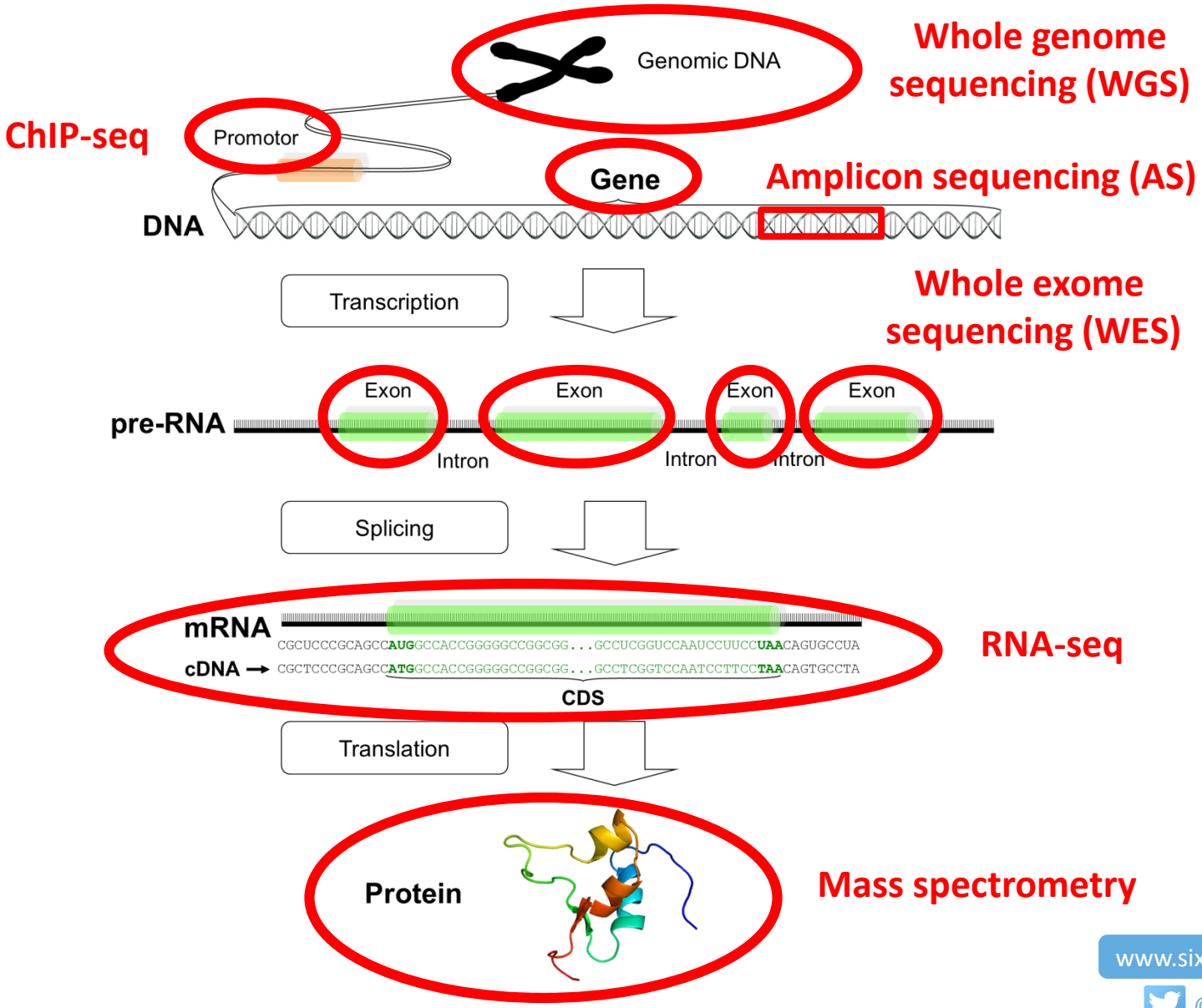
@SixthResearcher



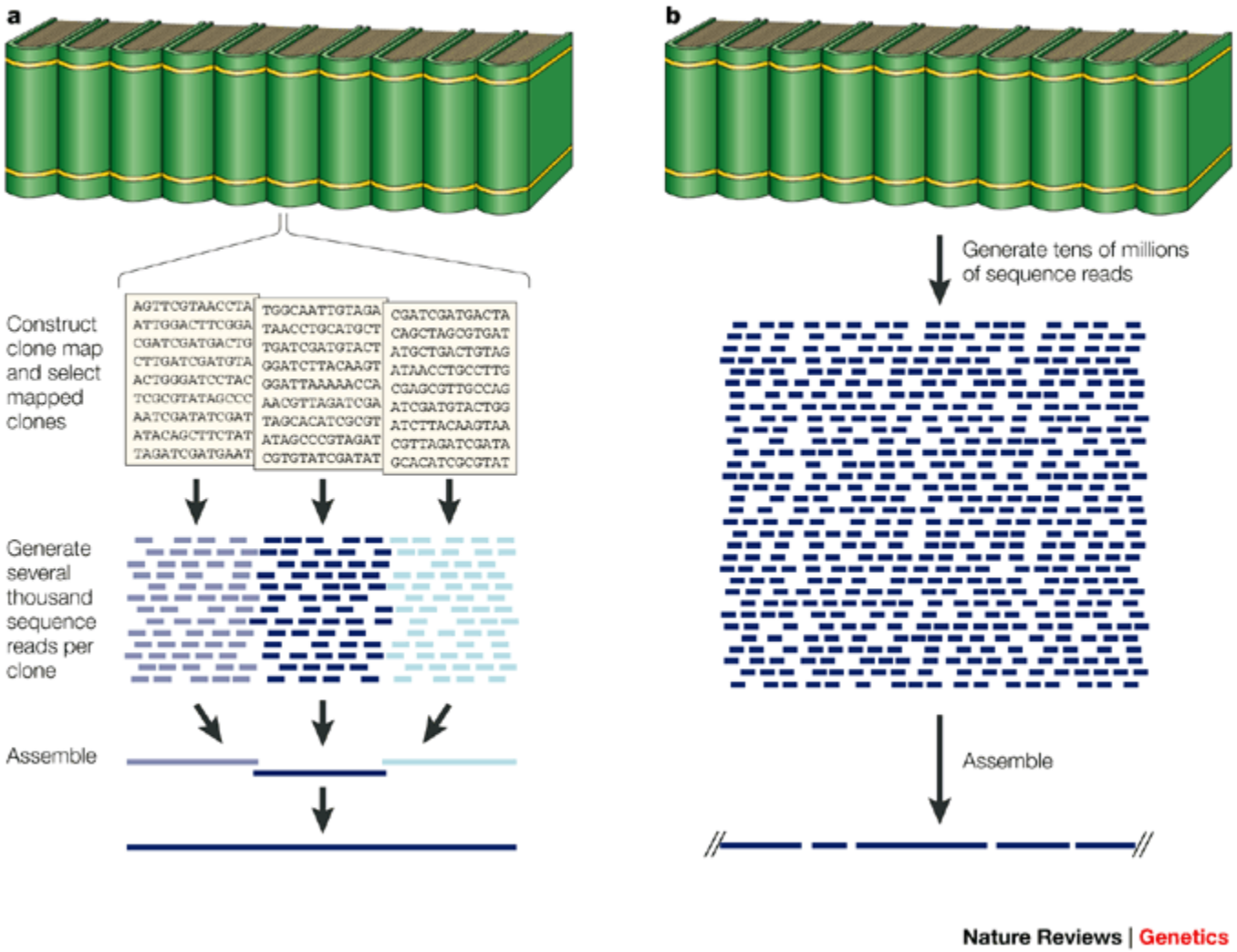
- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

What do we want to sequence?

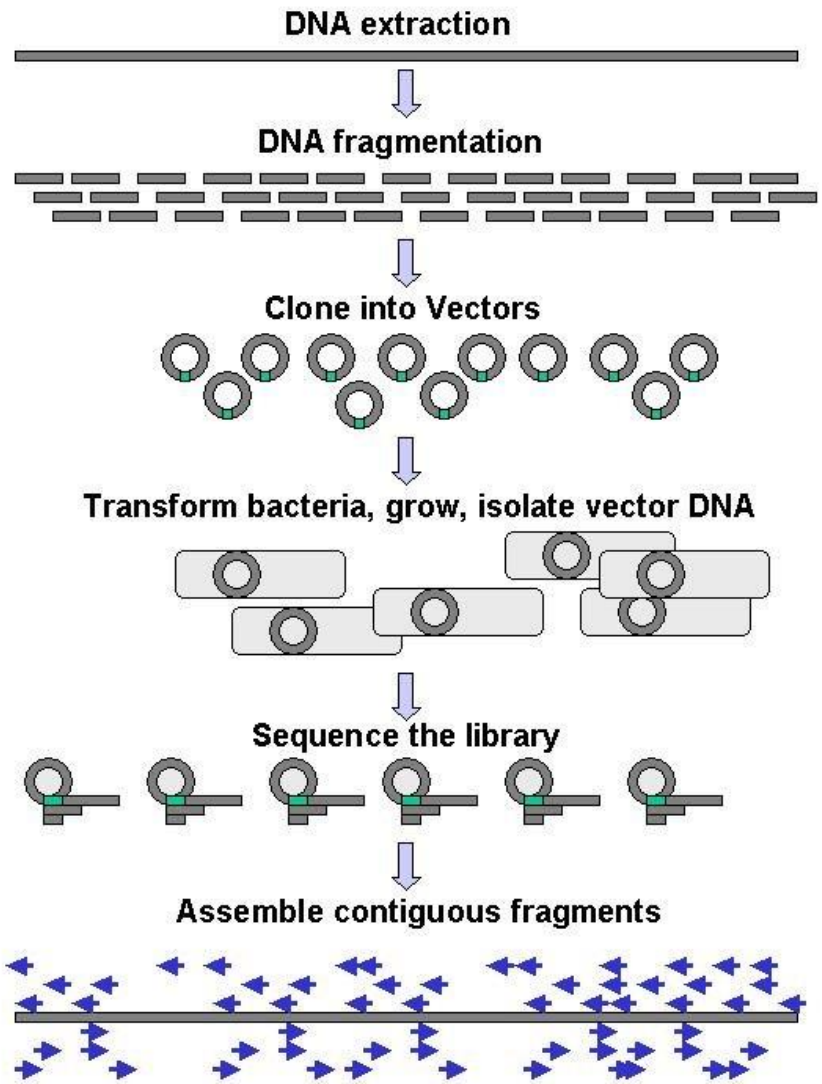


How do we want to sequence?

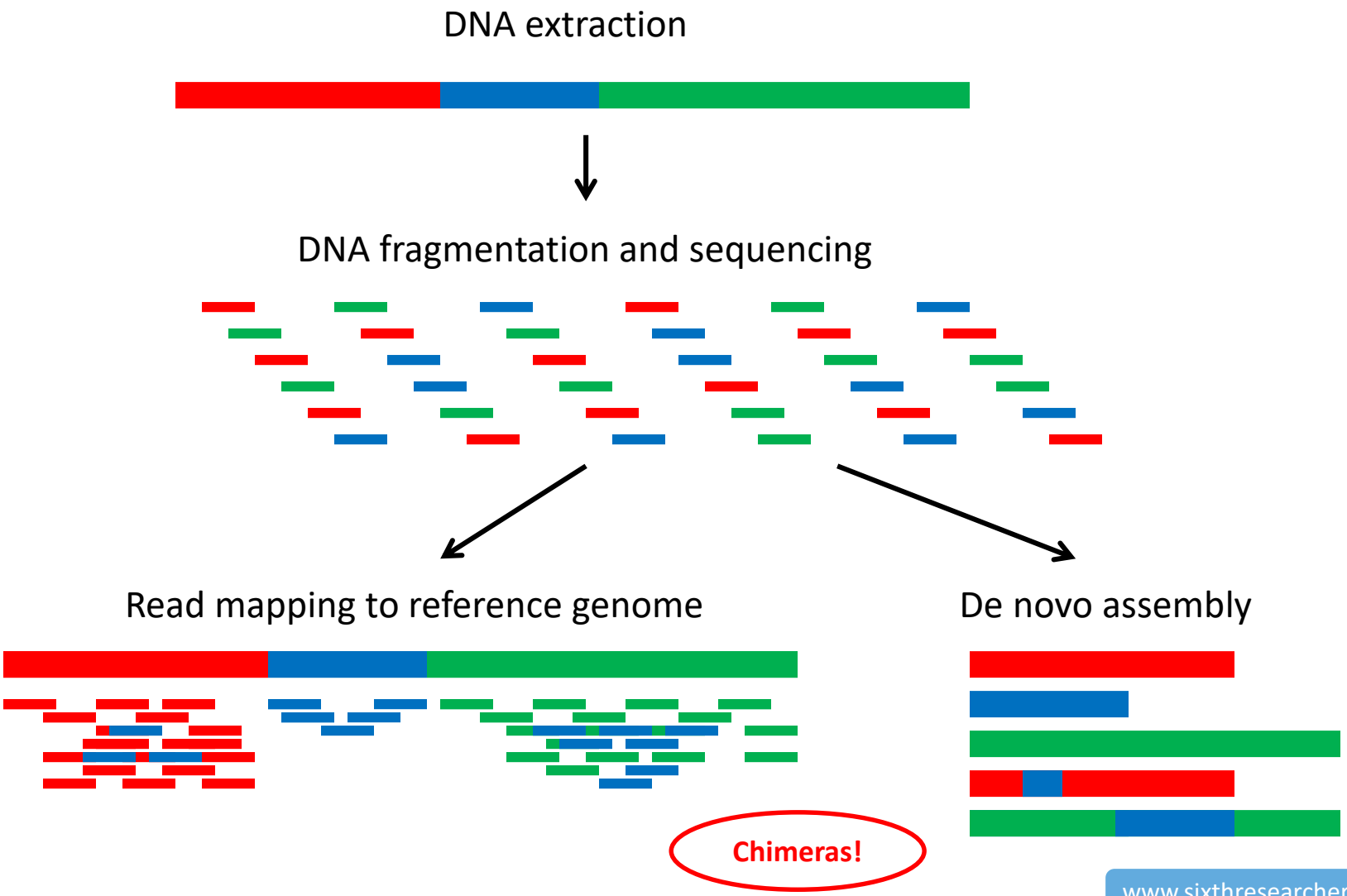


Green,E.D. (2001) Strategies for the systematic sequencing of complex genomes. *Nat. Rev. Genet.*, **2**, 573–583.

Metagenomics - Shotgun sequencing

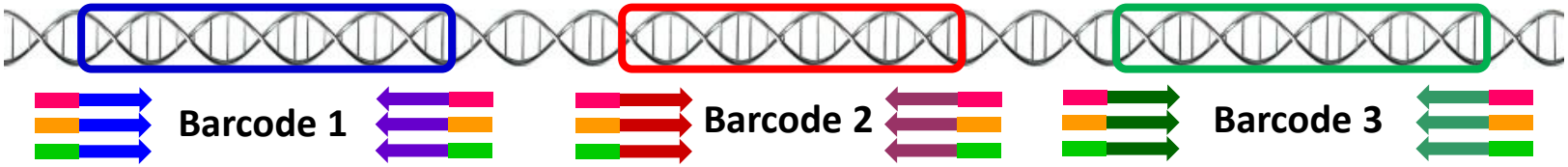


Metagenomics - High-throughput sequencing

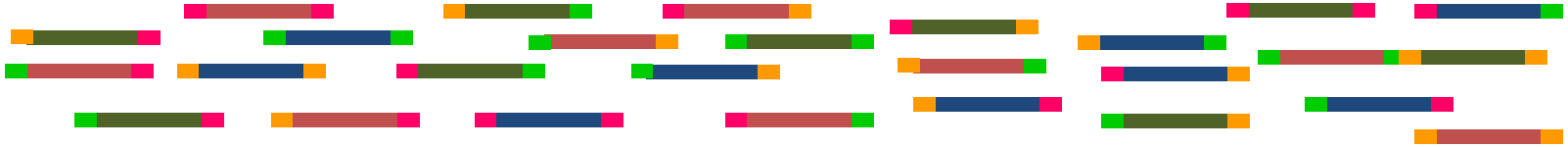


Metabarcoding - Amplicon sequencing

1. PCR amplification and sample tagging



2. Sequencing of PCR products



3. De-multiplexing of reads

	Samples					
	1	2	3	4	5	6
Barcode 1						
Barcode 2						
Barcode 3						

Metabarcoding vs Metagenomics

	DNA Barcoding	Genomics
Species Number	All (or most)	1 (or few)
Gene Region Number	1 (or few)	All (or most)

Kress, W. J., & Erickson, D. L. (2008). DNA barcodes: genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(8), 2761–2.

Metabarcoding vs Metagenomics

	DNA Metabarcoding	Metagenomics
Taxonomic resolution	+ COI sufficient!	+ +
PCR based = Primer bias	— — / 12S / 16S? whobbles? — ?	+ ? explore pot. biases
Taxa missed	— <20%	+
Abundance	— —	— higher potential
Reference database	+ COI / — others	— others / + can use COI
Cost	+	— 10x / — —100x
Potential:	Improved primers	MT enrichment Maybe abundance?
	Short term?	Long term?

http://luckylion.de/2015/10/11/dna-metabarcoding-vs-metagenomics/

- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

- **Where?**

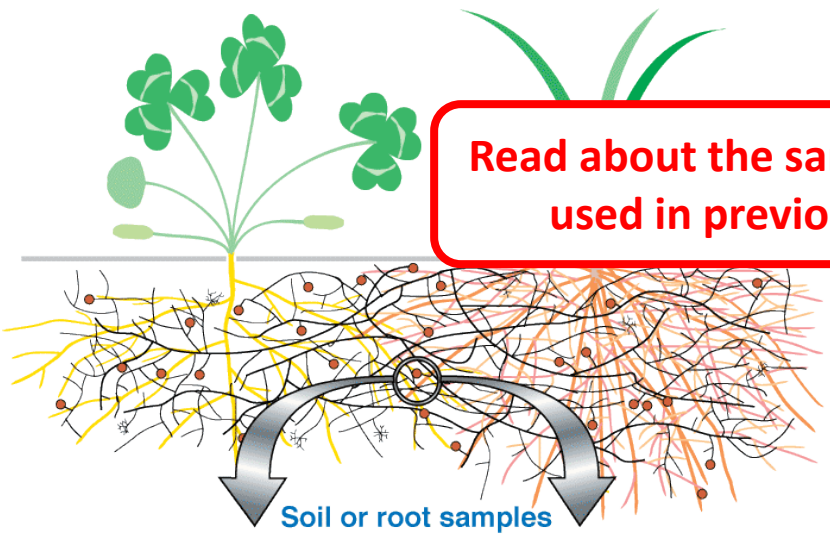
  Variability in abundance within soil and plant.
  Consider vertical and horizontal distribution of fungi.

- **When?**

  Temporal dynamics over short and long term.
  For complete community census, sample across multiple seasons.

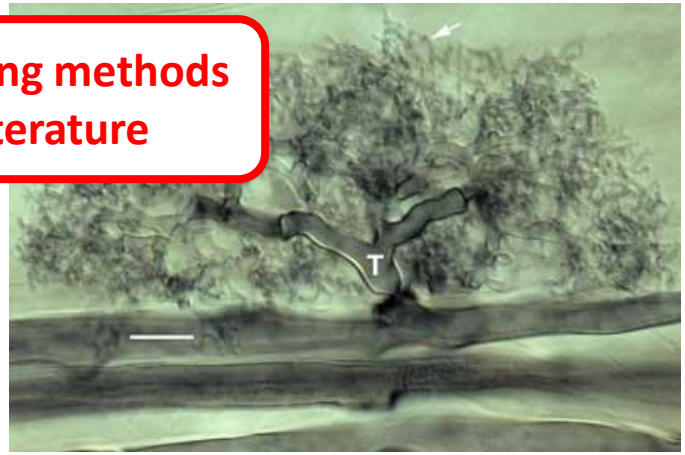
- **How many? How much?**

  Perform power analysis to determine optimal sample size and quantity.



Read about the sampling methods used in previous literature

Arbuscular mycorrhizal fungi



Hart,M.M. *et al.* (2015) Navigating the labyrinth: A guide to sequence-based, community ecology of arbuscular mycorrhizal fungi. *New Phytol.*, **207**, 235–247.

- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

➢ Sample preservation

Sample preservation methods may result in a significant loss of DNA.
Snap-freezing in liquid nitrogen (fast and convenient in the lab but not in field)
Other methods: Ethanol storage, silica-gel drying, freeze-drying, oven-drying at low heat, storage in DNA extraction buffer…

➢ DNA/RNA isolation

Traditional phenol/chloroform extraction.
Modern extraction kits.
Researcher fatigue may result in later samples being handled less efficiently.
Samples processed early in the protocol will be exposed to variable conditions longer.

➢ Internal controls

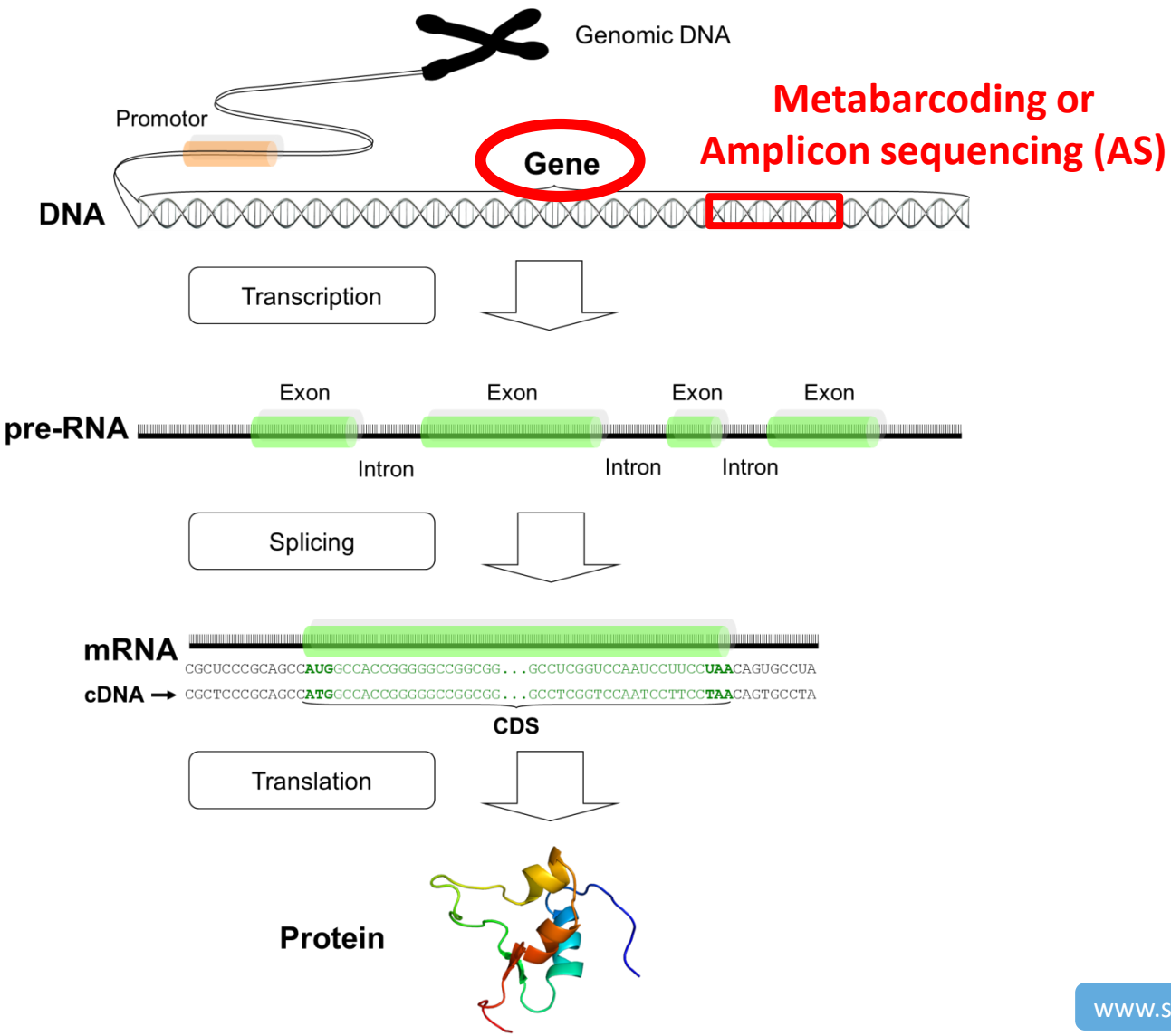
Internal standards, as a initial known quantity of DNA, will provide a measure of DNA yield.
Especially important for samples that originate from different environments.
Should be used to quantify DNA/RNA recovery.
Will validate the accuracy of results from further analyses.
A 'blank' sample (negative control) will help to control contaminations during the process.

Sometimes our DNA of interest will be rare compared with other DNAs present in the samples

Hart,M.M. *et al.* (2015) Navigating the labyrinth: A guide to sequence-based, community ecology of arbuscular mycorrhizal fungi. *New Phytol.*, **207**, 235–247.

- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

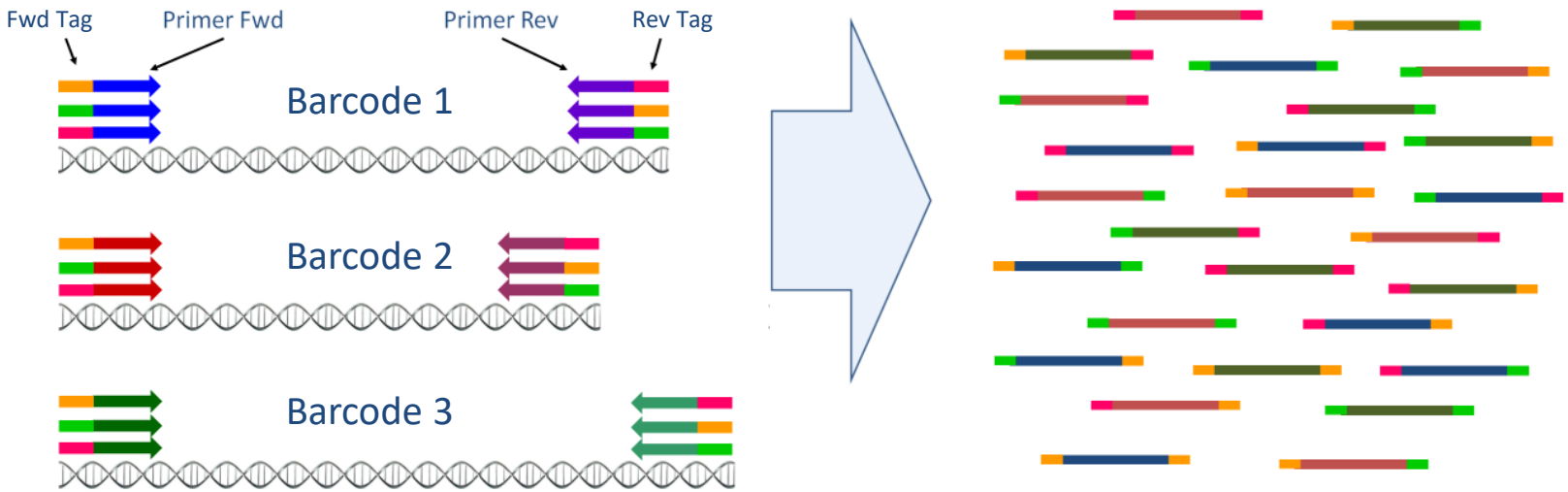
What do we want to sequence?



Metabarcoding - Amplicon sequencing

Choose barcodes, design primers and add tags

PCR amplification and sequencing



Barcodes, markers and tags

A DNA BARCODE is…

a standardized short sequence of DNA (400–800 bp) that in principle should be easily generated and characterized for all species on the planet. A massive on-line digital library of barcodes will serve as a standard to which any DNA barcode sequence of an unidentified environmental sample from sea, soil, air, etc. can be matched.

Savolainen et al. 2005

A GENETIC MARKER is…

a specific gene or DNA sequence that produces a detectable trait with a known location on a chromosome and that can be used to study family and population, identification of cells, species or individual.

www.biotecharticles.com

So… a DNA barcode is a type of genetic marker.

A DNA TAG is…

A unique short DNA sequence that identifies unambiguously a sample. DNA tags are usually ligated after PCR amplification or directly included in one or both primers.

Which barcode to choose?

A perfect barcode should…

- ✓ be present in all the organisms, in all the cells
- ✓ have variable sequence among different species
- ✓ be conserved among individuals of the same species
- ✓ be easy to amplify with conserved flanking sites
- ✓ be not too long for sequencing

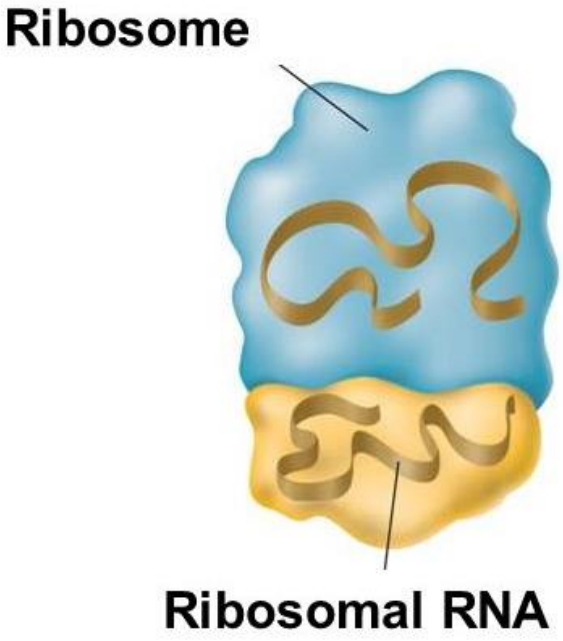
Which barcode to choose?

Ribosomes contain two major rRNAs and 50 or more proteins.

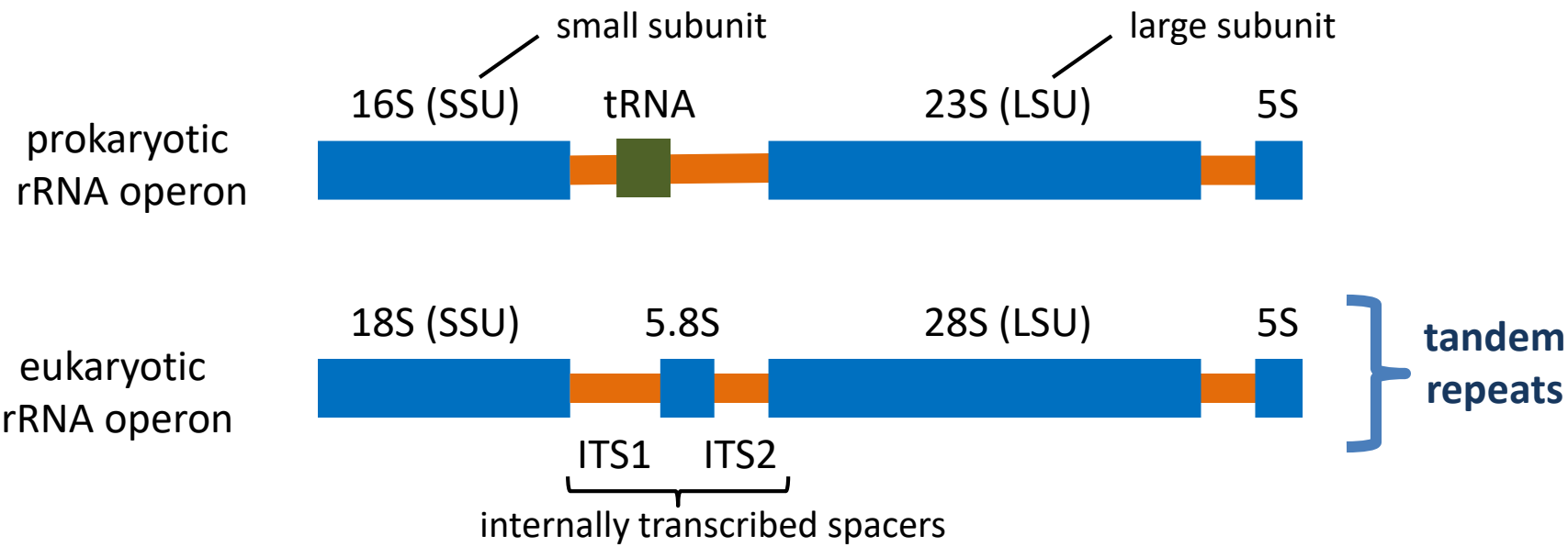
The ribosomal RNAs form two subunits, the **large subunit (LSU)** and **small subunit (SSU).**

rRNA is one of only a few gene products **present in all organisms and in all cells**.

For this reason, genes that encode the rRNA (rDNA) are **very good barcodes** to identify an organism's taxonomic group, calculate related groups, and estimate rates of species divergence.

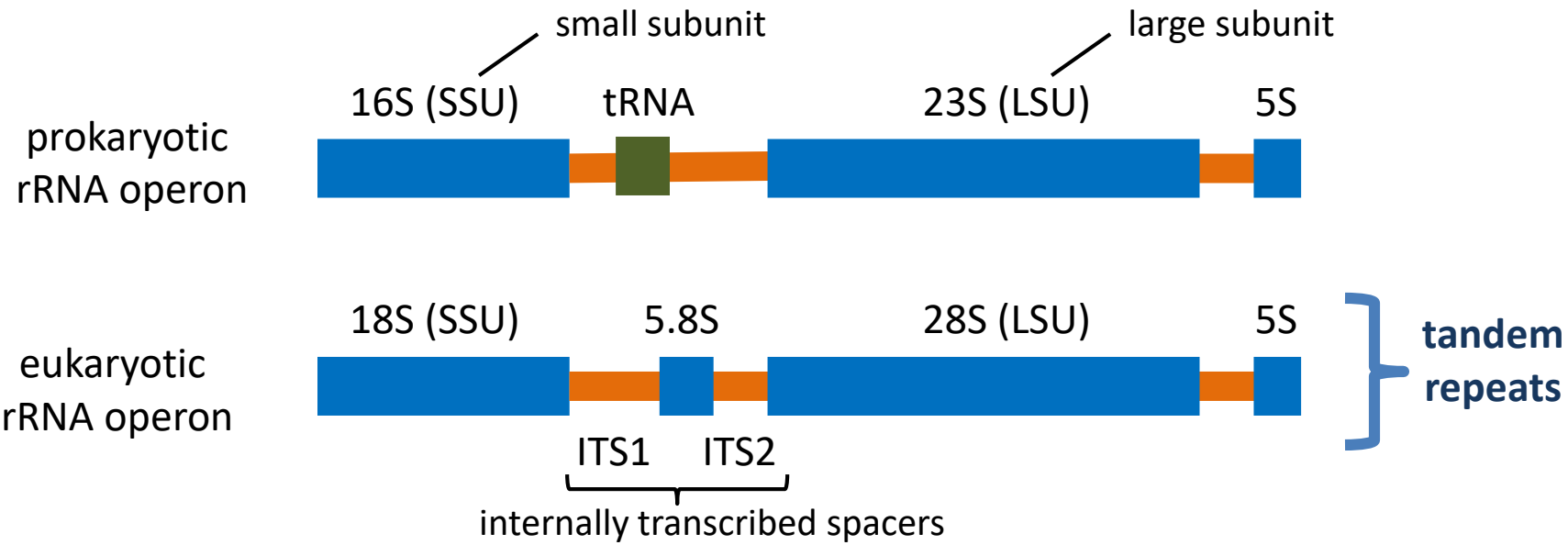


Which barcode to choose?



Type	LSU	SSU
prokaryotic	5S - 120 bp 23S - 2906 bp	16S - 1542 bp
eukaryotic	5S - 121 bp 5.8S - 156 bp 28S - 5070 bp	18S - 1869 bp

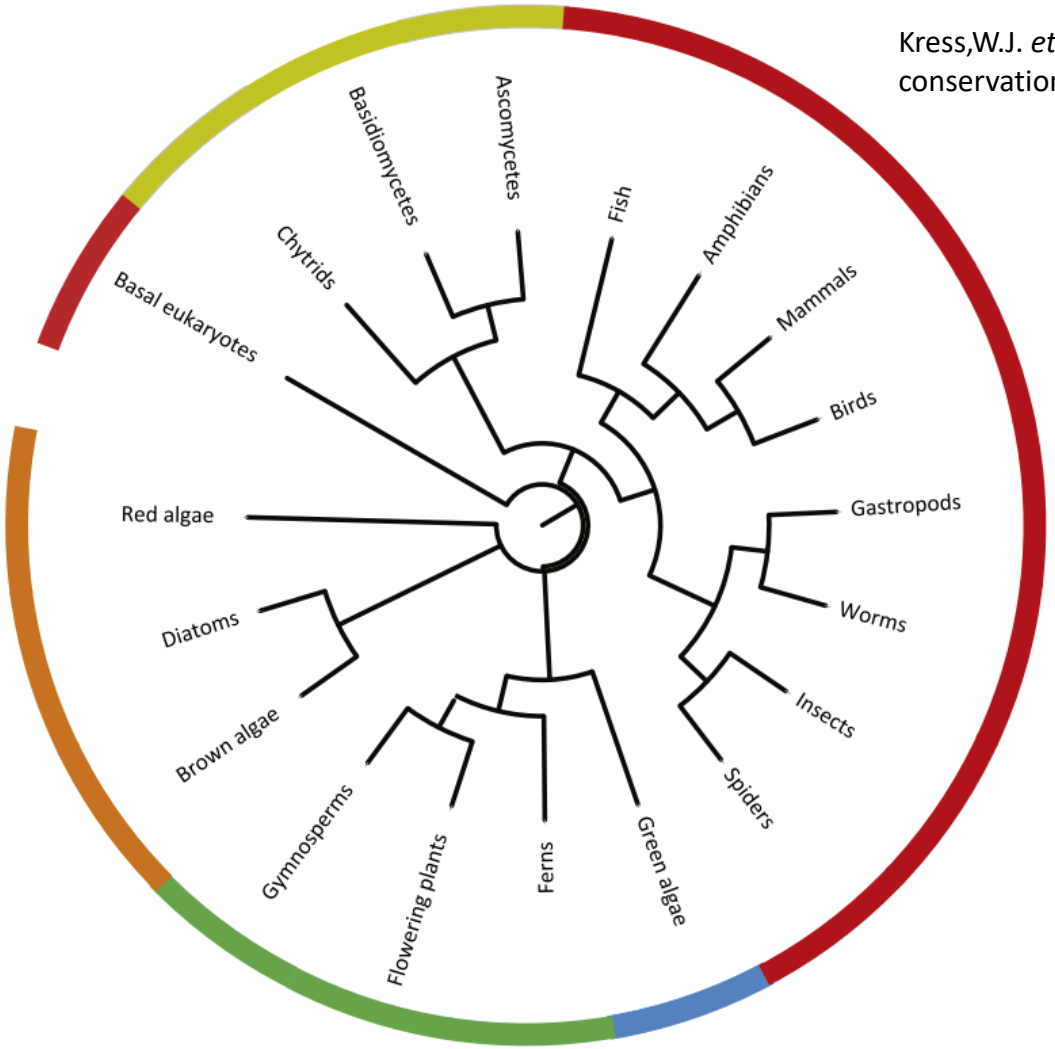
Which barcode to choose?



The ribosomal operon offers the greatest resolution when used as a whole.

Unfortunately, the ribosomal operon is in excess of 5500 bp (prokaryotic), which is intractable for Sanger sequencing and for current NGS technologies.

Which barcode to choose?



Kress,W.J. *et al.* (2014) DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.*, **30**, 25–35.

Tree of life

Key: Color	Clade	Primary barcode(s)	Secondary barcode(s)
	Animals	CO1	CO1, 16S
	Fungi	ITS	LSU D1/D2
	Green algae	*tuf*A	LSU D2/D3
	Land plants	*rbc*L/matK	*psb*A-*trn*H/ITS
	Algae	CO1-5P	LSU D2/D3
	Bacteria/ Archae	16S	RIF

CO1: cytochrome c oxidase subunit 1
ITS: internally transcribed spacer
LSU: large subunit rRNA
D1/D2/D3: divergent domains
RIF: DnaA replication initiation factor

http://www.barcodeoflife.org/

Which barcode to choose?

- **Ideally**, a single DNA barcode (also called marker) would be used to recognize organisms at organizational levels from genotype to kingdom.

- **In reality**, there is no de facto best sequence target that would achieve all aims.

Discriminating taxa at the species level requires a more variable sequence (barcode) than at the genus or family level.

Most studies have focused only on identifying taxa, but protein-encoding genes with known functions may become important functional barcodes for future community surveys.

Hart,M.M. *et al.* (2015) Navigating the labyrinth: A guide to sequence-based, community ecology of arbuscular mycorrhizal fungi. *New Phytol.*, **207**, 235–247.

Primer design and barcoding

Barcode	Forward primer	Reverse primer
Microbial 16S rRNA V3-V5 region	CCGTCAATTCMTTTRAGT	CTGCTGCCTCCCGTAGG

Sample	Forward tag	Reverse tag
S001	AACGCG	AAGACA
S002	TCACTC	CGTCAC
S003	CTTGGT	TTGAGT
S004	TGGAAC	TAACAT
S005	CGAATC	GGTCGA
…	…	…

Which sequencing technology to choose?

Sanger ABI

Ion Torrent

454

Illumina

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500
Ion Torrent	2nd	200-400	1	8.2×10^7	2–4 h	0.1
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57
Illumina MiSeq	2nd	2 x 300	0.1	2.5×10^7 (paired)	4-55 h	0.15
Illumina HiSeq 2500 (High Output)	2nd	2 × 125	0.1	8×10^9 (paired)	7–60 h	0.03
Illumina HiSeq 2500 (Rapid Run)	2nd	2 × 250	0.1	1.2×10^9 (paired)	1–6 days	0.04
SOLiD 5500×1	2nd	2 × 60	5	8×10^8	6 days	0.11
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90

Rhoads,A. and Au,K.F. (2015) PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.*, **13**, 278–289.

What do we have after sequencing?

- Each sequencing technology outputs different kind of results.

Technology	Output format	Description
Sanger ABI	ABI	It contains the 'trace data' i.e. the probabilities of the 4 bases along the sequencing run, together with the sequence, as deduced from that data.
454 (Roche)	SFF	Binary format that provides flowgrams or measurements that estimate the length of the next homopolymer stretch in the sequence (i.e., in "AAATGG", "AAA" is a 3-mer stretch of A's).
Ion Torrent	BAM	Binary form of the SAM format. It contains the information for each sequence about where/how it aligns or not to a reference.
Illumina	FASTQ	Text-based format for storing both sequences and its corresponding quality scores.

The most accepted NGS standard format is FASTQ.

FASTQ read format

There are many tools to convert the different formats to FASTQ, e.g.:

http://sequenceconversion.bugaco.com/converter/biology/sequences/

A FASTQ file has the following look:

1st line: @ IDENTIFIER
2nd line: SEQUENCE
3rd line: + OPTIONAL
4th line: QUALITY

```
@M01530:20:000000000-A89BL:1:1102:17014:3847
TCACTCGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGTTCCTGGAGAGAC
+
A2FHGGGGGGGGG?FE//><CGC>..-@CCD.<0=.<-;DGAACFBFB0?DGGGC

@M01530:20:000000000-A89BL:1:1102:17988:3900
CCGGAAGAGTGTCATTTCTCCAACGGGACGGAGCAGATACGGTTCCTGGACAGAT
+
>EFHGGGGEGEGGGAFGGHGGGGCFECGGGGGGHDDGGGHHGGGHGHHHGGGGGG

@M01530:20:000000000-A89BL:1:1102:19310:3936 1:N:0:1
AACCGAGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGTTCCTGGACAGAT
+
GGHHGHHADGFGGDCFGDF@CEGCFFBGEFFHGFF<CFHFGDGDGE0/A>/HF@E
```

FASTQ read format

There are many tools to convert the different formats to FASTQ, e.g.:

http://sequenceconversion.bugaco.com/converter/biology/sequences/

A FASTQ file has the following look:

1st line: @ IDENTIFIER → @M01530:20:000000000-A89BL:1:1102:17014:3847

2nd line: SEQUENCE → TCACTCGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGTTCCTGGAGAGAC

3rd line: + OPTIONAL → +

4th line: QUALITY → A2FHGGGGGGGGG?FE//><CGC>..-@CCD.<0=.<-;DGAACFBFB0?DGGGC

Each ASCII char represents a Phred quality score:

17 32 37 39

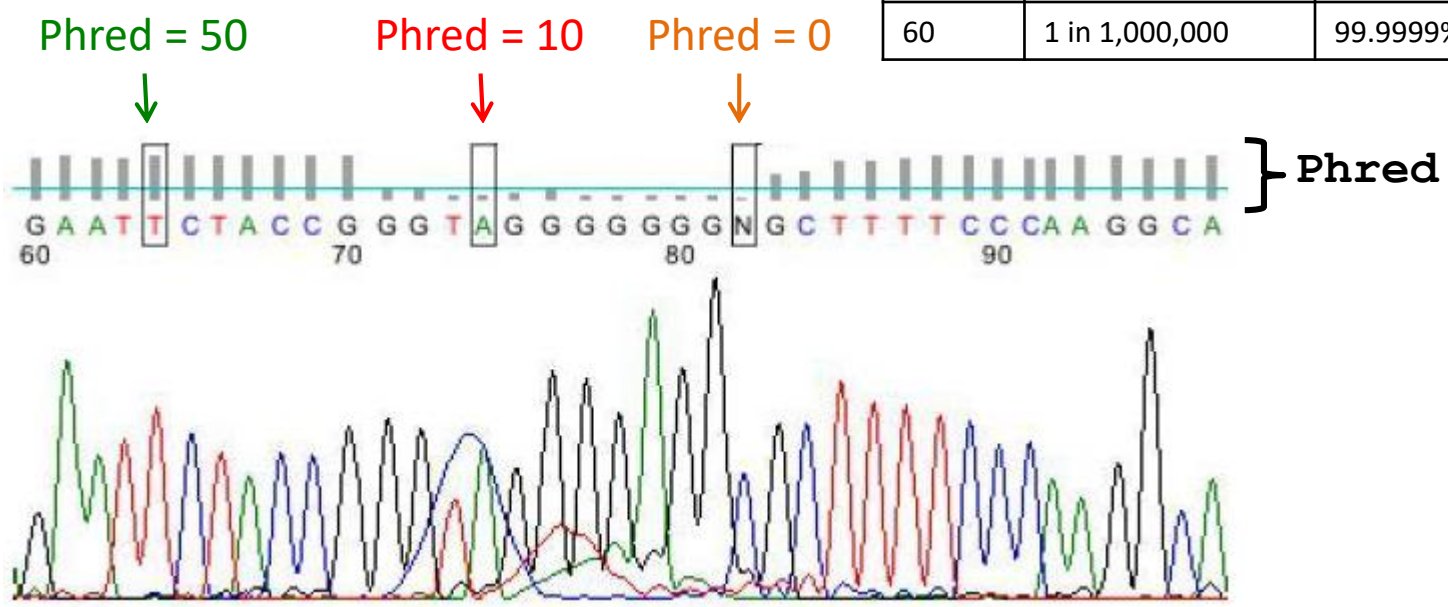
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                              |    |         |
0.......................26...31.......40
```

Sequencing quality score

Phred quality score:

$$Q = -10 \log_{10} P$$

Phred	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



Wikipedia

Read type and lengths

454 and IonTorrent

- Reads do not have fix lengths:

```
>G4S72XW01AM8OM rank=0000036 x=147.0 y=2340.0 length=89
TTCTCGACGATTCCTGCAGCAGATGATAGTTACATAGTCCAGGCAAGTTCTGCAAGCAGTTCAAAGCGGAGAGTTAAGCCGAATGAAGT
>G4S72XW01ALTYX rank=0000041 x=131.0 y=2151.0 length=86
CCGTCCACGATTCCTGCAGCAGATGATAGTTACATAGTCCAGGCAAGTTCTGCAAGCAGTTCAAGCGGAGAGTTAAGCCGAATCGA
>G4S72XW01AVHCV rank=0000065 x=241.0 y=1805.5 length=89
TTCTCGACGATTCCTGCAGCAGATGATAGTTACATAGTCCAGGCAAGTTCTGCAAGCAGTTCAAAGCGGAGAGTTAAGCCGAATACGTG
>G4S72XW01AQG7N rank=0000069 x=184.5 y=1809.0 length=65
CCGTCCATCTCCGTGTCCCGGCCCGTATCGCCTCCCTACTGTGCTTGAACACCCTGCGCTACGTG
>G4S72XW01ANE2G rank=0000071 x=149.5 y=2422.0 length=227
TTGCAAGCAGGTTGCTCAGGCCCACTTGGTCACTCTGTGCATTGCCCTTGGCAATCCGTGTGTTCCGTTTCCAATACCCCGGCCCCTCCTGCTCTATCCATGGC
GCTCGCGGCTCCATCCTCGGCTTCGGGGCGTCGCTGTCAAAGCGCACGAACTGCGTGTCCTCCACATAGCCCACTTCCATATGCCGGGGCTCCCCTCCGGGGCC
GGGACACGGAGGTACACTT
>G4S72XW01ALTSD rank=0000078 x=131.5 y=1915.0 length=260
TTCTCGGAGTGTCATTTCTCCAACGAGACGGAGCTGGTGCGGTTCCTGGAAAGATACATCTACAACCGGGAGGAGTACGTGCGCTTCGACAGCGACGTGGGGGA
GTACCGCGCGGTGAGCGAGCTGGGGCGGCGGTCAGCCGAGTACTGGAACAGACAGAAGGACCTCCTGGAGCAGAAGCGGGGACAGGTGGACAACTACTGCCGAC
ACAACTATGGGGTTGGTGAGAGCTTCACTGTGGAGCGGAGAGTTGACTGCTT
>G4S72XW01APU23 rank=0000079 x=177.5 y=1805.0 length=54
CCGCTCTCCGTGTCCCGGCCCTGAGCTATGTGCTTGAACACCCTGCGCGCTGGA
>G4S72XW01AL62B rank=0000091 x=135.5 y=2737.0 length=221
TTCTCGACCTCCGTGTCCCGGCCCGGAGGGAGCCCGGCATATGGAAGTGGGCTATGTGGAGGACACGCAGTTCGTGCGCTTTGACAGCGACGCCCCGAAGCCGA
GGATGGAGCCGCGAGCGCCATAGATAGAGCAGGGGGCCGGGTAGTTGAACGGAACACACGGATTGCCAAGGGCGAATGCACAGAGTGACCAAGTGGGCCTGAGC
AACCTGCGTTCCA
>G4S72XW01AR49R rank=0000093 x=203.0 y=1821.0 length=87
CCGTCCACGATTCCTGCAGCAGATGATAGTTACATAGTCCAGGCAAGTTCTGCAAGCAGTTCAAAGCGGAGAGTTAAGCCGAATCGA
>G4S72XW01AO8U8 rank=0000107 x=170.0 y=1682.0 length=229
TTGCAAGCGCAGGGTGTTCAAGCACATTCGGTGATTCTGTGCAGAGTTCTCGGAAATCCGTGTGTTTTGCTCCCAATACTCCGGACCCTCCAGCCCCATCCACG
GCGCCCGCGGCTCCTGTCTGGGATTCTTTGCGTCGCTGTCGAAGCGCATGAACTGGGTGTCGTCCACGTAGCCCACGGAGATGAAGCGGGACTCCCGGAGGCCG
GGCCGGGACACGGAGATGTAG
```

Read type and lengths

Illumina

- Reads are fix length but usually are paired (two files):

R1 file:

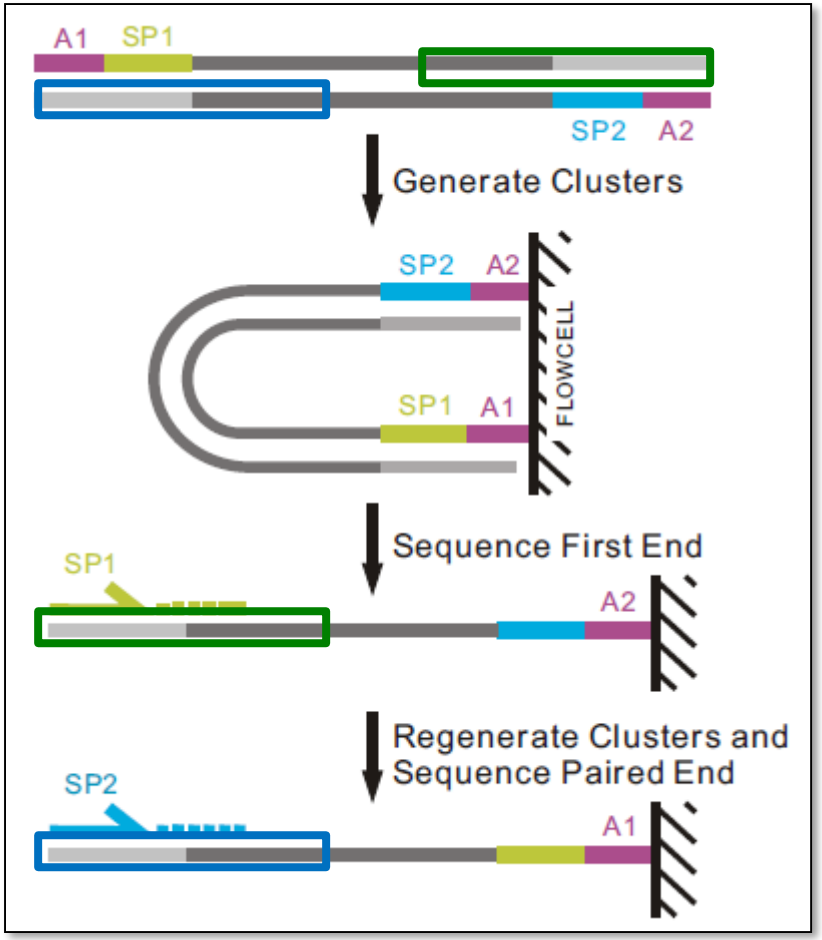
```
>M01530:20:000000000-A89BL:1:1101:14684:1732 1:N:0:1
CTTGGTGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGCTCCTACACAGATACATCTACAACCGGGAGGAGGTCTCGC
>M01530:20:000000000-A89BL:1:1101:18776:1733 1:N:0:1
TCAGAGGAGCGGGGGTCTCCACACCATACAATATGTTACCGGCTGTGACCTCCTGTCCGACGGGAGCGTCCGTGGATCCTT
>M01530:20:000000000-A89BL:1:1101:15484:1734 1:N:0:1
GGGGACGTCGGCATGTTCTGGTTCCGATAAACAGACACTATTTACCGCCGCTACCGCGGGCAACGGCGACAACTTCACCA
>M01530:20:000000000-A89BL:1:1101:18291:1819 1:N:0:1
AGTGTTGAGTGTCATTTCTCCAACGGGACGGAGCGGATACGGTTCCTGGACAGATACTTCTACAACCGGGAGGAGTACGTGCG
```

R2 file:

```
>M01530:20:000000000-A89BL:1:1101:14684:1732 2:N:0:1
TTGAGTTCACCTCTCCGCTCCACAGTGAAGCTCTCGACAACCCCATAGTTGTGTCTGCACACAGTGTCCACCTCGGCCCGC
>M01530:20:000000000-A89BL:1:1101:18776:1733 2:N:0:1
AACCGATGCGCTCCAGCTCCTTCTGCCCGTATCCGACGTATTTCTGGAGCTCTTCCGGGCACTCGTGCTTCAGGTAATTCG
>M01530:20:000000000-A89BL:1:1101:15484:1734 2:N:0:1
TCAGCAGTTTAATCACTGTTGCACTGGTCAACACTGGAATGGCGAGGCGCTGTACTTCTTCCAACAGCACTTTCACCATTAA
>M01530:20:000000000-A89BL:1:1101:18291:1819 2:N:0:1
GAACTATCACCTCTCCGCTCCACAGTGAAGCTCTCAACAACCCCGTAGTTGTGTCGGCAGTAGTTGTCCACCGTGGCCCGC
```

Read type and lengths

Illumina



Read type and lengths

Illumina

- And read ends may overlap:

R1 file / R2 file

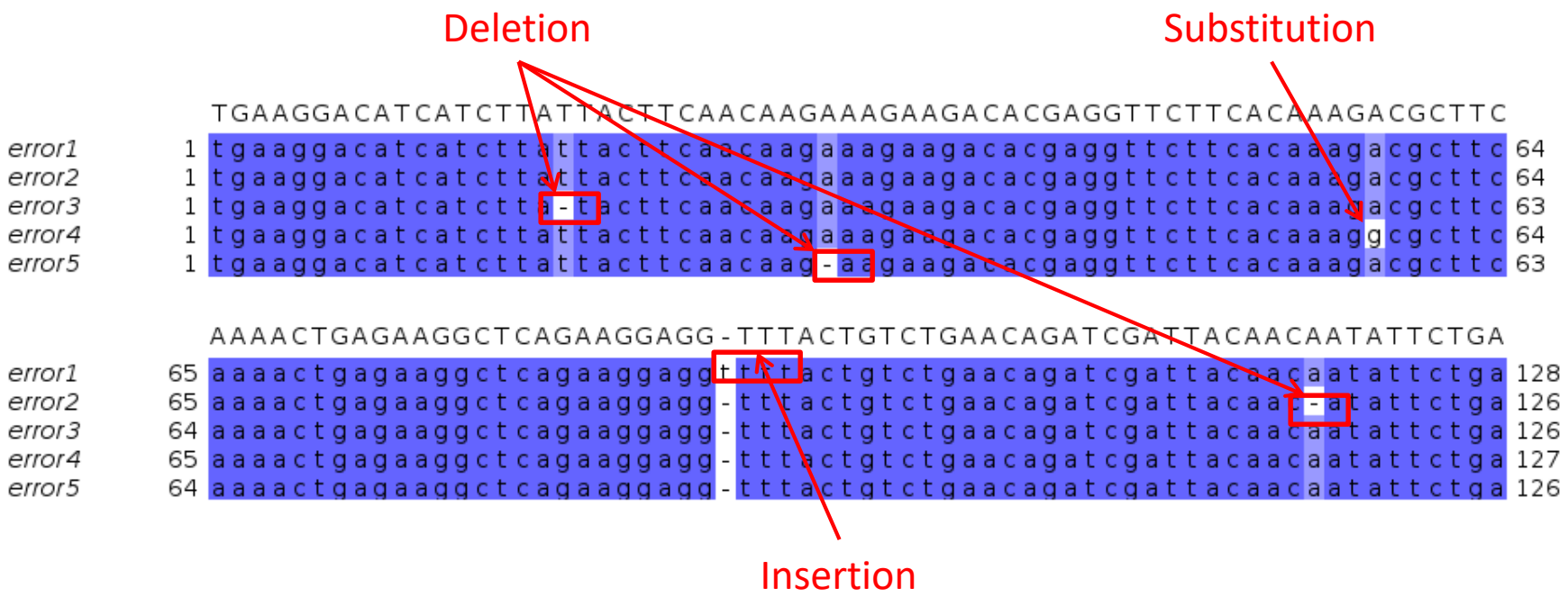
Reverse complementary

```
>M01530:20:000000000-A89BL:1:1101:14684:1732 1:N:0:1
CTTGGTGAGTGTCATTTCTCCAACGGGACGGAGCGGGTGCGGCTCCTACACAGATACATCTACAACCGGGAGGAGGTCTCGC
          >M01530:20:000000000-A89BL:1:1101:14684:1732 2:N:0:1
          GTCATTTCTCCAACGGGACGGAGCGGGTGCGGCTCCTACACAGATACATCTACAACCGGGAGGAGGTCTCGCCGTACGGCTA

>M01530:20:000000000-A89BL:1:1101:18776:1733 1:N:0:1
TCAGAGGAGCGGGGGTCTCCACACCATACAATATGTTACCGGCTGTGACCTCCTGTCCGACGGGAGCGTCCGTGGATCCTT
             >M01530:20:000000000-A89BL:1:1101:18776:1733 2:N:0:1
             GGTCTCCACACCATACAATATGTTACCGGCTGTGACCTCCTGTCCGACGGGAGCGTCCGTGGATCCTTCTAAGCCTCAGGCC

>M01530:20:000000000-A89BL:1:1101:15484:1734 1:N:0:1
GGGGACGTCGGCATGTTCTGGTTCCGATAAACAGACACTATTTACCGCCGCTACCGCGGGCAACGGCGACAACTTCACCA
      >M01530:20:000000000-A89BL:1:1101:15484:1734 2:N:0:1
      GTCGGCATGTTCTGGTTCCGATAAACAGACACTATTTACCGCCGCTACCGCGGGCAACGGCGACAACTTCACCACAGCGG

>M01530:20:000000000-A89BL:1:1101:18291:1819 1:N:0:1
AGTGTTGAGTGTCATTTCTCCAACGGGACGGAGCGGATACGGTTCCTGGACAGATACTTCTACAACCGGGAGGAGTACGTGCG
>M01530:20:000000000-A89BL:1:1101:18291:1819 2:N:0:1
         TGTCATTTCTCCAACGGGACGGAGCGGATACGGTTCCTGGACAGATACTTCTACAACCGGGAGGAGTACGTGCGATTACCGCT
```

Sequencing errors

454 and IonTorrent

- 1% of sequencing errors, mostly indels in homopolymer regions.



Sometimes there are more reads with errors than without!!!!!

Sequencing errors

Illumina

- <1% of sequencing errors, mostly random substitutions.

Substitution

```
                CTCTCCATGTATTACAACAAGCTGGAATACGCCAGGTTTGACAGCAACGTGGGTAAATATGT
error1      1   ctctccatgtattacaacaagctggaatacgccaggtttgacggcaacgtgggtaaatatgt  62
error2      1   ctctccatgtattacaacaagccggaatacgccaggtttgacagcaacgtgggtaaatatgt  62
error3      1   ctctccatgtattacaacaagctggaatacgccaggtttggcagcaacgtgggtaaatatgt  62
error4      1   ctctccatgtattacaacaagctggaatacgccaggtttgacagcaacgtgggtaaatatgt  62
error5      1   ctctccatgtattacaacaagctggaatacgccaggtttgacagcaacgtgggtaaatatgt  62

                TGGATACACGACGTATGGAGTGAAGAACGCTGAACGCTGGAACAAAGACACGTCAGAGATCG
error1     63   tggatacacgacgtatggagtgaagaacgctgaacgctggaacaaagacacgtcagagatcg 124
error2     63   tggatacacgacgtatggagtgaagaacgctgaacgctggaacaaagacacgtcagagatcg 124
error3     63   tggatacacgacgtatggagtgaagaacgctgaacgctggaacaaagacacgtcagagatcg 124
error4     63   tgggtacacgacgtatggagtgaagaacgctgaacgctggaacaaagacacgtcagagatcg 124
error5     63   tggatacacgacgtatggagtgaggaacgctgaacgctggaacaaagacacgtcagagatcg 124
```

As errors are random, the consensus sequence will be correct.

Other errors

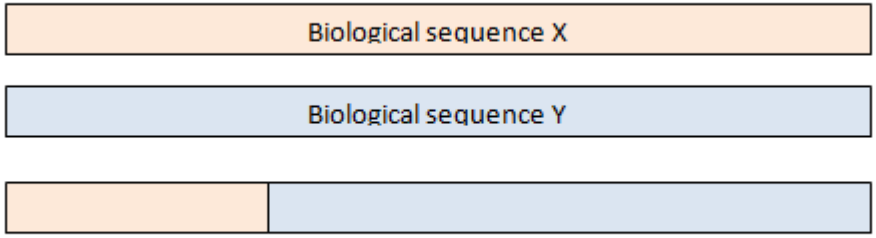
➢ PCR errors

Most commercially available Taq polymerases introduce errors at the rate of 1 point mutation every 1000 nts.

Solution: higher fidelity polymerases such as Pfu or Phusion High-Fidelity generating 10-100 times fewer errors respectively.

➢ Chimeras

Chimeras are sequences formed from two or more biological sequences joined together.



Chimera formed from X and Y

Solutions: - Reduce the number of PCR cycles.

- Increase the annealing temperature.

http://drive5.com/usearch/manual/chimera_formation.html

Error correction strategies

- **Filtering**: removes suspicious reads

  Problem: we can lost most of the reads, and with them most of the information.

  Also we can discard correct reads by error.

- **Clustering**: corrects erroneous reads

  Problem: it can be hard to discriminate among erroneous reads and correct ones.

- Experimental design
- Sampling
- Sample processing
- Sequencing
- Sequence processing

Operational taxonomic unit (OTU)

- **Theoretically, an OTU is a taxonomic level** of sampling selected by the user to be used in a study, such as individuals, populations, species, genera, or bacterial strains (Sokal and Sneath, 1963).

- **Practically, an OTU is a cluster of similar sequence variants** of the barcode (16S, ITS, etc.). Each of these cluster is intended to represent a taxonomic unit of a bacteria species or genus depending on the sequence similarity threshold.

An OTU cluster is usually defined by variants with a 97% of sequence identity.

Stackebrandt and Goebel (1994)

BUT…

- Some species have genes that are >97% similar, giving merged OTUs containing multiple species.
- A single species may have paralogs that are <97% similar, causing the species to be split across two or more OTUs.
- Some clusters, even a majority, may be spurious due to artifacts including read errors and chimeras.

http://www.drive5.com/usearch/manual/otu_definition.html

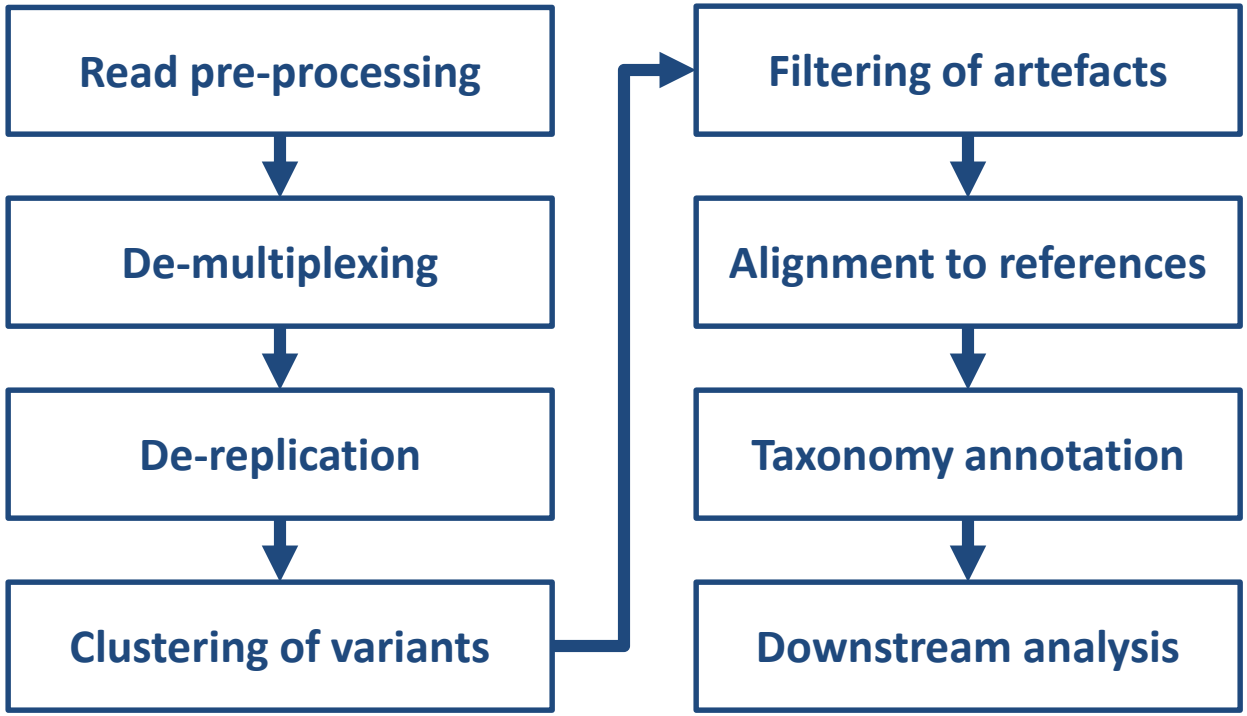
Pros and Cons

OTUs	Taxonomy
Novel organisms	Universal names
Insufficient taxonomy	Meaning associated with names
Does not lump together all order or family-level classifications	Independent of clustering width and algorithm
Many names based on phenotype rather than genotype	Historically well-studied are split New areas are lumped

Susan Huse. STAMPS 2016 Course. Introduction to OTU Clustering.

Analysis pipeline

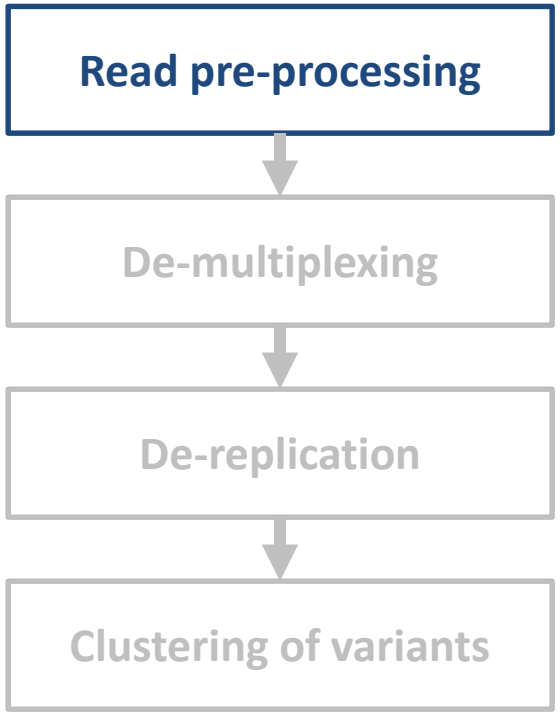
These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR, MICCA and AmpliTAXO



Oulas,A. *et al.* (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights*, **9**, 75–88.

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR, MICCA and AmpliTAXO



1. Read pre-processing

If reads are paired-end type (e.g. Illumina), an initial step consists of merging overlapping paired reads into single reads is required.

Anomalous reads are removed and when reads have different lengths (e.g. 454) they are also trimmed to a fix length to make easier further processing (alignment and clustering).

2. De-multiplexing

Organizes the multiplexed reads into amplicons (single PCR products) based on the different barcodes (primers) and tags (samples) used.

3. De-replication

Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

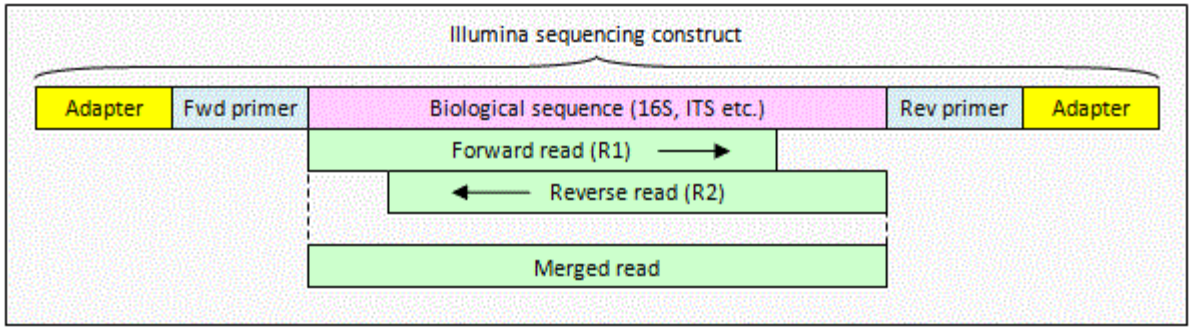
4. Clustering of variants

Variants are clustered based on a user-defined similarity threshold. This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

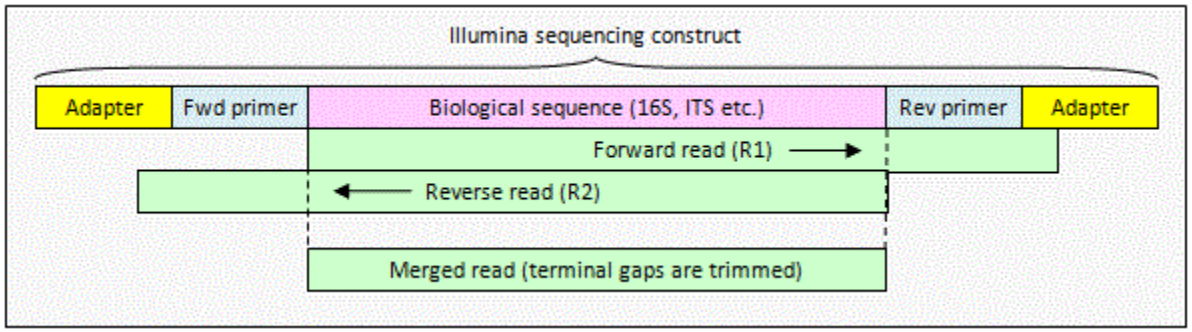
Read pre-processing

If reads are paired-end type (e.g. Illumina), an initial step consists of merging overlapping paired reads into single reads is required.

- Illumina paired read with overlap:



- Illumina paired reads with staggered overlap:



http://drive5.com/usearch/manual/read_layout.html

Read pre-processing

Detection and removal of suspicious reads.

Primer: CCGTCAATTCMTTTRA
Barcode: AATGGTAC

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

Read pre-processing

Detection and removal of suspicious reads.

Primer: CCGTCAATTCMTTTRA
Barcode: AATGGTAC

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

Read pre-processing

When reads have different lengths they are trimmed to a fix length to make easier further processing (alignment and clustering).

Primer: CCGTCAATTCMTTTRA
Barcode: AATGGTAC

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

Read pre-processing

When reads have different lengths they are trimmed to a fix length to make easier further processing (alignment and clustering).

Primer: CCGTCAATTCMTTTRA
Barcode: AATGGTAC

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

Read pre-processing

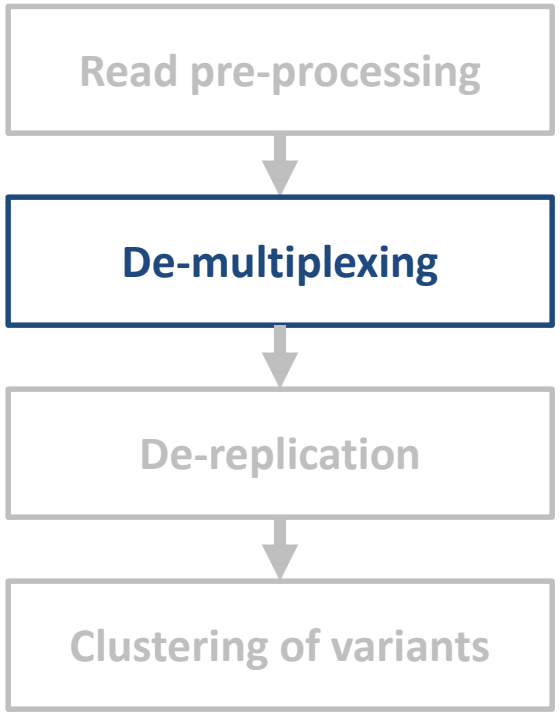
When reads have different lengths they are trimmed to a fix length to make easier further processing (alignment and clustering).

Primer: CCGTCAATTCMTTTRA
Barcode: AATGGTAC

```
>GQY1XT001A6MUA
AATGGTACCCGTCAATTCATTTGATCTTGCGGTTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTCATA
>GQY1XT001BTRWS
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCGTTTACGGCGTGGACTACCAGTCGCACTCGAGCTGCA
>GQY1XT001BBPBR
AATGGTACCCGTCAATTCATTTGACGTTGCCCCCCGTTTACTGTGCGGACTACCAGTCGCACTCAAGGCCCC
>GQY1XT001BDDE9
AATGGTACCCGTCAATTCCTTTAATCTTGCGGGTCGTTTACGGCGTGGACTACCAGTCGCACTCCAGTTACA
>GQY1XT001CIUF3
AATGGTACCCGTCAATTCCTTTGATCTTGCGGGCCTTTACGGCGTGGACTACCAGGCGCCCTCCAGCCCGGC
>GQY1XT001B44ZE
AATGGTACCCGTCAATTCATTTAACCTTGCGGGGTTTTACCGCGTGGACTACCAGGCGCCCTCAAGAAGAAC
>GQY1XT001CIW3P
AATGGTACCCGTCAATTCATTTGACGTTGCCTCTCGTTTACTGCGTGGACTACCAGTCGCACTCAAGGCCCC
>GQY1XT001A731D
AATGGTACCCGTCAATTCATTTAACGTTGCCCCCGTTACTGCGTGGACTACCAGGGGCAATCAAGACTGCCA
```

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR, MICCA and AmpliTAXO



1. Read pre-processing

   If reads are paired-end type (e.g. Illumina), an initial step consists of merging overlapping paired reads into single reads is required.

   Anomalous reads are removed and when reads have different lengths (e.g. 454) they are also trimmed to a fix length to make easier further processing (alignment and clustering).

2. De-multiplexing

   Organizes the multiplexed reads into amplicons (single PCR products) based on the different barcodes (primers) and tags (samples) used.

3. De-replication

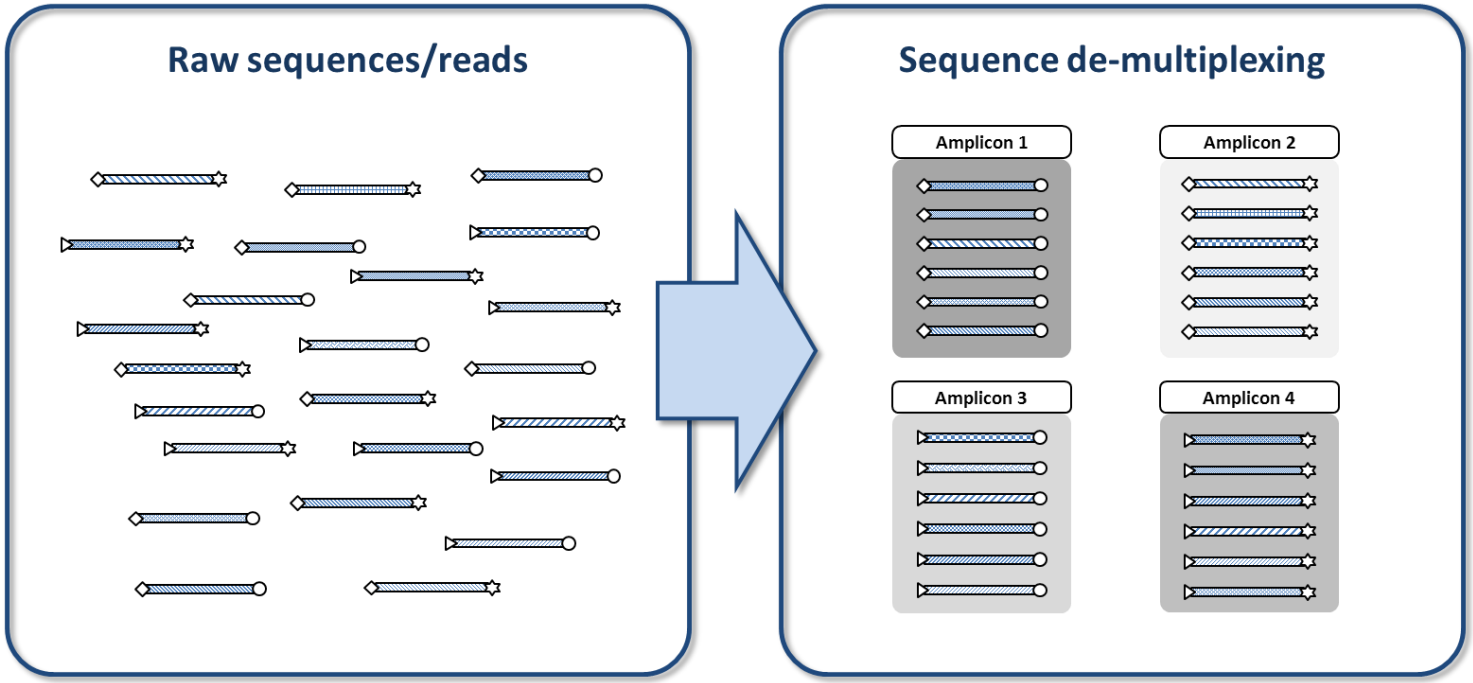
   Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

4. Clustering of variants

   Variants are clustered based on a user-defined similarity threshold. This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

De-multiplexing

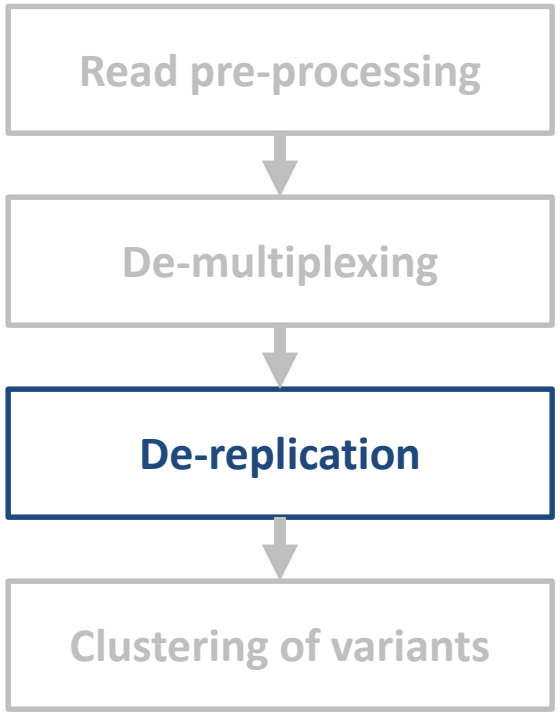
Organizes the multiplexed reads into amplicons (single PCR products) based on the different barcodes (primers) and tags (samples) used.



1 Amplicon = 1 Sample

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR, MICCA and AmpliTAXO



1. Read pre-processing

   If reads are paired-end type (e.g. Illumina), an initial step consists of merging overlapping paired reads into single reads is required.

   Anomalous reads are removed and when reads have different lengths (e.g. 454) they are also trimmed to a fix length to make easier further processing (alignment and clustering).

2. De-multiplexing

   Organizes the multiplexed reads into amplicons (single PCR products) based on the different barcodes (primers) and tags (samples) used.

3. De-replication

   Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

4. Clustering of variants

   Variants are clustered based on a user-defined similarity threshold. This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

De-replication

Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

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

De-replication

Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

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

De-replication

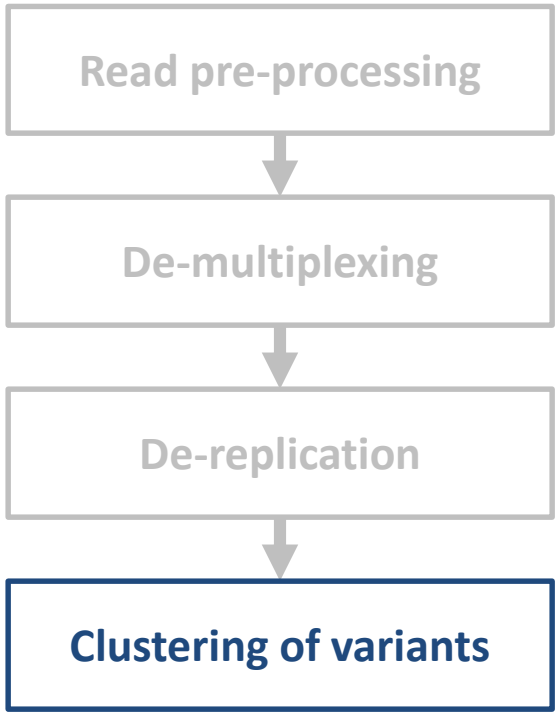
Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

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

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

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR, MICCA and AmpliTAXO



1. Read pre-processing

   If reads are paired-end type (e.g. Illumina), an initial step consists of merging overlapping paired reads into single reads is required.

   Anomalous reads are removed and when reads have different lengths (e.g. 454) they are also trimmed to a fix length to make easier further processing (alignment and clustering).

2. De-multiplexing

   Organizes the multiplexed reads into amplicons (single PCR products) based on the different barcodes (primers) and tags (samples) used.

3. De-replication

   Redundant reads are annotated as unique sequences (variants) and their abundances (depths).

4. Clustering of variants

   Variants are clustered based on a user-defined similarity threshold. This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

Clustering of variants

Variants are clustered based on a user-defined similarity threshold.

This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

```
>*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
```

Clustering of variants

Variants are clustered based on a user-defined similarity threshold.

This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

```
>*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
```

Clustering of variants

Variants are clustered based on a user-defined similarity threshold.

This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

```
>*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
```

Clustering of variants

Variants are clustered based on a user-defined similarity threshold.

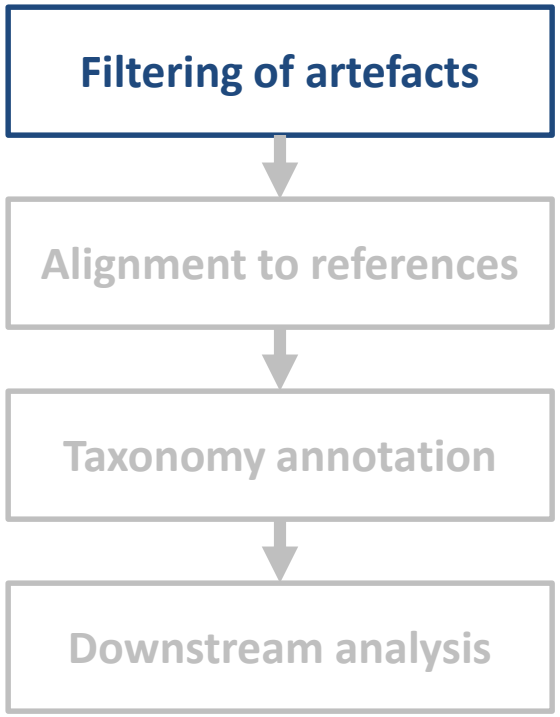
This step is crucial to group redundant sequences due to sequencing and PCR errors into unique variants that will be representative of single OTUs.

```
>*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
```

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR and AmpliTAXO



5. Filtering of artefacts

Detection and removal of artefactual variants left after clustering: chimeras, contaminants, PCR errors…

6. Alignment to references

Clustered variants (OTUs) are aligned against a database of reference sequences, e.g. Greengenes, SILVA…

7. Taxonomy annotation

Taxonomy annotations from databases will be assigned to OTUs. In an ideal scenario, each OTU will correspond to a unique species taxonomy assignment.

8. Downstream analysis

OTU table and taxonomy results can be subject of further analyses: alpha diversity measurements and rarefaction plots, beta diversity and ordination plots, taxonomy heatmaps…

Filtering of artefacts

Detection and removal of artefactual variants left after clustering: chimeras, contaminants…

```
>*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
GTTCTTCCTAATATCTACGCATTTCACCGCTACACTAGGAATTCCACTTTCCTCTCCGATACTCTAGATTGG
>#16S-0000058 | depth=12 | freq=0.66
TTCAGTCGCTCCCCTAGCTTTCGCACTTCAGCGTCAGTTGCCGTAAGCCAGAGAGCCGCTTTCGCCACCGGT
GTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTCTTGCACTCAAGTTAAA
>*16S-0000098 | depth=10 | freq=0.55
TTTAGTCCTGTTCGCTCCCCACGCTTTCGCTCCTCAGCGTCAGTAACGGCCCAGAGACCCGCCTTCGCCACC
GGTGTTCTTCCTGATATCTGCGCATTCCACCGCTACACCAGGAGTTCCAGCCTCCCCTACCGCACTCAAGCC
>#16S-0000295 | depth=2 | freq=0.11
TTCACGATACCCACGCTTTCGAGCATCAGCGTCAGTTGCGCTACAGTAAGCTGCCTTCGCAATCGGAGTTCT
TCGTGATATCTAAGCATTTCACCGCTACACCACGAATTCCGCCTACTTTCGGCGCACTCAAGCCCCCCAGTT
>#16S-0000021 | depth=1 | freq=0.06
TTCAACGTTCGCTCCCCTGGCTTTCGCGCCTCAGCGTCAGTTTTCGTCCAGAAAGTCGCCTTCGCCACTGGT
GTTCTTCCTAATATCTACGCATTTCACCGCTACACTAGGAATTCCACTTTCCTCTCCGATACTCTAGATCAG
```

Filtering of artefacts

Detection and removal of artefactual variants left after clustering: chimeras, contaminants…

```
>*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
GTTCTTCCTAATATCTACGCATTTCACCGCTACACTAGGAATTCCACTTTCCTCTCCGATACTCTAGATTGG
>#16S-0000058 | depth=12 | freq=0.66
TTCAGTCGCTCCCCTAGCTTTCGCACTTCAGCGTCAGTTGCCGTAAGCCAGAGAGCCGCTTTCGCCACCGGT
GTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTCTTGCACTCAAGTTAAA
>*16S-0000098 | depth=10 | freq=0.55
TTTAGTCCTGTTCGCTCCCCACGCTTTCGCTCCTCAGCGTCAGTAACGGCCCAGAGACCCGCCTTCGCCACC
GGTGTTCTTCCTGATATCTGCGCATTCCACCGCTACACCAGGAGTTCCAGCCTCCCCTACCGCACTCAAGCC
>#16S-0000295 | depth=2 | freq=0.11
TTCACGATACCCACGCTTTCGAGCATCAGCGTCAGTTGCGCTACAGTAAGCTGCCTTCGCAATCGGAGTTCT
TCGTGATATCTAAGCATTTCACCGCTACACCACGAATTCCGCCTACTTTCGGCGCACTCAAGCCCCCCAGTT
>#16S-0000021 | depth=1 | freq=0.06
TTCAACGTTCGCTCCCCTGGCTTTCGCGCCTCAGCGTCAGTTTTCGTCCAGAAAGTCGCCTTCGCCACTGGT
GTTCTTCCTAATATCTACGCATTTCACCGCTACACTAGGAATTCCACTTTCCTCTCCGATACTCTAGATCAG
```

Filtering of artefacts

Detection and removal of artefactual variants left after clustering: chimeras, contaminants…

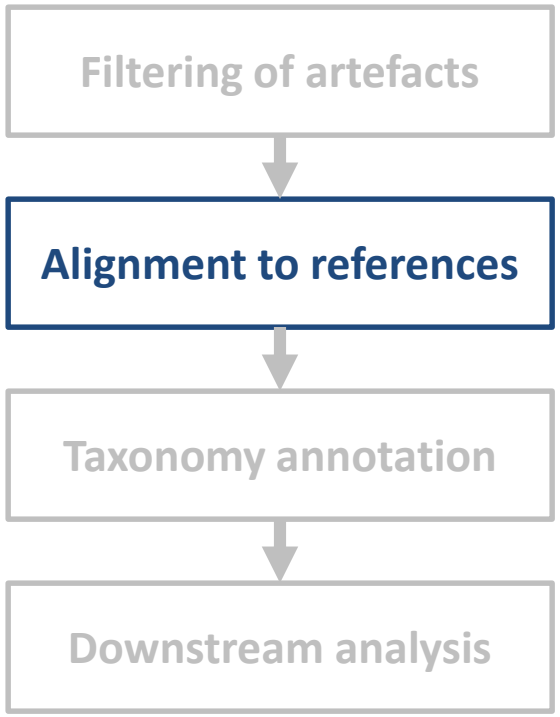
```
>*16S-0000011 | depth=44 | freq=2.42
TTCAGTCGCTCCCCTAGCTTTCGCACTTCAGCGTCAGTTGCCGTCCAGTGAACTATCTTCATCATCGGCATT
CCTGCACATATCTACGAATTTCACCTCTACTCGTGCAGTTCCGTCCACCTCTCCAGCACTCTAGCCAAACAG
>*16S-0000076 | depth=33 | freq=1.82
TTCAATGTTTGCTCCCCACGCTTTCGAGCCTCAGCGTCAGTTACAAGCCAGAGAGCCGCTTTCGCCACCGGT
GTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTCTTGCACTCAAGTTAAA
>*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
TCGTGATATCTAAGCATTTCACCGCTACAGCACGAATTCCGCCTAGTTTCGGCGCACTCAAGCCCCCCAGTT
>#16S-0000021 | depth=1 | freq=0.06
TTCAACGTTCGCTCCCCTGGCTTTCGCGCCTCAGCGTCAGTTTTCGTCCAGAAAGTCGCCTTCGCCACTGGT
GTTCTTCCTAATATCTACGCATTTCACCGCTACACTAGGAATTCCACTTTCCTCTCCGATACTCTAGATCAG
```

Chimera

Contaminations

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR and AmpliTAXO



5. Filtering of artefacts

   Detection and removal of artefactual variants left after clustering: chimeras, contaminants, PCR errors…

6. Alignment to references

   Clustered variants (OTUs) are aligned against a database of reference sequences, e.g. Greengenes, SILVA…

7. Taxonomy annotation

   Taxonomy annotations from databases will be assigned to OTUs. In an ideal scenario, each OTU will correspond to a unique species taxonomy assignment.

8. Downstream analysis

   OTU table and taxonomy results can be subject of further analyses: alpha diversity measurements and rarefaction plots, beta diversity and ordination plots, taxonomy heatmaps…

Alignment to references

After clustering and filtering variants, the retrieved OTUs are aligned against a database of reference sequences, e.g. Greengenes, SILVA…

```
>*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

Alignment to references

AY053482.1

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

```
Query  1    TTCAACCTTGCGGTCGTACTCCCCAGGCGGAGTGCTTAATGCGTTAGCTGCGGCACTAAA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  882  TTCAACCTTGCGGTCGTACTCCCCAGGCGGAGTGCTTAATGCGTTAGCTGCGGCACTAAA  823

Query  61   CCCCGGAAAGGGTCTAACACCTAGCACTCATCGTTTACGGCGTGGACTACCAGGGTATCT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  822  CCCCGGAAAGGGTCTAACACCTAGCACTCATCGTTTACGGCGTGGACTACCAGGGTATCT  763

Query  121  AATCCTGTTTGCTCCCCACGCTTTCGAGCCTCAGCGTCAGTTACAAGCCAGAGAGCCGCT  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  762  AATCCTGTTTGCTCCCCACGCTTTCGAGCCTCAGCGTCAGTTACAAGCCAGAGAGCCGCT  703

Query  181  TTCGCCACCGGTGTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACT  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  702  TTCGCCACCGGTGTTCCTCCATATATCTACGCATTTCACCGCTACACATGGAATTCCACT  643

Query  241  CTCCCCTCTTGCACTCAAGTTAAACAGTTTCCAAAGCGTACTATGGTTAAGCCACAGCCT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  642  CTCCCCTCTTGCACTCAAGTTAAACAGTTTCCAAAGCGTACTATGGTTAAGCCACAGCCT  583

Query  301  TTAACTTCAGACTTATCT  318
            ||||||||||||||||||
Sbjct  582  TTAACTTCAGACTTATCT  565
```

Alignment to references

CP001685.1

Sequence ID: lcl|Query_210571 **Length:** 1510 **Number of Matches:** 1

Range 1: 560 to 872 Graphics

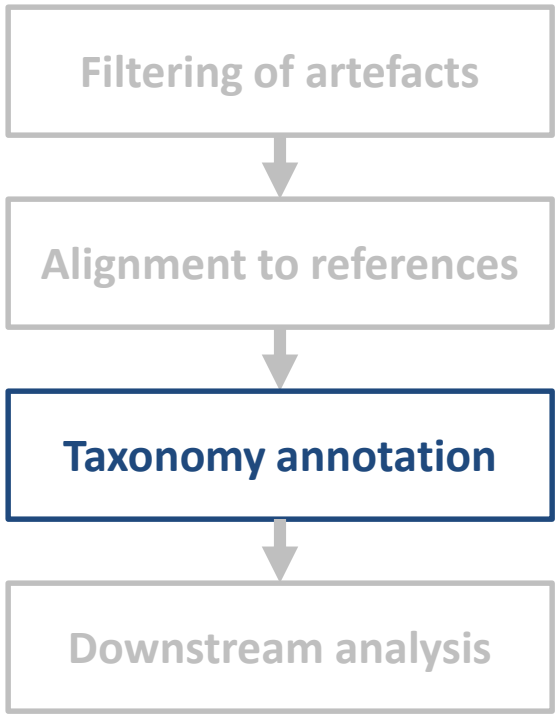
Score	Expect	Identities	Gaps	Strand
490 bits(265)	2e-142	297/313(95%)	0/313(0%)	Plus/Minus

```
Query  1    TTCAGCCTTGCGGCCGTACTCCCCAGGCGGATTACTTATCGCATTCGCTTCGGCACAGAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||| |||||||||| |||
Sbjct  872  TTCAGCCTTGCGGCCGTACTCCCCAGGCGGATTACTTATCGCATTAGCTTCGGCACGGAC  813

Query  61   AGTCTTCCTGCCCACACCCAGTAATCATCGTTTACGGCCGGGACTACCAGGGTATCTAAT  120
            | ||||                                                |||||||
Sbjct  812  ACTCTT                                             ATCTAAT  753

Query  121  CCTGTT                                             TATCTTC  180
            ||||||                                              |||||||
Sbjct  752  CCTGTT                                             TATCTTC  693

Query  181  ATCATCGGCATTCCTGCACATATCTACGAATTTCACCTCTACTCGTGCAGTTCCGTCCAC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  692  ATCATCGGCATTCCTGCACATATCTACGAATTTCACCTCTACTCGTGCAGTTCCGTCCAC  633

Query  241  CTCTCCGGTACTCCAGCCTATCAGTTTCAAAGGCAGGCCTGCGGTTGAGCCGCAGGTTTT  300
            |||||| | |||| |||| | ||||||| | ||||||| |||||||||||||||| |||||
Sbjct  632  CTCTCCAGCACTCTAGCCAAACAGTTTCCAGGGCAGGCTTGCGGTTGAGCCGCAAGTTTT  573

Query  301  CACCCCTGACTTG  313
            |||||| ||||||
Sbjct  572  CACCCCAGACTTG  560
```

Around 95-97% of identity is required in the alignment of an OTU sequence to a database reference

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR and AmpliTAXO



5. Filtering of artefacts

   Detection and removal of artefactual variants left after clustering: chimeras, contaminants, PCR errors…

6. Alignment to references

   Clustered variants (OTUs) are aligned against a database of reference sequences, e.g. Greengenes, SILVA…

7. Taxonomy annotation

   Taxonomy annotations from databases will be assigned to OTUs. In an ideal scenario, each OTU will correspond to a unique species taxonomy assignment.

8. Downstream analysis

   OTU table and taxonomy results can be subject of further analyses: alpha diversity measurements and rarefaction plots, beta diversity and ordination plots, taxonomy heatmaps…

Taxonomy annotation

Taxonomy annotations from databases will be assigned to OTU sequences.

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

CP001685.1;tax=k:Bacteria,p:Fusobacteria,c:Fusobacteria (class),o:Fusobacteriales,
f:Fusobacteriaceae,g:Leptotrichia,s:buccalis

Sequence ID: lcl|Query_210571 **Length:** 1510 **Number of Matches:** 1

Range 1: 560 to 872 Graphics

Score	Expect	Identities	Gaps	Strand
490 bits(265)	2e-142	297/313(95%)	0/313(0%)	Plus/Minus

Taxonomy annotation

In an ideal scenario, each OTU sequence will have a taxonomy assignment.

OTU representative sequences

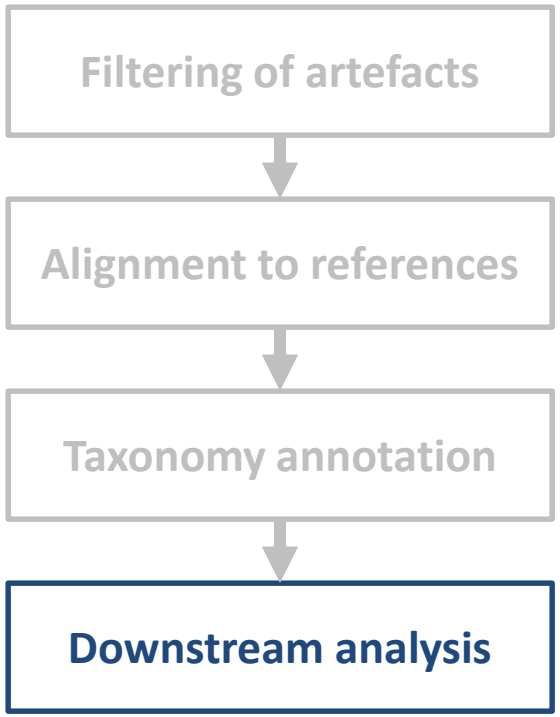
OTU taxonomy assignments

Samples and OTU frequencies

			COUNT_OTUS	19	14	13	15	14
SEQUENCES	MEAN_FR	SAMPLES	OTU	SRS052681	SRS042606	SRS042483	SRS048589	SRS051454
16S: TTCAACCTTGCGGTCG	0.0393	5	tax=k:Bacteria,p:Firmicutes,c:Bacilli,o:Lactobacillales,f:Streptococcaceae,g:Streptococcus,s:pseudopneumoniae;	0.0061	0.0893	0.0094	0.0089	0.083
16S: TTCATACTTGCGTACG	0.0707	5	tax=k:Bacteria,p:Fusobacteria,c:Fusobacteria (class),o:Fusobacteriales,f:Fusobacteriaceae,g:Fusobacterium;	0.0683	0.0918	0.0031	0.0529	0.1374
16S: TTCACCGTTGCCGGCG	0.0557	5	tax=k:Bacteria,p:Bacteroidetes,c:Bacteroidia,o:Bacteroidales,f:Porphyromonadaceae,g:clone,s:HF001;	0.0463	0.0494	0.0682	0.0575	0.057
16S: TTTAGCCTTGCGGCCG	0.0815	2	tax=k:Bacteria,p:Actinobacteria,c:Actinobacteria (class),o:Actinomycetales,f:Corynebacteriaceae,g:Corynebacterium,s:matruchotii;	0.0628			0.1001	
16S: TTTAATCTTGCGACCG	0.0291	5	tax=k:Bacteria,p:Proteobacteria,c:Betaproteobacteria,o:Neisseriales,f:Neisseriaceae,g:Neisseria;	0.0061	0.0963	0.0055	0.0065	0.0311
16S: TTCAACCTTGCGGTCG	0.0246	5	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Veillonellaceae,g:Veillonella;	0.0226	0.0374	0.0086	0.0168	0.0376
16S: TTCACCGTTGCCGGCG	0.0124	4	tax=k:Bacteria,p:Bacteroidetes,c:Bacteroidia,o:Bacteroidales;	0.0127	0.0089	0.0071		0.0207
16S: TTCAGCCTTGCGGCCG	0.0093	5	tax=k:Bacteria,p:Fusobacteria,c:Fusobacteria (class),o:Fusobacteriales,f:Fusobacteriaceae,g:Leptotrichia,s:buccalis;	0.0242	0.0038	0.0031	0.0098	0.0058
16S: TTTAGCCTTGCGGCCG	0.0127	3	tax=k:Bacteria,p:Actinobacteria,c:Actinobacteria (class),o:Actinomycetales,f:Actinomycetaceae,g:Actinomyces,s:odontolyticus;	0.0039	0.031	0.0031		
16S: TTCACCGTTGCCGGCG	0.008	4	tax=k:Bacteria,p:Bacteroidetes,c:Bacteroidia,o:Bacteroidales,f:Prevotellaceae,g:Prevotella;	0.0072	0.0127	0.0055		0.0065
16S: TTCAACCTTGCGGTCG	0.0141	2	tax=k:Bacteria,p:Firmicutes,c:Bacilli,o:Lactobacillales,f:Enterococcaceae,g:Enterococcus;		0.0177			0.0104
16S: TTCATTCTTGCGAACG	0.0093	3	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Lachnospiraceae;		0.0139		0.0042	0.0097
16S: TTCATTCTTGCGAACG	0.0086	3	tax=k:Bacteria,p:Actinobacteria;	0.0121		0.0063	0.0075	
16S: TTTAGCCTTGCGGCCG	0.0115	2	tax=k:Bacteria,p:Actinobacteria,c:Actinobacteria (class),o:Actinomycetales,f:Actinomycetaceae,g:Actinomyces,s:oris;	0.0182			0.0047	
16S: TTCACACTTGCGTGCG	0.0114	2	tax=k:Bacteria,p:Bacteroidetes,c:Flavobacteria,o:Flavobacteriales,f:Flavobacteriaceae,g:Capnocytophaga,s:sputigena;			0.0047		0.0181
16S: TTCATTCTTGCGAACG	0.0093	2	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Lachnospiraceae,g:Oribacterium,s:sp. oral taxon 078;		0.0095			0.0091
16S: TTCACTCTTGCGAGCG	0.006	3	tax=k:Bacteria,p:Bacteroidetes,c:Flavobacteria,o:Flavobacteriales,f:Flavobacteriaceae;	0.0033	0.0063		0.0084	
16S: TTCACCGTTGCCGGCG	0.0051	3	tax=k:Bacteria,p:Bacteroidetes,c:Bacteroidia,o:Bacteroidales,f:Prevotellaceae,g:Prevotella,s:nigrescens;	0.0066	0.0057	0.0031		
16S: TTCAGCCTTGCGGCCG	0.007	2	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Veillonellaceae,g:Selenomonas,s:noxia;	0.0083			0.0056	
16S: TTCAGTGTTGCCACCG	0.0055	2	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Clostridiales Family XI. Incertae Sedis,s:Parvimonas micra;	0.0077				0.0032
16S: TTCACCCTTGCGGGCA	0.0077	1	tax=k:Bacteria,p:Spirochaetes,c:Spirochaetes (class),o:Spirochaetales,f:Spirochaetaceae,g:Treponema,s:socranskii;	0.0077				
16S: TTTAATCTTGCGACCG	0.0075	1	tax=k:Bacteria,p:Proteobacteria,c:Betaproteobacteria,o:Burkholderiales,f:Burkholderiaceae;				0.0075	
16S: TTCAGTCTTGCGACCG	0.0061	1	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Veillonellaceae,g:Selenomonas;	0.0061				
16S: TTCAACCTTGCGGCCG	0.0061	1	tax=k:Bacteria,p:Proteobacteria,c:Betaproteobacteria,o:Burkholderiales,f:Comamonadaceae;				0.0061	
16S: TTCATTCTTGCGAACG	0.005	1	tax=k:Bacteria,p:Bacteroidetes,c:Bacteroidia,o:Bacteroidales,f:Porphyromonadaceae;	0.005				
16S: TTTAACCTTGCGGTCG	0.0039	1	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales;			0.0039		
16S: TTTATTCTTGCGAACG	0.0037	1	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Eubacteriaceae,g:Eubacterium;				0.0037	
16S: TTCATTCTTGCGAACG	0.0032	1	tax=k:Bacteria,p:Firmicutes,c:Clostridia,o:Clostridiales,f:Lachnospiraceae,g:Catonella;					0.0032

Analysis pipeline

These are the general steps shared by the most used metagenomics analysis tools: UPARSE, QIIME, MOTHUR and AmpliTAXO



5. Filtering of artefacts
   Detection and removal of artefactual variants left after clustering: chimeras, contaminants, PCR errors…

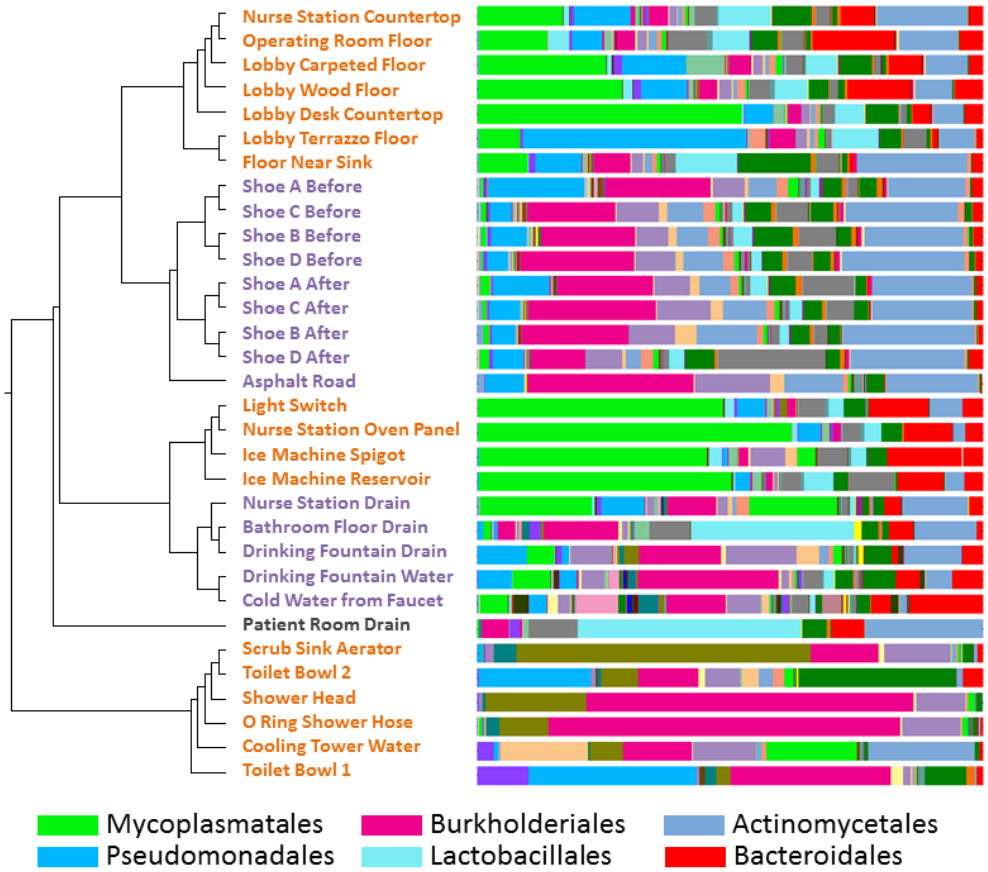
6. Alignment to references
   Clustered variants (OTUs) are aligned against a database of reference sequences, e.g. Greengenes, SILVA…

7. Taxonomy annotation
   Taxonomy annotations from databases will be assigned to OTUs. In an ideal scenario, each OTU will correspond to a unique species taxonomy assignment.

8. Downstream analysis
   OTU table and taxonomy results can be subject of further analyses: alpha diversity measurements and rarefaction plots, beta diversity and ordination plots, taxonomy heatmaps…

Downstream analysis

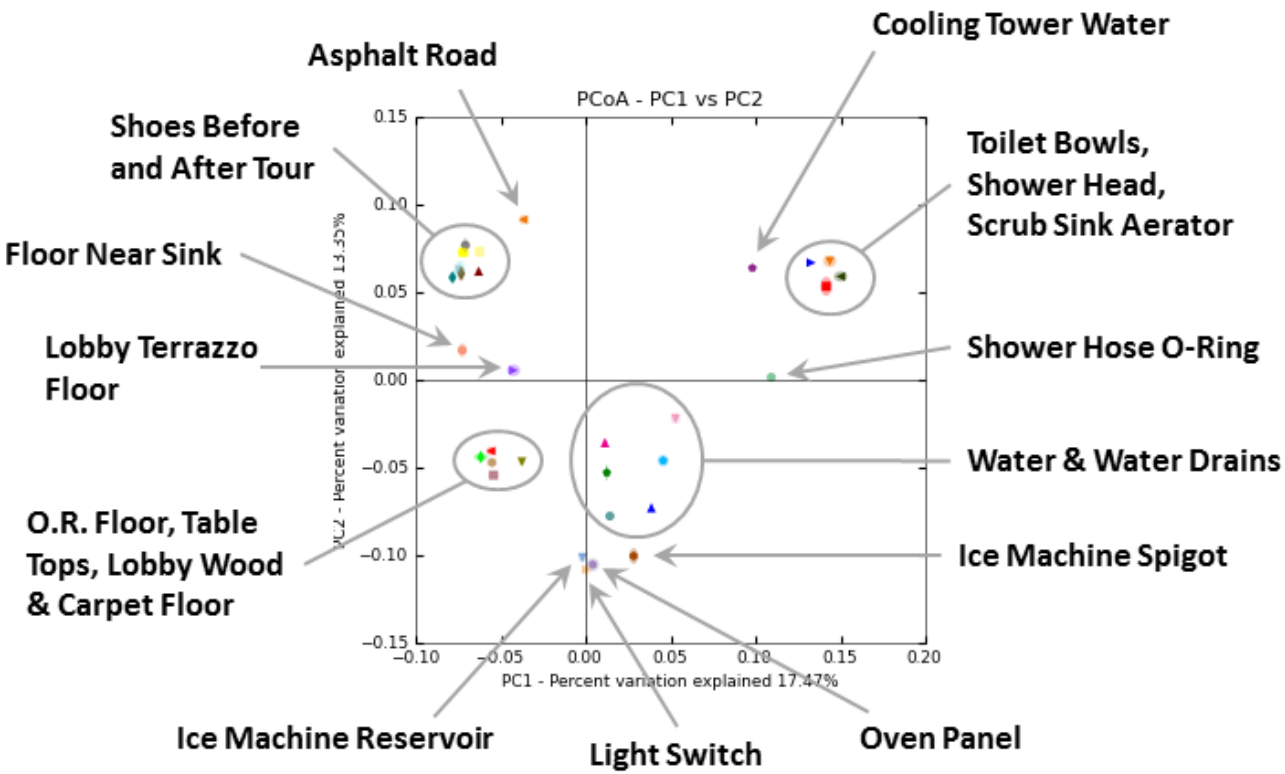
Taxonomy summaries:



http://hospitalmicrobiome.com/construction-samples/

Downstream analysis

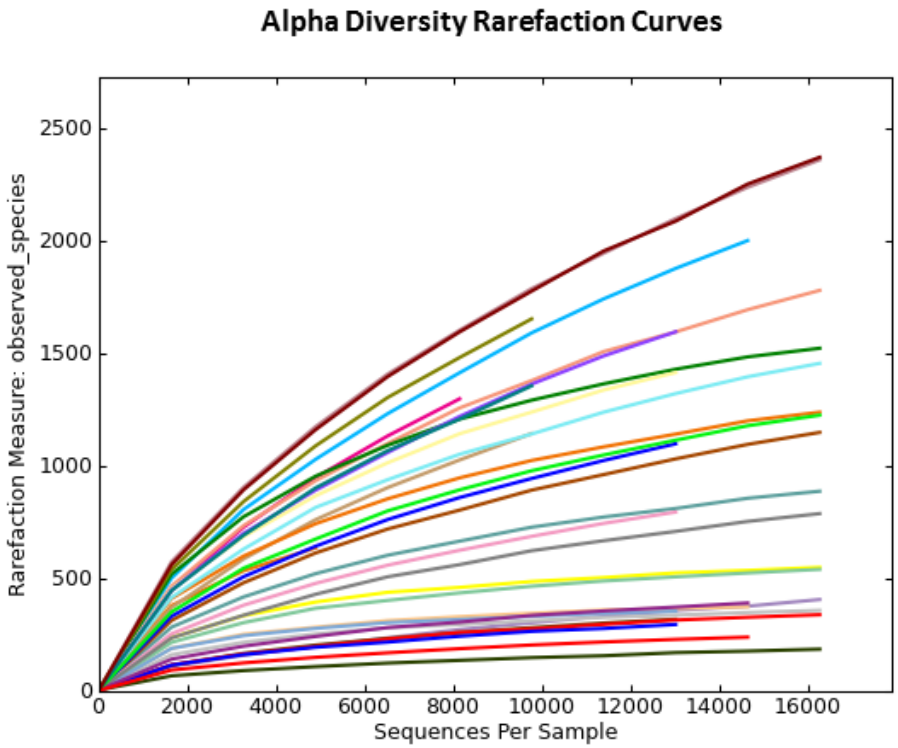
Principal Coordinate Analysis (PCoA):

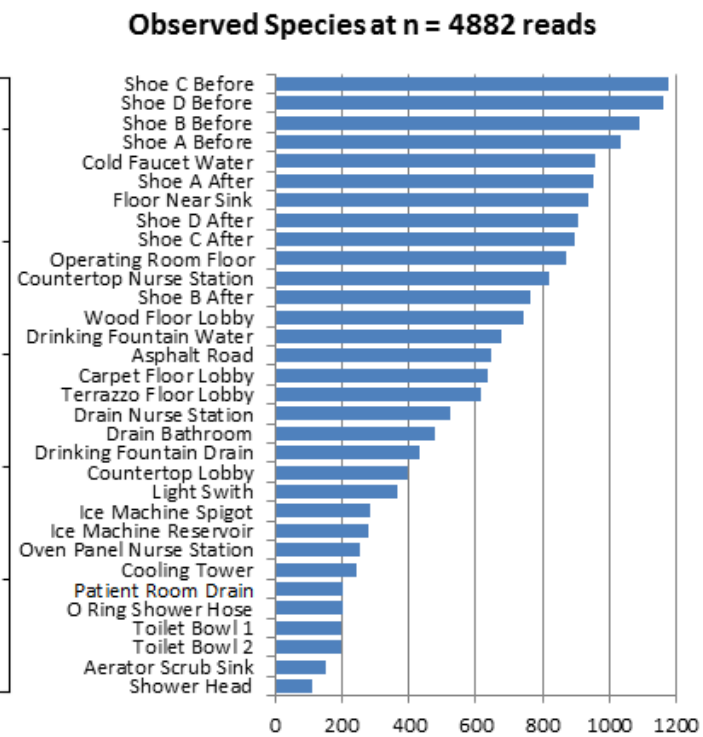


http://hospitalmicrobiome.com/construction-samples/

Downstream analysis

Alpha diversity measurements and rarefaction plots:





http://hospitalmicrobiome.com/construction-samples/

Interesting materials

- Materials from Strategies and Techniques for Analyzing Microbial Population Structure Course
  https://stamps.mbl.edu/index.php/Sue_Huse

- SSU Metagenomics (UPARSE)
  http://drive5.com/ssu.html

- MOTHUR manual
  http://www.mothur.org/wiki/Mothur_manual

- QIIME overview tutorial
  http://www.wernerlab.org/teaching/qiime/overview


www.sixthresearcher.com
@SixthResearcher