Introduction to Next Generation Sequencing (NGS) and Metabarcoding

Anders K. Krabberød

Department of Biosciences/ Norwegian Sequencing center

University of Oslo

a.k.krabberod@ibv.uio.no

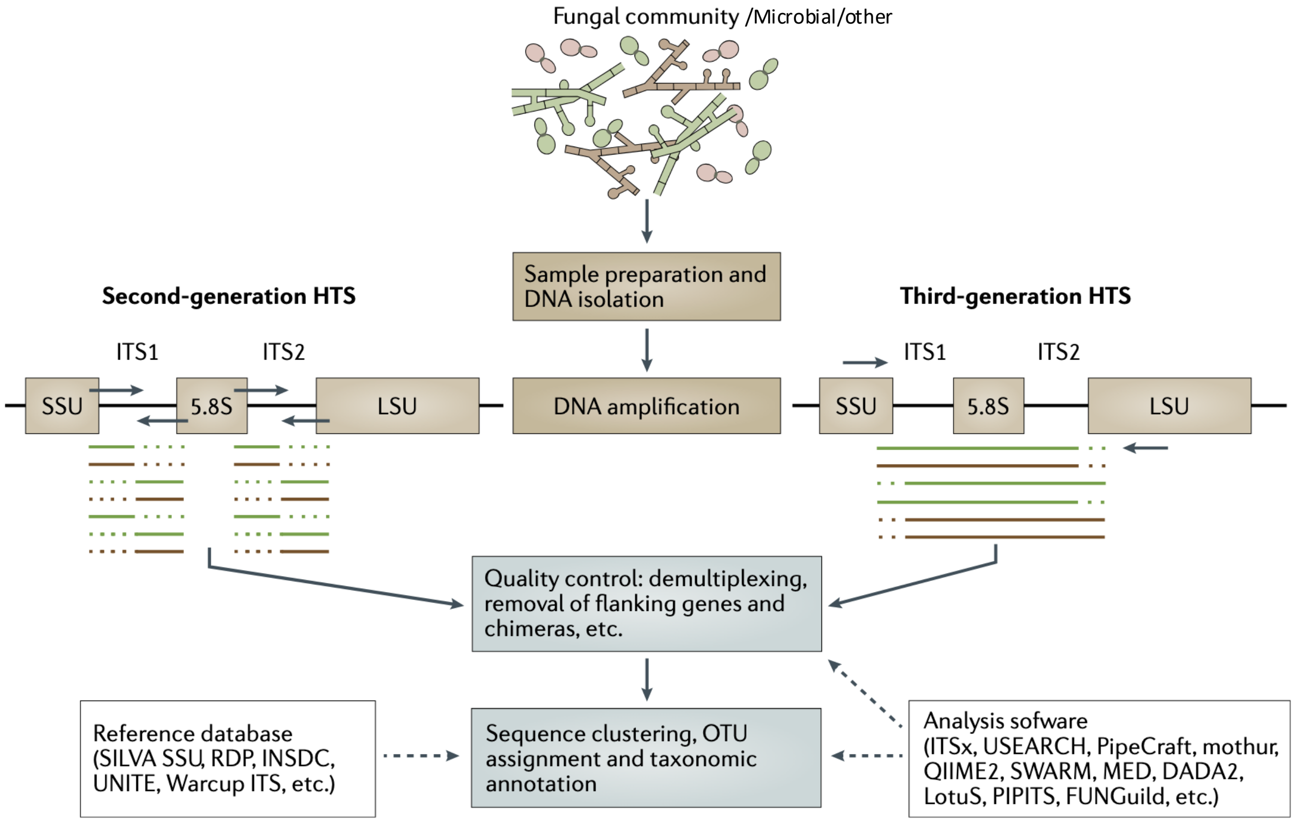
AB332 - 2024

UNIVERSITY
OF OSLO

Some important terms

- Metabarcoding
- Amplicons
- OTUs (and ASVs)
- High-throughput Sequencing (HTS)
- Next Generation Sequencing (NGS)
- Third-generation Sequencing

Metabarcoding



Fungal community /Microbial/other

Sample preparation and DNA isolation

DNA amplification

Second-generation HTS

Third-generation HTS

Quality control: demultiplexing, removal of flanking genes and chimeras, etc.

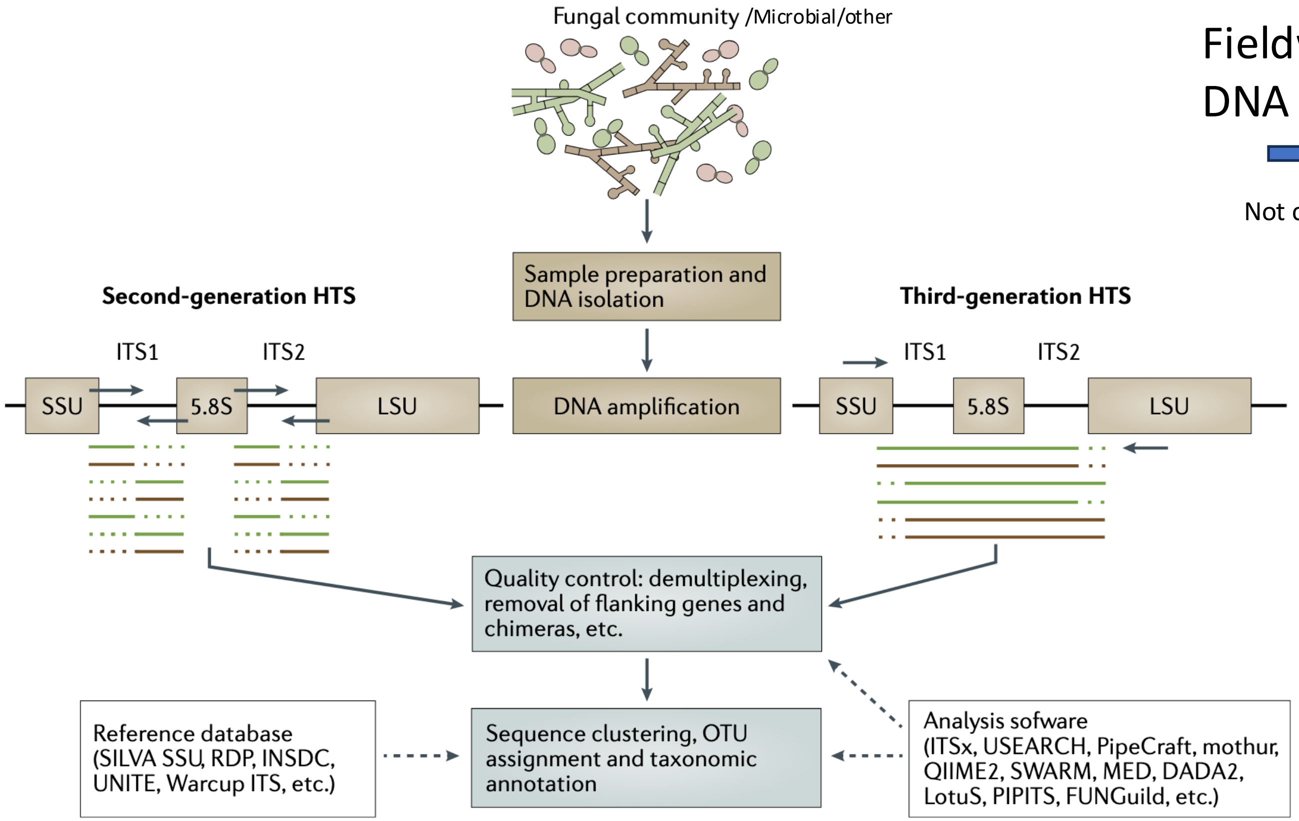
Sequence clustering, OTU assignment and taxonomic annotation

Reference database (SILVA SSU, RDP, INSDC, UNITE, Warcup ITS, etc.)

Analysis sofware (ITSx, USEARCH, PipeCraft, mothur, QIIME2, SWARM, MED, DADA2, LotuS, PIPITS, FUNGuild, etc.)

Nillson et al. 2019

Metabarcoding



Fungal community /Microbial/other

Sample preparation and DNA isolation

DNA amplification

Second-generation HTS

Third-generation HTS

Quality control: demultiplexing, removal of flanking genes and chimeras, etc.

Reference database (SILVA SSU, RDP, INSDC, UNITE, Warcup ITS, etc.)

Sequence clustering, OTU assignment and taxonomic annotation

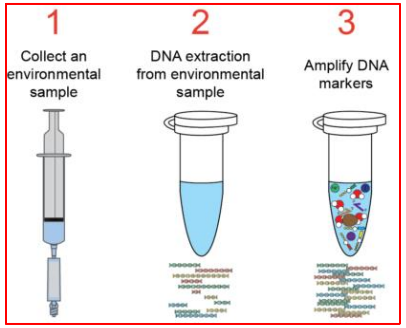
Analysis sofware (ITSx, USEARCH, PipeCraft, mothur, QIIME2, SWARM, MED, DADA2, LotuS, PIPITS, FUNGuild, etc.)

Nillson et al. 2019

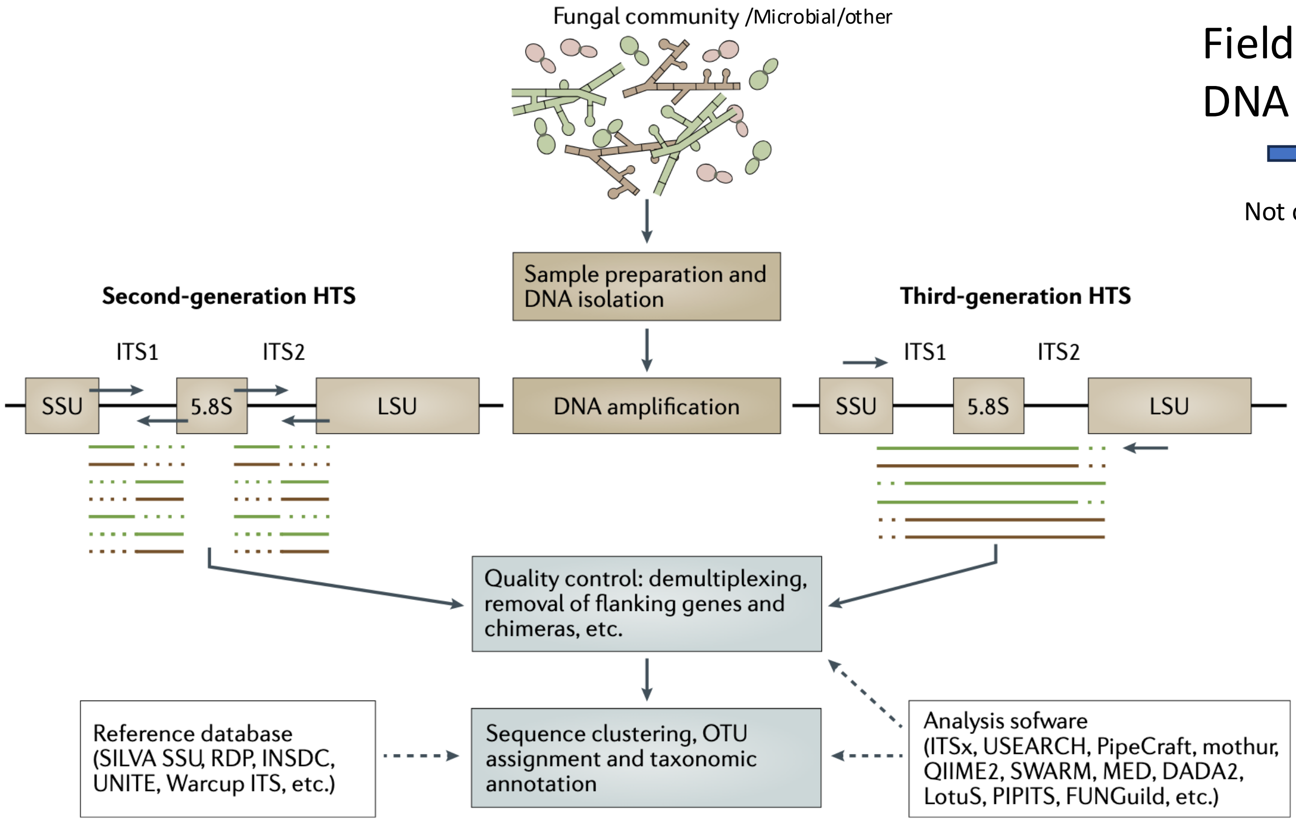
Fieldwork/labwork/ DNA extraction etc →

Not covered in this lecture

Abdelfattah et al.,2018


1
Collect an environmental sample
2
DNA extraction from environmental sample
3
Amplify DNA markers

Metabarcoding



Fungal community /Microbial/other

Sample preparation and DNA isolation

Second-generation HTS

Third-generation HTS

ITS1

ITS2

SSU

5.8S

LSU

DNA amplification

Quality control: demultiplexing, removal of flanking genes and chimeras, etc.

Reference database (SILVA SSU, RDP, INSDC, UNITE, Warcup ITS, etc.)

Sequence clustering, OTU assignment and taxonomic annotation

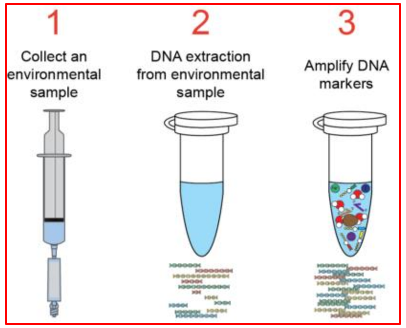
Analysis sofware (ITSx, USEARCH, PipeCraft, mothur, QIIME2, SWARM, MED, DADA2, LotuS, PIPITS, FUNGuild, etc.)

Nillson et al. 2019

Fieldwork/labwork/ DNA extraction etc

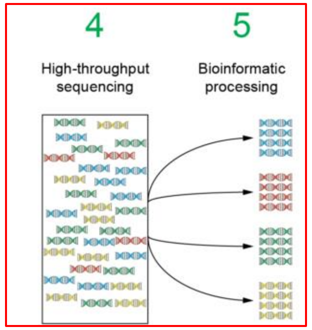
Not covered in this lecture

Abdelfattah et al.,2018

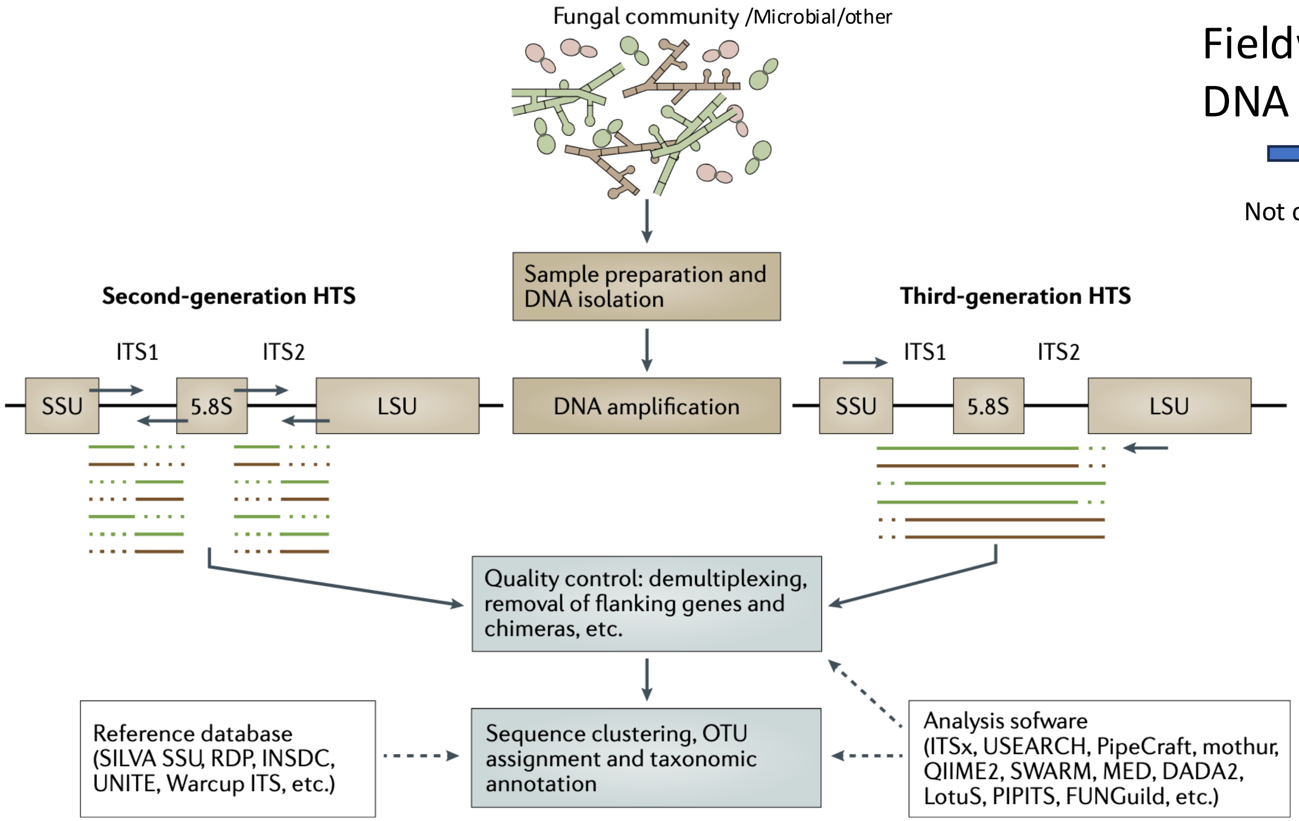

1
Collect an environmental sample
2
DNA extraction from environmental sample
3
Amplify DNA markers

Sequencing / Bioinformatics

Tuesday



Metabarcoding



Fungal community /Microbial/other

Sample preparation and DNA isolation

DNA amplification

Second-generation HTS

Third-generation HTS

Quality control: demultiplexing, removal of flanking genes and chimeras, etc.

Sequence clustering, OTU assignment and taxonomic annotation

Reference database (SILVA SSU, RDP, INSDC, UNITE, Warcup ITS, etc.)

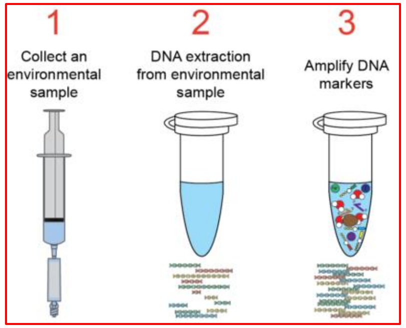
Analysis sofware (ITSx, USEARCH, PipeCraft, mothur, QIIME2, SWARM, MED, DADA2, LotuS, PIPITS, FUNGuild, etc.)

Nillson et al. 2019

Fieldwork/labwork/ DNA extraction etc

Not covered in this lecture

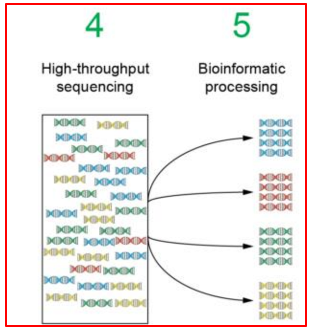
Abdelfattah et al.,2018



1
Collect an environmental sample
2
DNA extraction from environmental sample
3
Amplify DNA markers

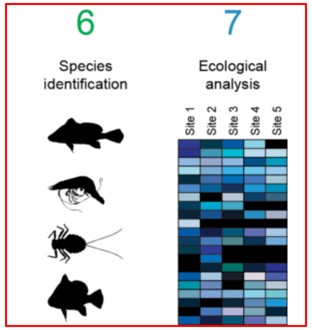
Sequencing / Bioinformatics

Tuesday

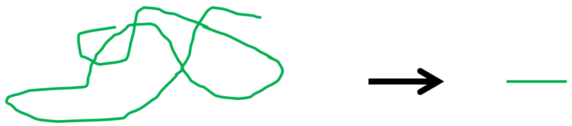


Ecological analysis

Computer Labs
Wed. and Thursday

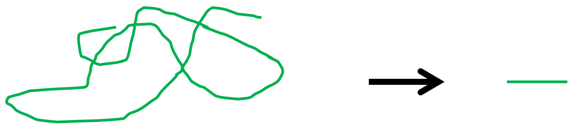


DNA Barcoding



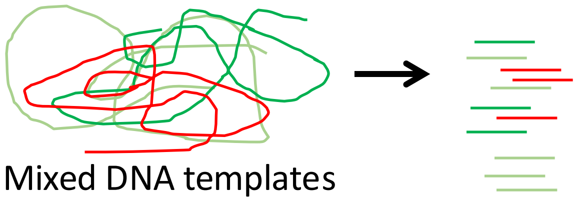
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

DNA Barcoding



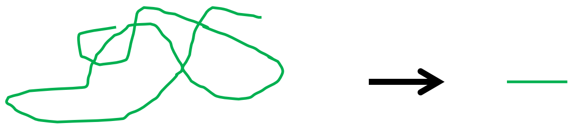
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

Metabarcoding



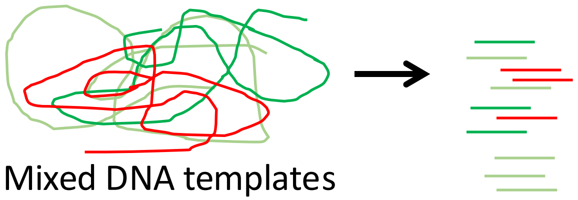
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

DNA Barcoding



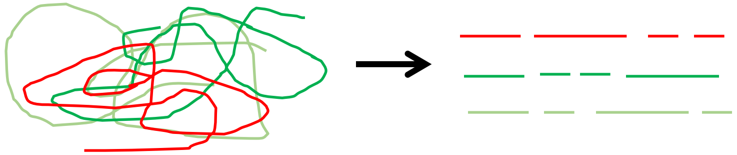
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

Metabarcoding



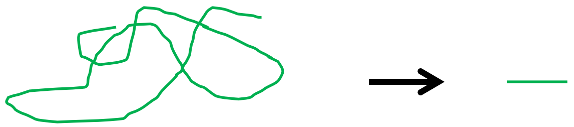
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

Metagenomics



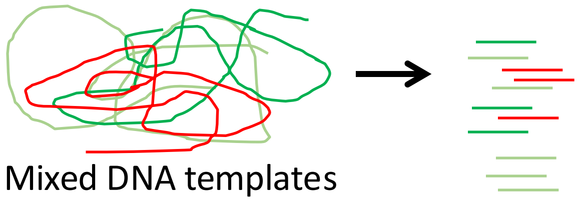
Genome-wide sequence variation in a community

DNA Barcoding



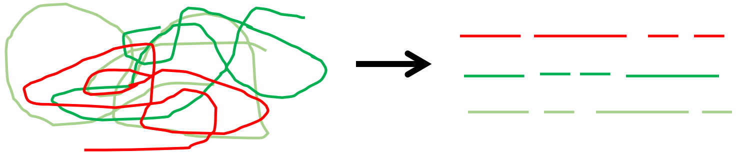
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

Metabarcoding



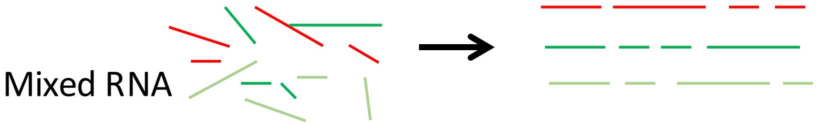
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

Metagenomics



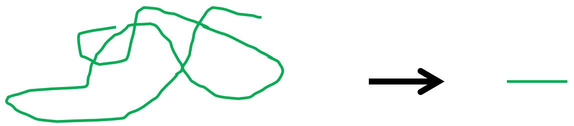
Genome-wide sequence variation in a community

Metatranscriptomics



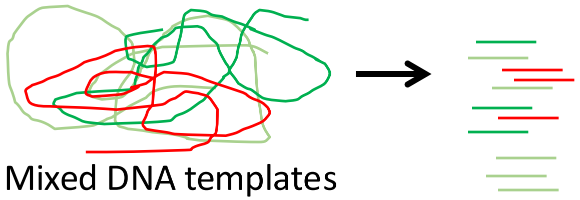
cDNA sequence variation in a community

DNA Barcoding



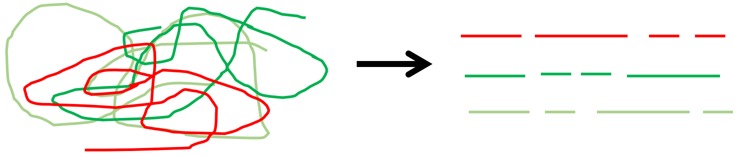
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

Metabarcoding



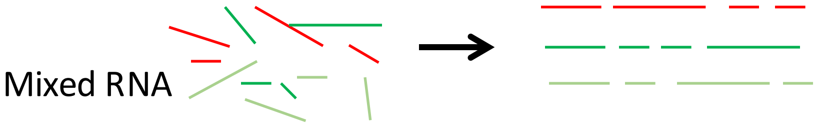
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

Metagenomics



Genome-wide sequence variation in a community

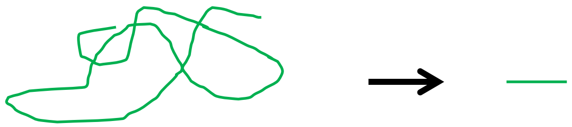
Metatranscriptomics



cDNA sequence variation in a community

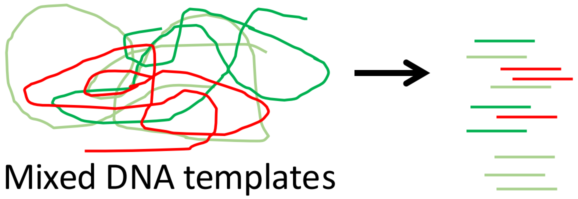
Who is active, and which genes?

DNA Barcoding



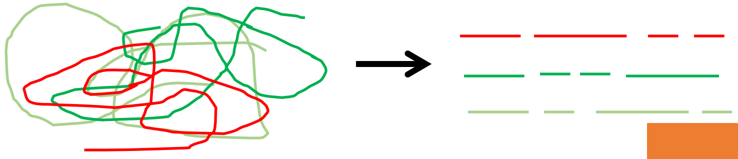
Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

Metabarcoding



Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

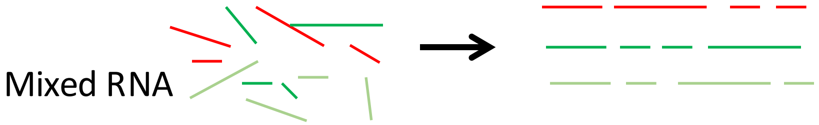
Metagenomics



Genome-wide sequence variation in a community

Which genes, and from whom?

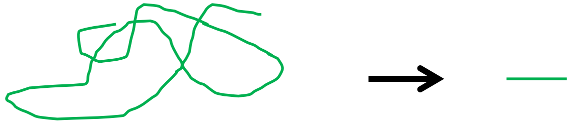
Metatranscriptomics



cDNA sequence variation in a community

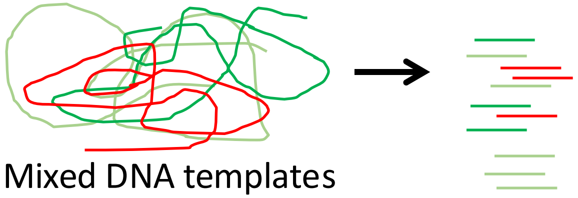
Who is active, and which genes?

DNA Barcoding



Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

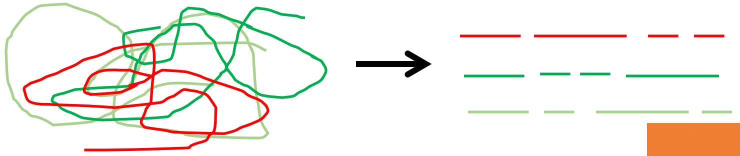
Metabarcoding



Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

Who is present?

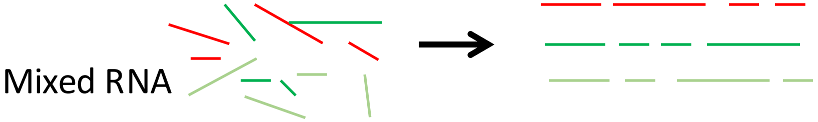
Metagenomics



Genome-wide sequence variation in a community

Which genes, and from whom?

Metatranscriptomics



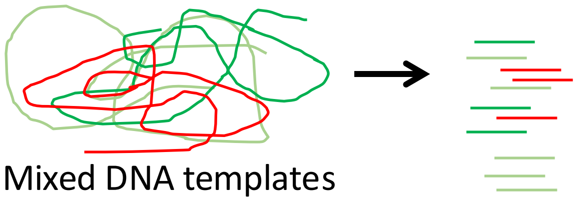
cDNA sequence variation in a community

Who is active, and which genes?

DNA Barcoding

Sequence variation in a single locus (e.g. ITS, 18S, COI) in a single specimen

Metabarcoding

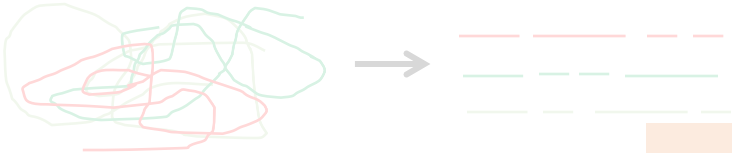


Mixed DNA templates

Sequence variation in a single locus (e.g. ITS, 18S, COI) in a community

Who is present?

Metagenomics

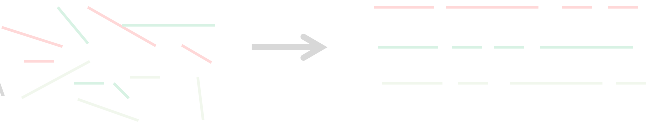


Genome-wide sequence variation in a community

Which genes, and from whom?

Metatranscriptomics

Mixed RNA



cDNA sequence variation in a community

Who is active, and which genes?

UiO : Universitetet i Oslo

Metabarcoding

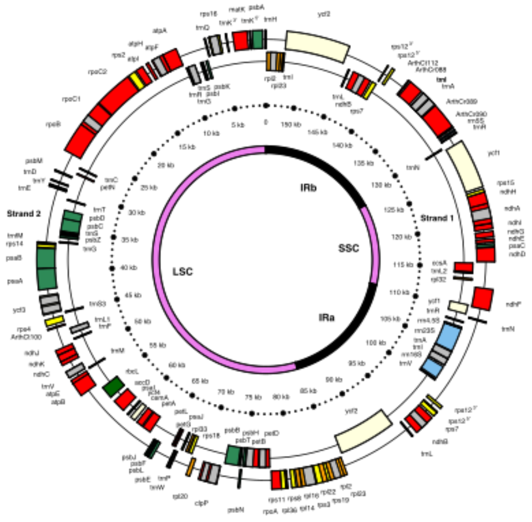
- Typical research questions:
  - Who are there?
  - Richness: How many taxa/species/OTUs (alpha/gamma diversity)?
    - **OUT:** Operational taxonomic unit -> the group of organisms currently being studied (Sokal & Sneath 1957), a modern take would be that OUT are pragmatic proxies for "species" at different taxonomic levels.
  - Compositional differences (beta diversity)?
  - Which processes and drivers are shaping the communities?
  - Co-occurrence patterns (possible interactions)

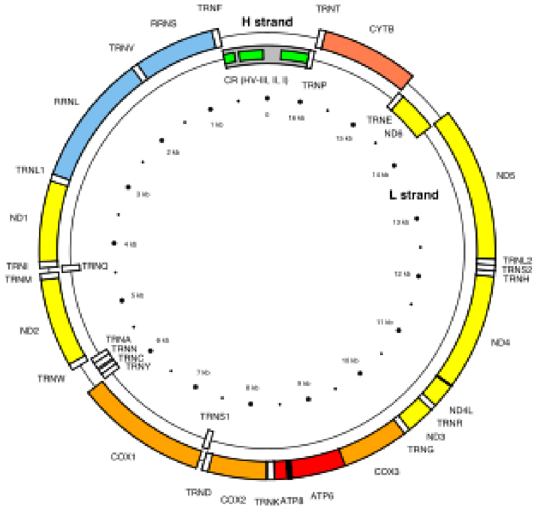
Markers in DNA metabarcoding

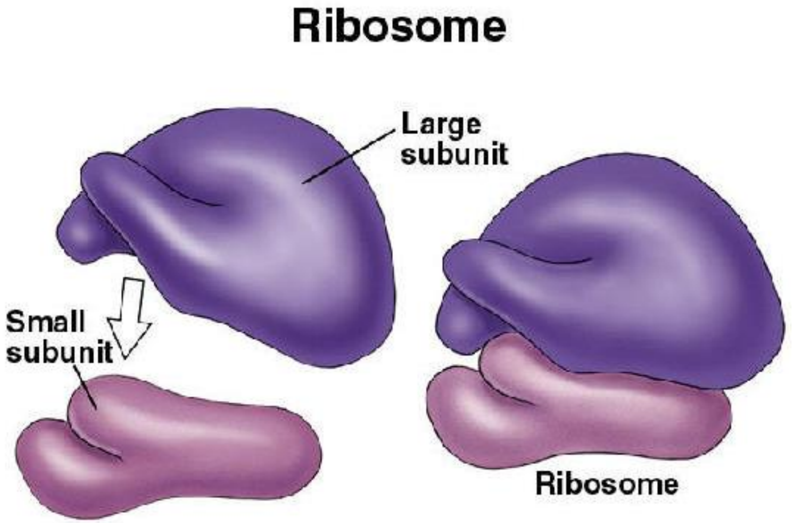
- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify in PCR
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length for all target organisms
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible to align (not always required)
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

Markers used in DNA metabarcoding

- Standard markers (<500 bp):
  - 18S: Eukaryotes
  - 16S: Bacteria/archaea
  - ITS: Fungi & plants
  - COI: Metazoa
  - *Rbc*L: Plants
  - *trn*L: Plants

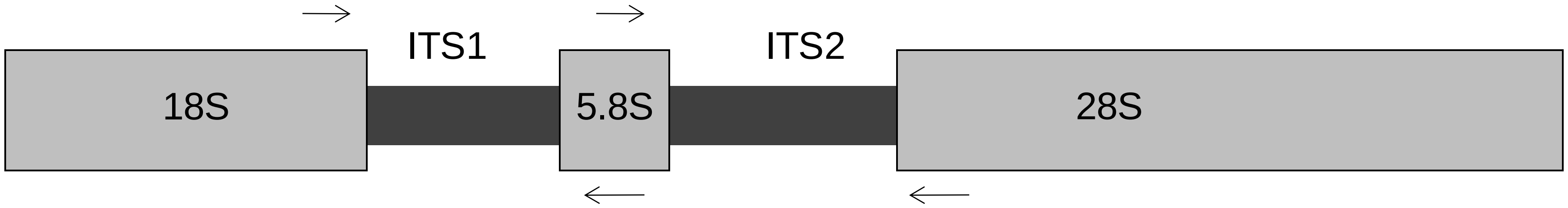






Ribosomal operon

- Often shrotened rRNA or rDNA
- 16S – 5S - 23S in prokaryotes
- 18S – 5.8S -28S in eukaryotes
  - S stands for Svedberg units; a unit of molecular size determined by centrifugation.
- Present across the Tree of Life!
- The full length is in the range of 5000-7000 (but with a lot of variation)
- Typically for Illumina sequencing a region of 300-450bp is used (e.g. V4).



The 18S marker

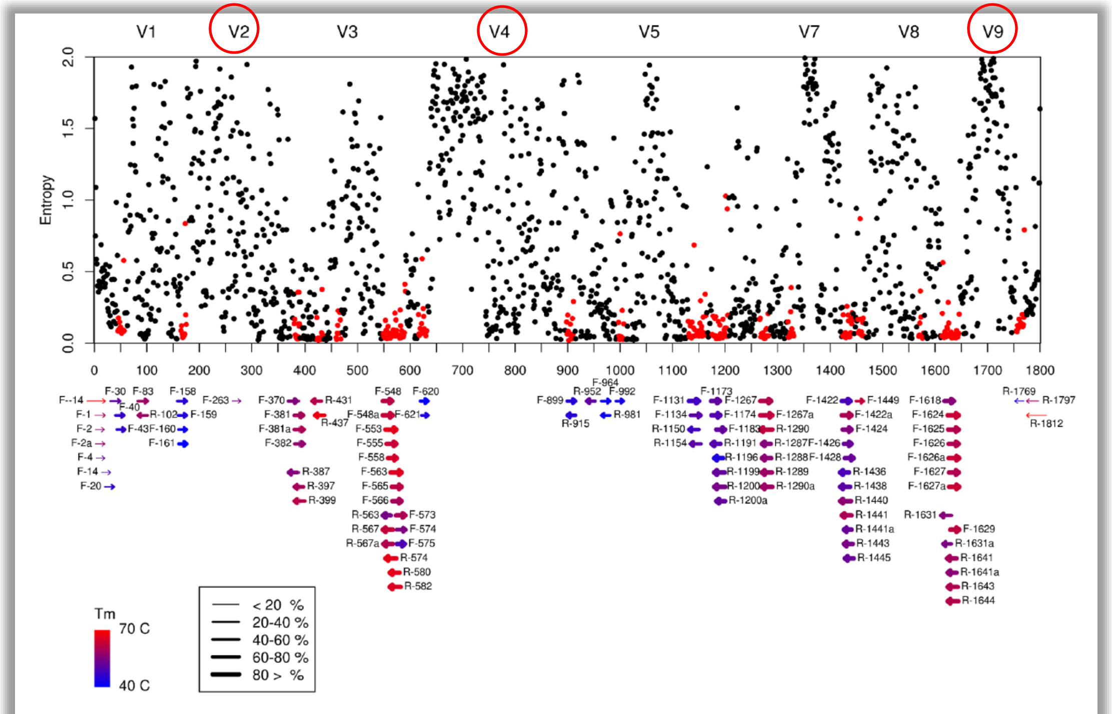
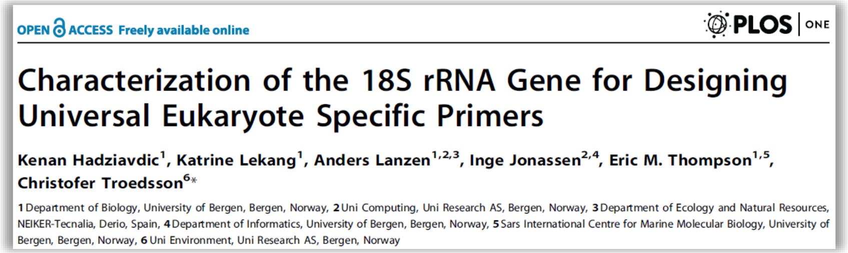


Figure 1. Nucleotide variability in 18S rRNA genes. Shannon entropy values of all eukaryotic alignment positions from the SILVA database along the 18S rRNA gene of *Saccharomyces cerevisiae*. Red dots mark consecutive nucleotide positions where at least 90% of ≥10 nt have entropy values lower than 0.2. The highly variable regions of 18S rRNA gene are denoted V1 to V9. In total, 100 primers targeting eukaryotes from the literature as well as those designed in this study are positioned along the reference sequence. The direction of the arrows indicates the orientation of the primers. The color denotes the melting temperature, and the thickness of the arrows represents the eukaryote universality of the primers. doi:10.1371/journal.pone.0087624.g001

- 18S is about 1700bp long
- Has 9 variable regions (named V1-V9)
- V4 (and some V9) are the most used for metabarcoding
- V4 has suitable length for Illumina sequencing (~450 bp)
- Primers exits that match most phyla

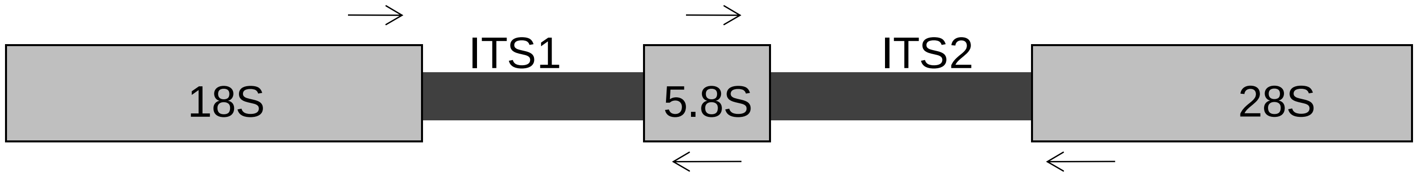


OPEN ACCESS Freely available online PLOS ONE

Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers

Kenan Hadziavdic[1], Katrine Lekang[1], Anders Lanzen[1,2,3], Inge Jonassen[2,4], Eric M. Thompson[1,5], Christofer Troedsson[6]*

1 Department of Biology, University of Bergen, Bergen, Norway, 2 Uni Computing, Uni Research AS, Bergen, Norway, 3 Department of Ecology and Natural Resources, NEIKER-Tecnalia, Derio, Spain, 4 Department of Informatics, University of Bergen, Bergen, Norway, 5 Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway, 6 Uni Environment, Uni Research AS, Bergen, Norway



How conserved/variable are the marker?

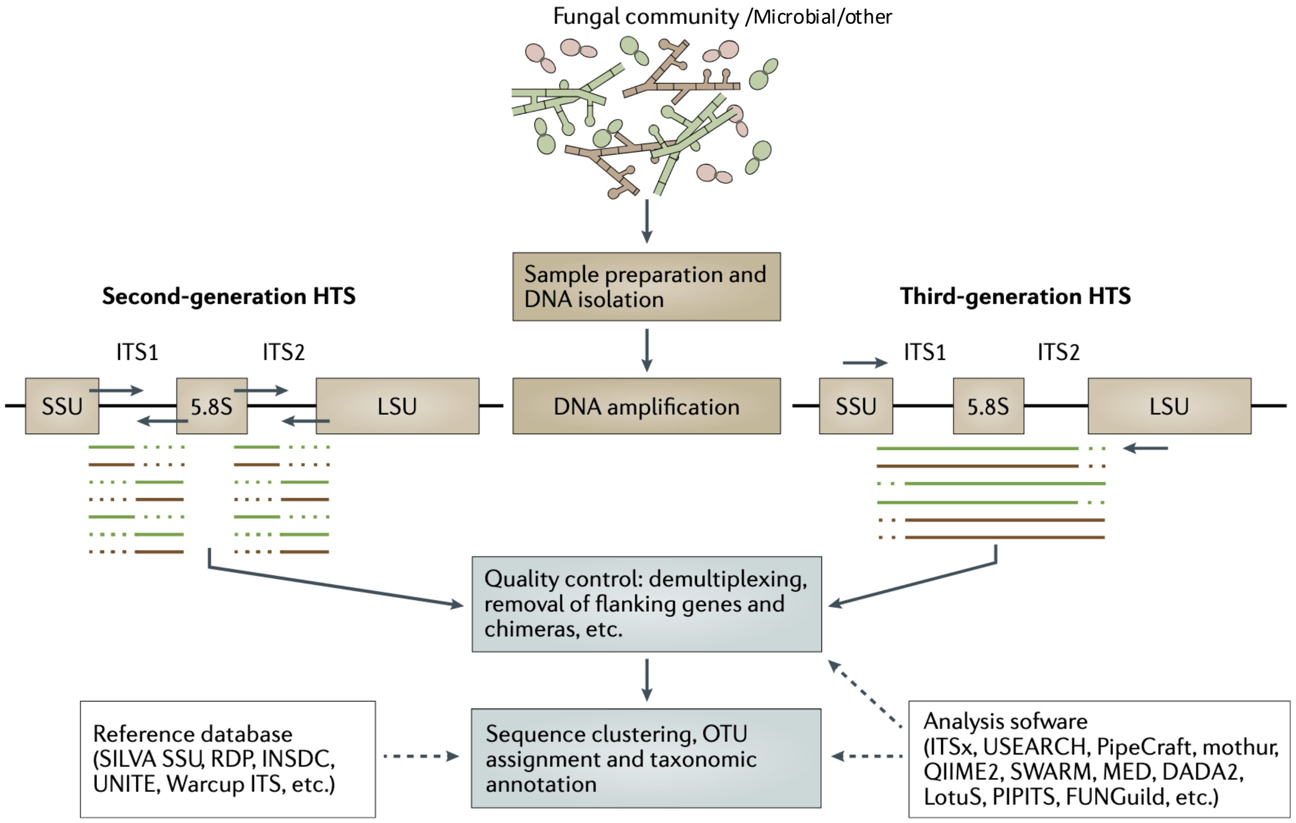
- 18S (and 16S): Low variability, low intraspecific variation, low interspecific variation

- ITS: High variability, high intraspecific variation, high ‘interspecific’ variation

⬇

- Impact how the bioinformatics analyses should be conducted

Metabarcoding



Fungal community /Microbial/other

Sample preparation and DNA isolation

DNA amplification

Second-generation HTS

Third-generation HTS

Quality control: demultiplexing, removal of flanking genes and chimeras, etc.

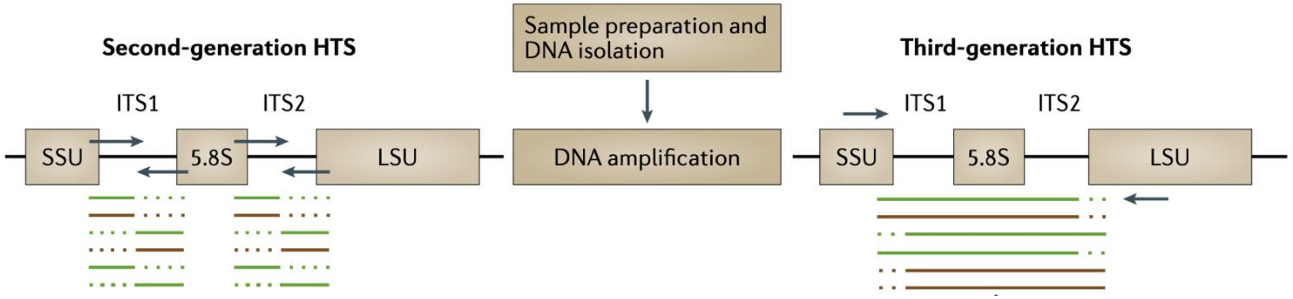
Sequence clustering, OTU assignment and taxonomic annotation

Reference database (SILVA SSU, RDP, INSDC, UNITE, Warcup ITS, etc.)

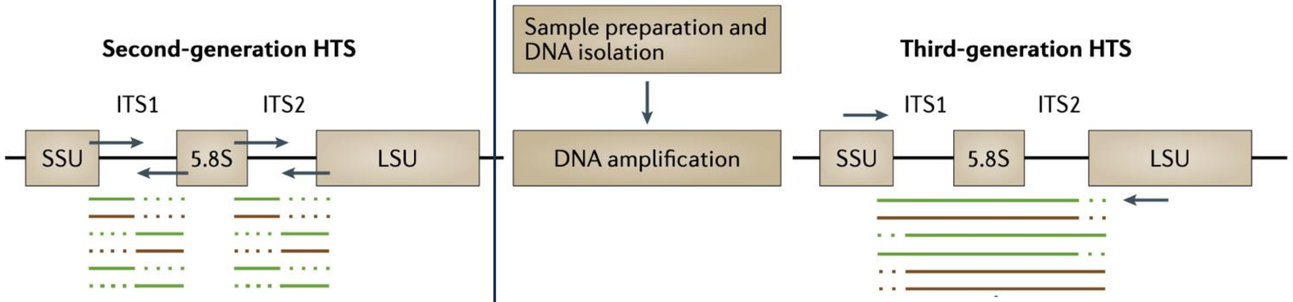
Analysis sofware (ITSx, USEARCH, PipeCraft, mothur, QIIME2, SWARM, MED, DADA2, LotuS, PIPITS, FUNGuild, etc.)

Nillson et al. 2019

Metabarcoding

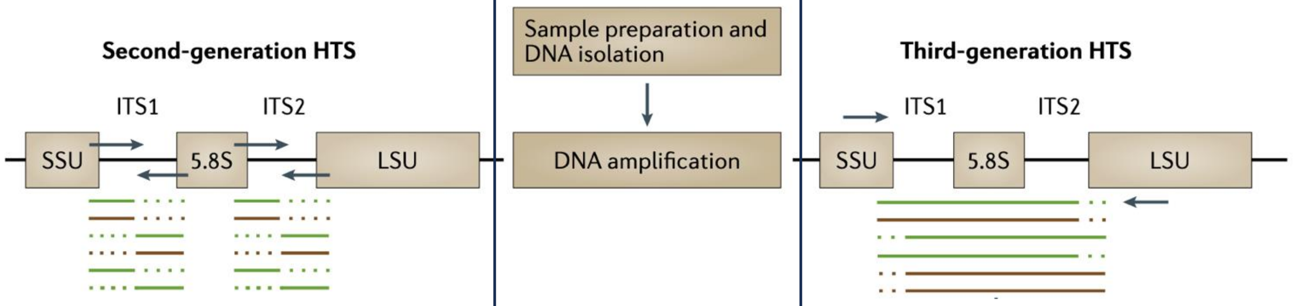


Metabarcoding



- Illumina
- 300bp – 600 bp

Metabarcoding



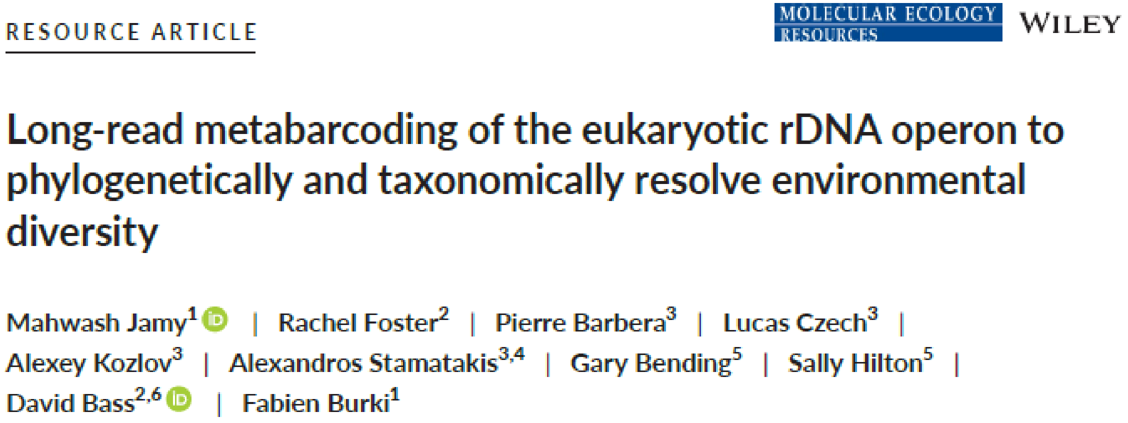
- Illumina
- 300bp – 600 bp

- PacBio
- Up tp 15 000 bp

Long-read metabarcoding

- Sequencing the full operon is possible with the development of sequencing technologies
- Stronger phylogenetic signal

- Comes with extra challenges
  - Harder to amplify longer regions
  - More chimeric sequences
  - Lower sequencing depth



RESOURCE ARTICLE

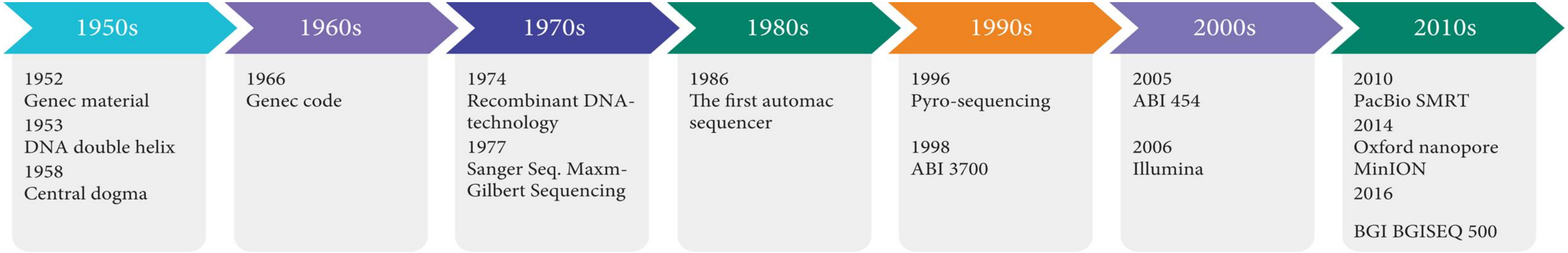
MOLECULAR ECOLOGY RESOURCES WILEY

Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity

Mahwash Jamy[1] | Rachel Foster[2] | Pierre Barbera[3] | Lucas Czech[3] | Alexey Kozlov[3] | Alexandros Stamatakis[3,4] | Gary Bending[5] | Sally Hilton[5] | David Bass[2,6] | Fabien Burki[1]

A short history of Sequencing

- Sequencing: determining the order of basepairs in a string of DNA (or RNA)
- The development started in the 1950’s after Watson and Crick described the structure of DNA
- The early methods were cumbersome (and dangerous) using radioactive material and adding individual nucleotides to a reaction one by one.
- The last few years the development has been phenomenal!



1950s

1952
Genetic material
1953
DNA double helix
1958
Central dogma

1960s

1966
Genetic code

1970s

1974
Recombinant DNA-technology
1977
Sanger Seq. Maxm-Gilbert Sequencing

1980s

1986
The first automatic sequencer

1990s

1996
Pyro-sequencing

1998
ABI 3700

2000s

2005
ABI 454

2006
Illumina

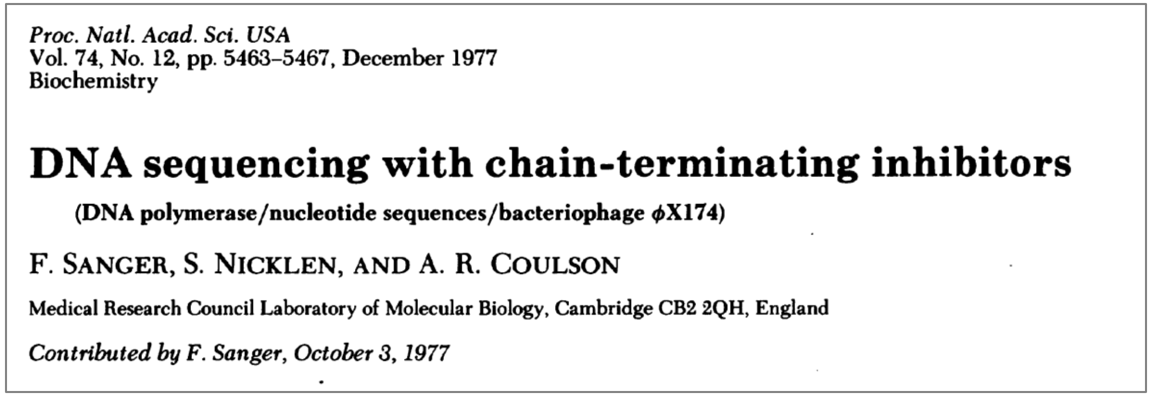
2010s

2010
PacBio SMRT
2014
Oxford nanopore MinION
2016

BGI BGISEQ 500

A short history of Sequencing

- F. Sanger et al. 1977
  - Short fragments
  - 15-200 nucleotides
  - Slooooow process



Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

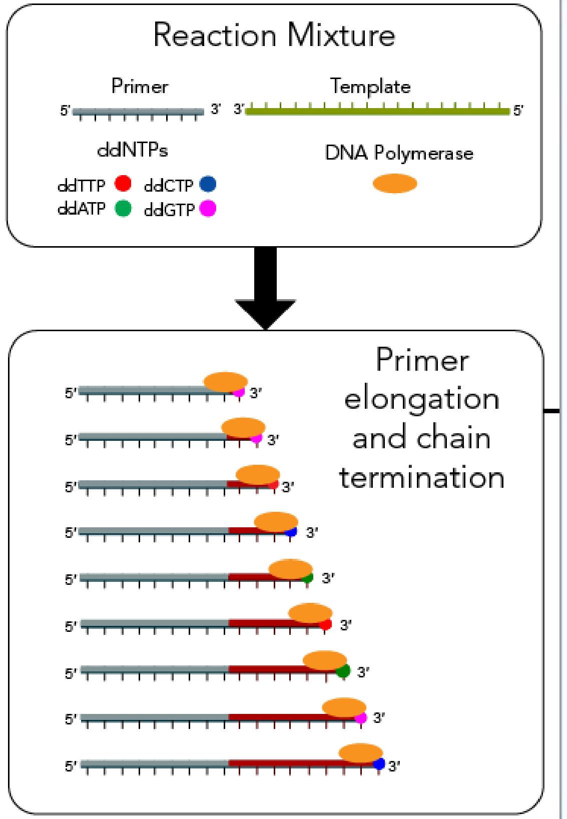
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

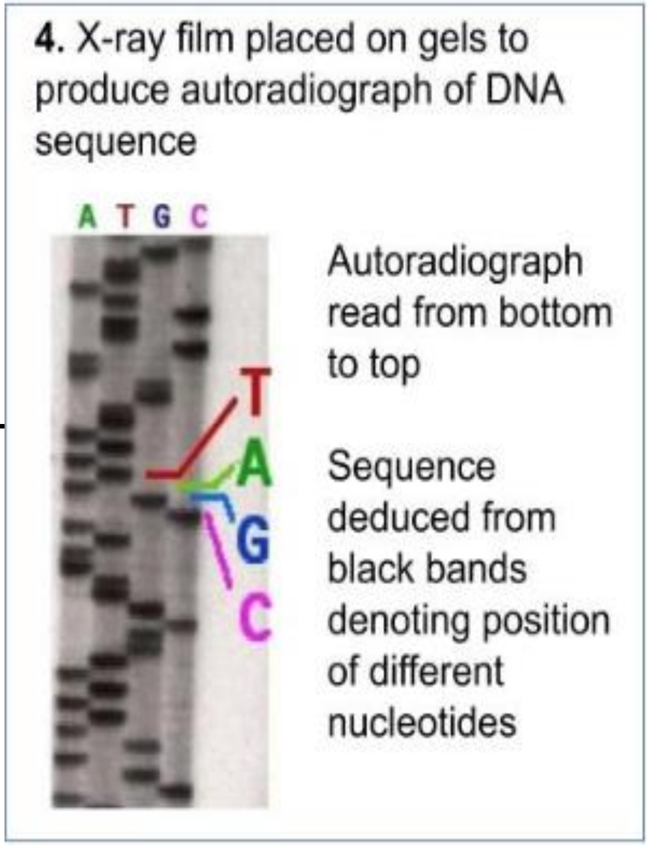
F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977



Primer elongation and chain termination



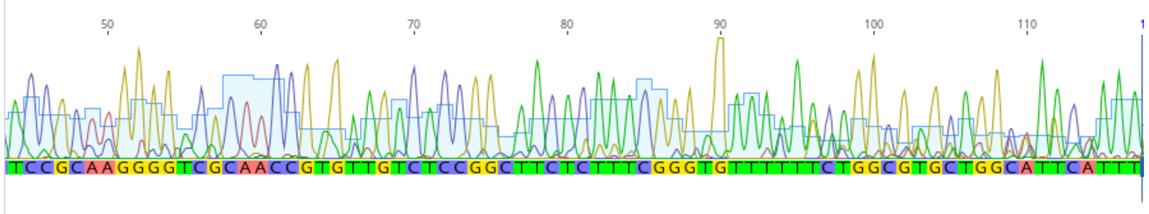
4. X-ray film placed on gels to produce autoradiograph of DNA sequence

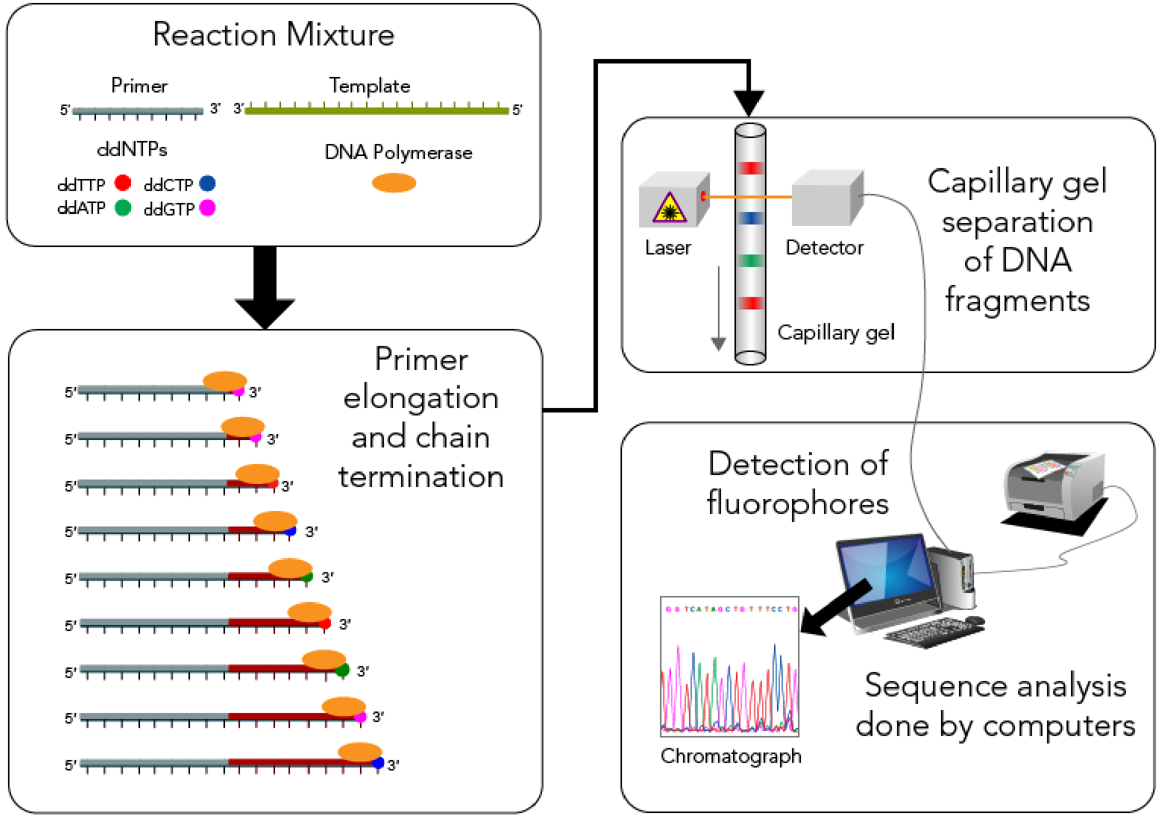
Autoradiograph read from bottom to top

Sequence deduced from black bands denoting position of different nucleotides

A short history of Sequencing

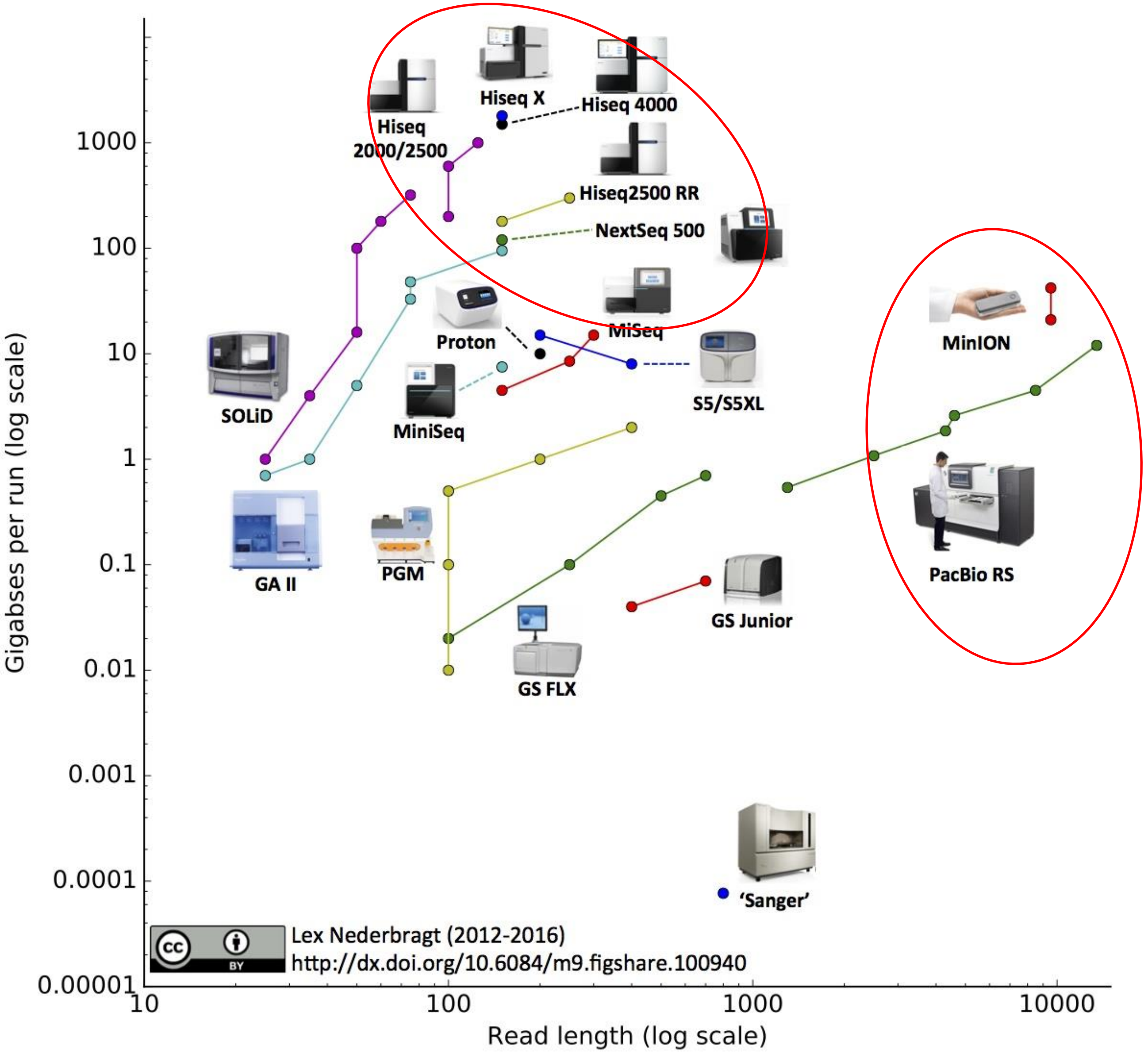
- F. Sanger et al. 1977
  - Short fragments
  - 15-200 nucleotides
  - Slooooow process
- Applied Biosystems automating the process in the late 80’s early 90’ with capillary electrophoresis, fluorescent dyes, and lasers.
- ”First generation sequencing”
  - 500—1000 nucleotides
  - Still slow, but faster than manual
  - The main technique for the human genome project
  - Sequencing 3 gigabases took 10 years
  - Still used, since it is very high quality and cheap (if you only want to look at a handful of sequences)




Reaction Mixture
Primer
Template
5'
3'
3'
5'
ddNTPs
ddTTP ●
ddATP ●
ddCTP ●
ddGTP ●
DNA Polymerase
Primer elongation and chain termination
Laser
Detector
Capillary gel
Capillary gel separation of DNA fragments
Detection of fluorophores
Chromatograph
Sequence analysis done by computers

High Throughput Sequencing (HTS)

- Early 2000’s and onwards
- High Throughput Sequencing (HTS) is a collective term
  - Next generation sequencing (NGS)
    - Short reads, (100-300bp)
    - but generates a huge amount of reads (in the billions)
    - 454 Roche
    - **Illumina (HiSeq, MiSeq, NovaSeq, NextSeq)**
    - Ion torrent
  - Third generation sequencing
    - Longer reads, (1000-100kbp)
    - not so many as NGS, but still in the 100k or millions
    - Oxford Nanopore (MinIon, GridIon,PromethIon, etc)
    - **PacBio (Sequel, Revio)**



Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

Illumina

- “Sequencing by synthesis”
- Short fragments
  - 150-300 bp in pairs
- Low error rate (0.1% - 0.5%)
- MiSeq output (2*300bp):
  - 25 million reads (15Gb)
- NextSeq (2*300 bp):
  - 1.2 billion reads (360Gb)
- NovaSeq 6000
  - 20 billion reads (6Tb)
- Other platforms exist

https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=1s

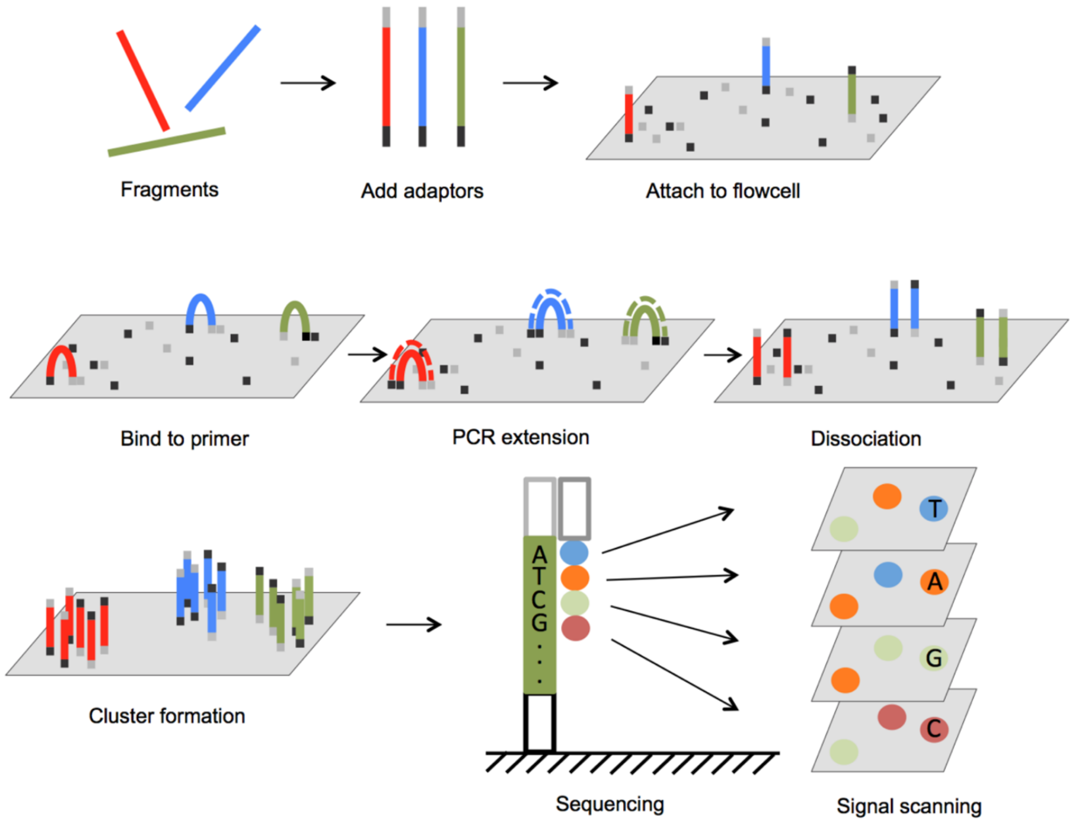


Figure 1: Principle of the illumina sequencing by synthesis (SBS) technology (Lu et al., 2016)

Illumina

- “Sequencing by synthesis”
- Short fragments
  - 150-300 bp in pairs
- Low error rate (0.1% - 0.5%)
- MiSeq output (2*300bp):
  - 25 million reads (15Gb)
- NextSeq (2*300 bp):
  - 1.2 billion reads (360Gb)
- NovaSeq 6000
  - 20 billion reads (6Tb)
- Other platforms exist

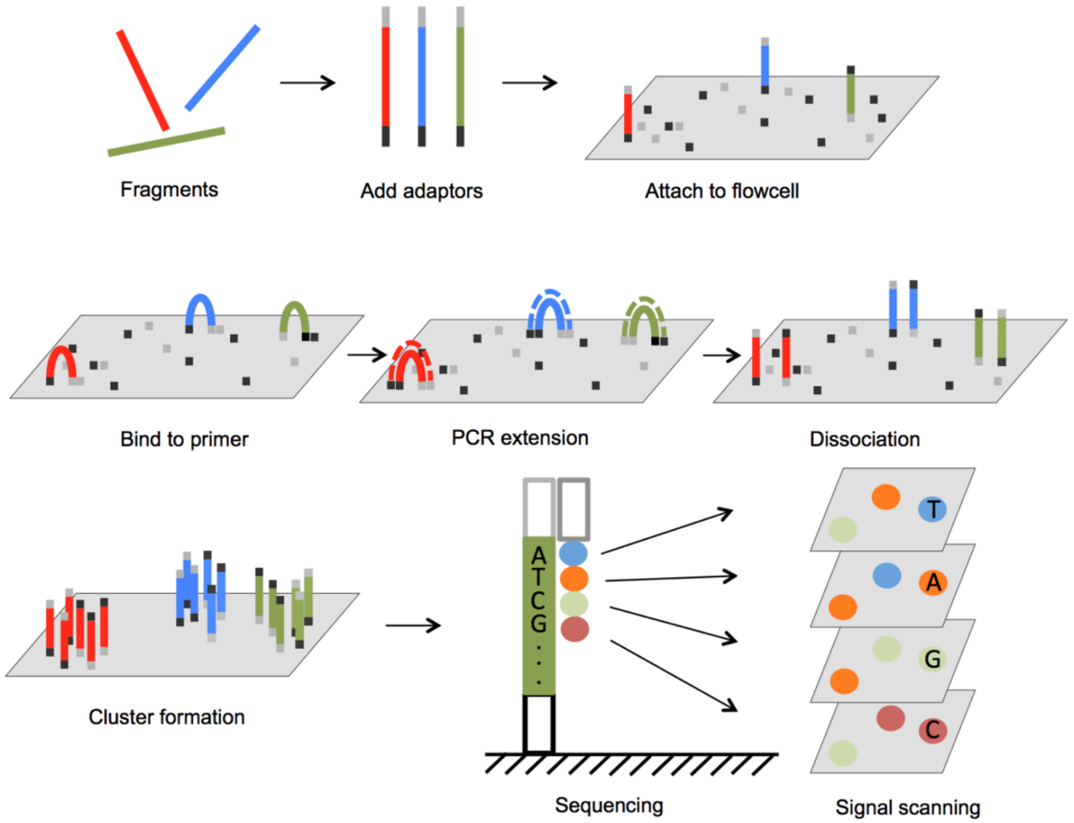


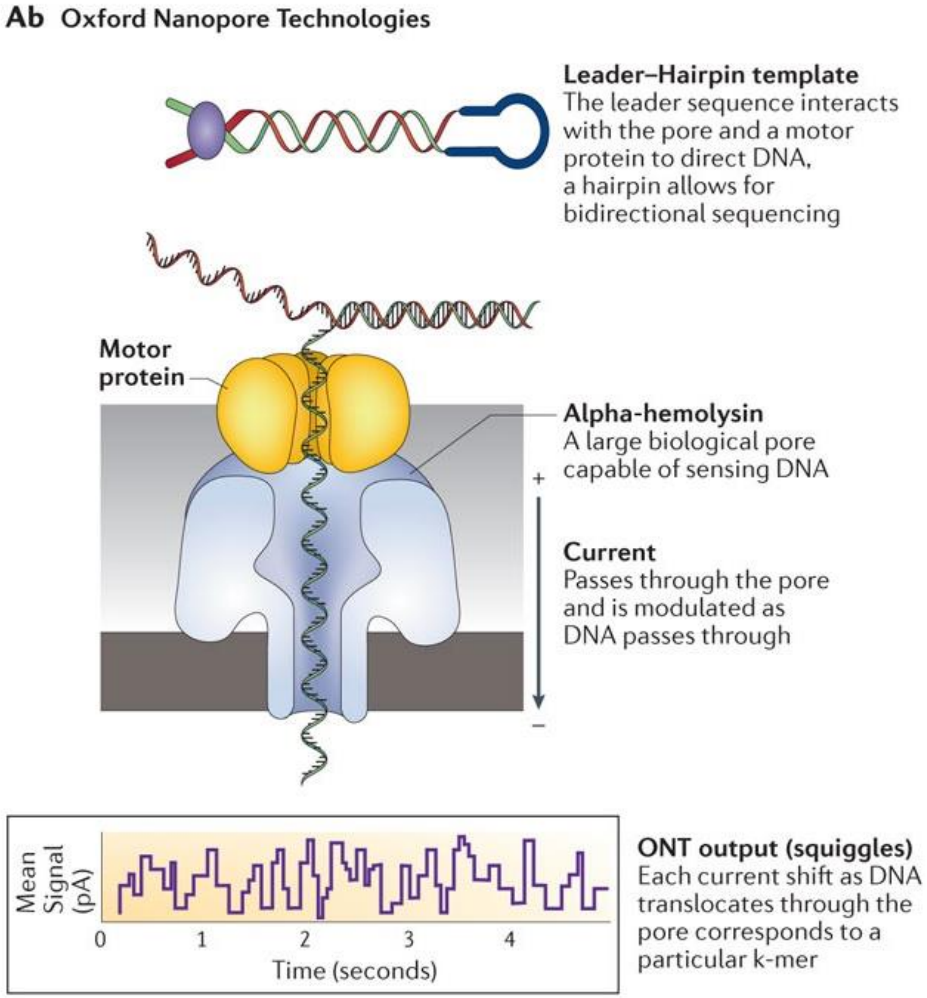
Figure 1: Principle of the illumina sequencing by synthesis (SBS) technology (Lu et al., 2016)

https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=1s

Goodwin, et al 2016

Oxford Nanopore

- Long to very long reads (10kb -100kb)
- Higher error rate, but it is improving
- Lower output than Illumina
  - MinION (50Gb)
  - PromethION (290Gb)
- Realtime sequencing
- Portable



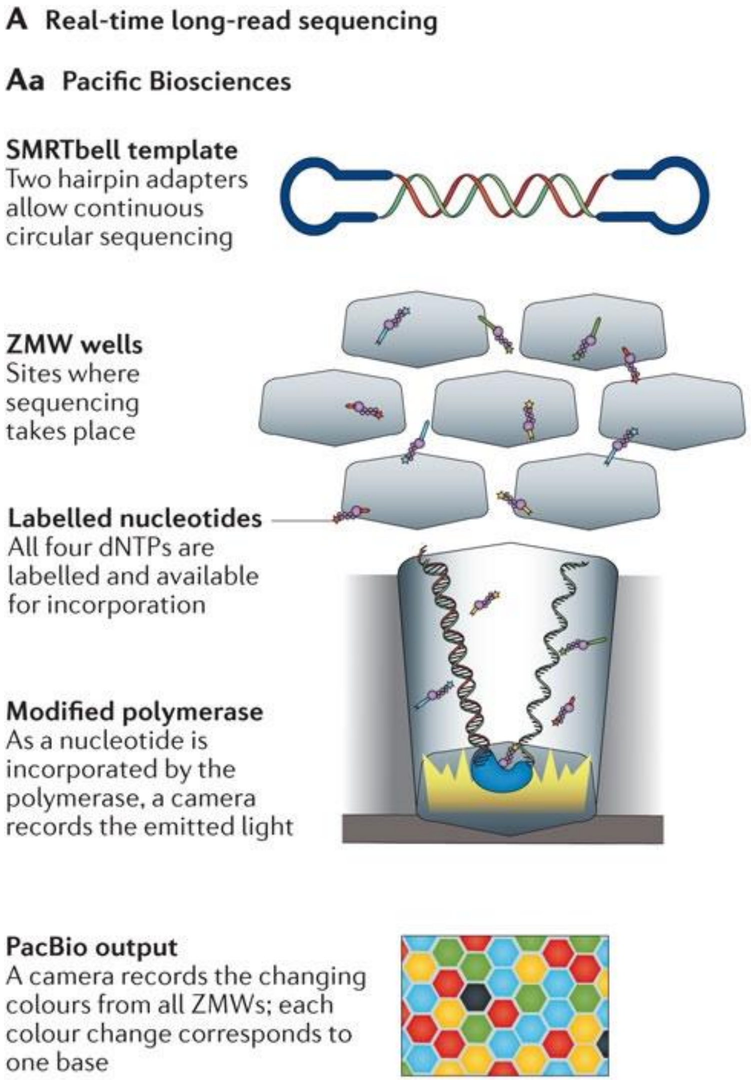
Ab Oxford Nanopore Technologies
Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing
Motor protein
Alpha-hemolysin
A large biological pore capable of sensing DNA
+
Current
Passes through the pore and is modulated as DNA passes through
–
Mean Signal (pA)
0
1
2
3
4
Time (seconds)
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

https://www.youtube.com/watch?v=RcP85JHLmnI&ab_channel=OxfordNanoporeTechnologies

Goodwin, et al 2016

PacBio

- SMRT-sequencing
  - Single-molecule Real Time
- Long reads (~15kb)
- Low error rate (0.1%)
- High output
  - Theoretical output:
  - Sequel II (up to 8M reads, 120 Gb)
  - Revio (up to 23M reads, 3Tb)
  - (Real output is ~75% of this)


A Real-time long-read sequencing
Aa Pacific Biosciences
SMRTbell template
Two hairpin adapters allow continuous circular sequencing
ZMW wells
Sites where sequencing takes place
Labelled nucleotides
All four dNTPs are labelled and available for incorporation
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base

https://www.youtube.com/watch?v=NHCJ8PtYCFc&ab_channel=PacBio

Goodwin, et al 2016

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

- These are the main steps when working with HTS data
- The details will vary depending on the sequencing technology used, the community under study, and the scientific answers being asked
- Some pipelines are built to do all steps for you automatically (Qiime, LotuS)
- Or you need to pick the relevant tool for your data

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Initial cleaning and quality controls

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

Post processing steps

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

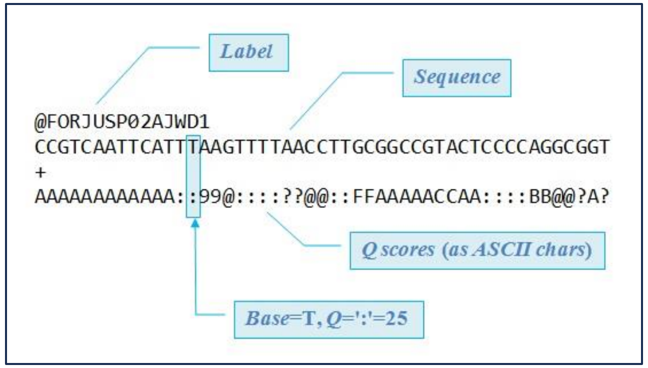
OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

The *fastq* format



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

The quality is in phred score

- 1-60, coded in ASCII characters
- 20 is 99% accuracy, 30 is 99.9%
- For a modern interpretation: https://fastqe.com/

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

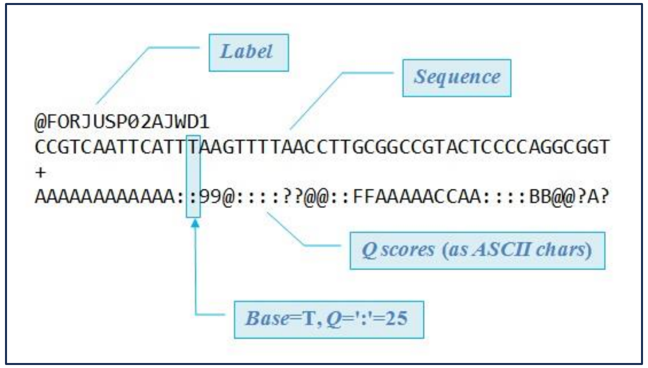
OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

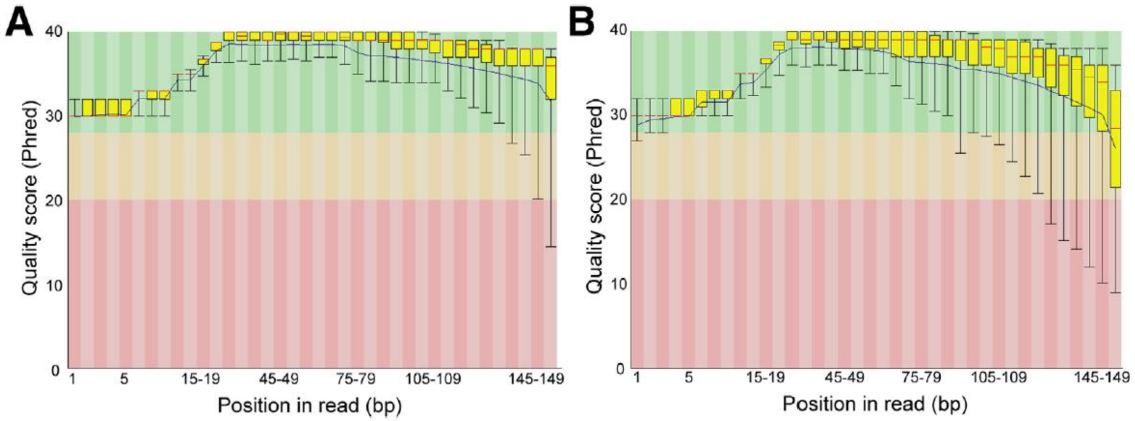
The *fastq* format



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

The quality is in phred score

- 1-60, coded in ASCII characters
- 20 is 99% accuracy, 30 is 99.9%
- For a modern interpretation: https://fastqe.com/



Remove

- Poor sequence quality
- Long/short sequences

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

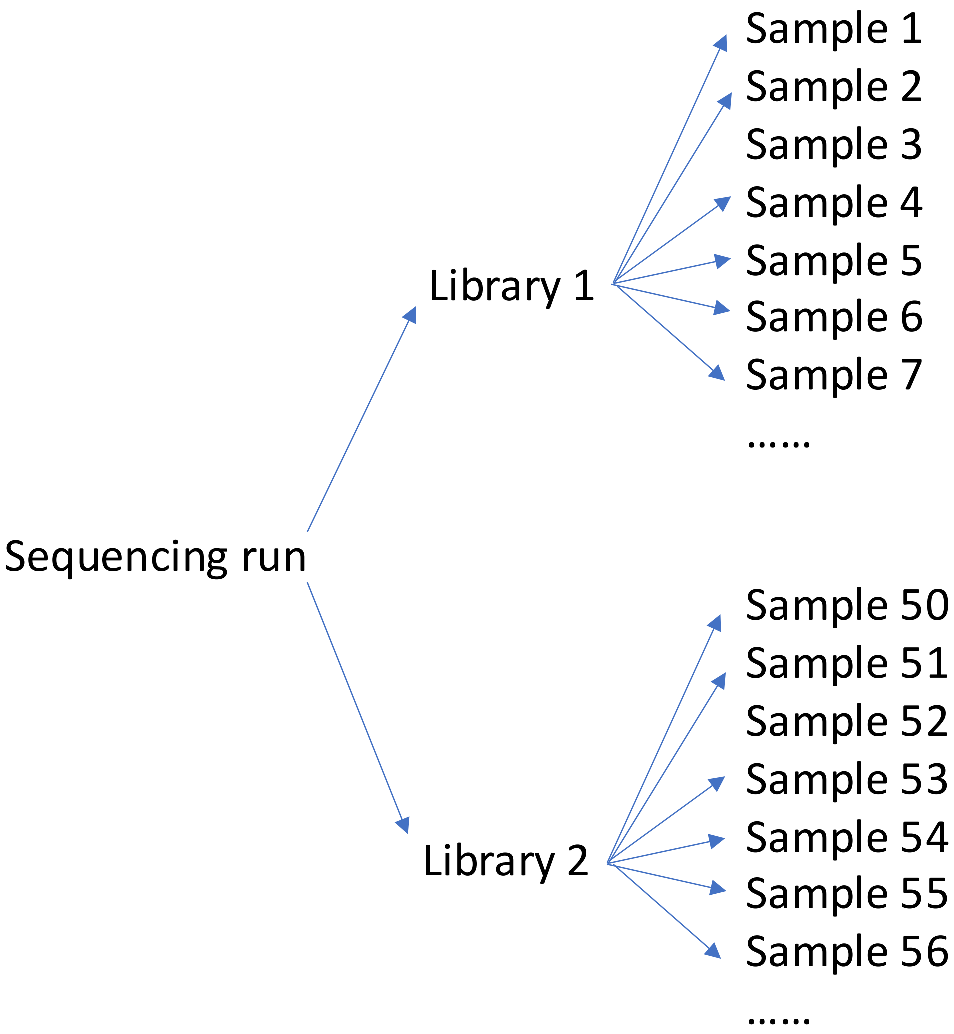
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

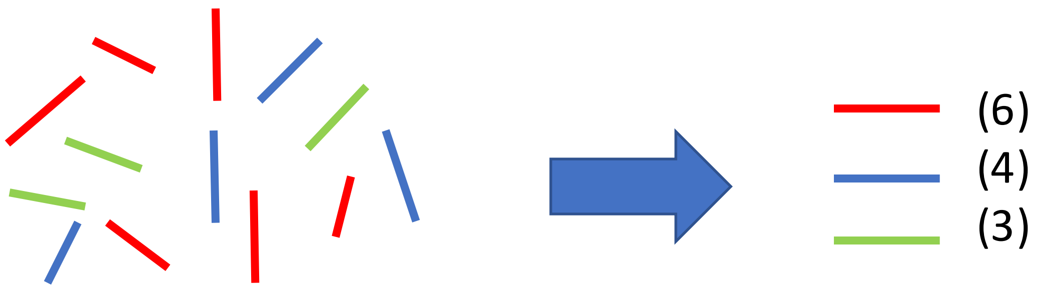
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



(k-mer based analyses)

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

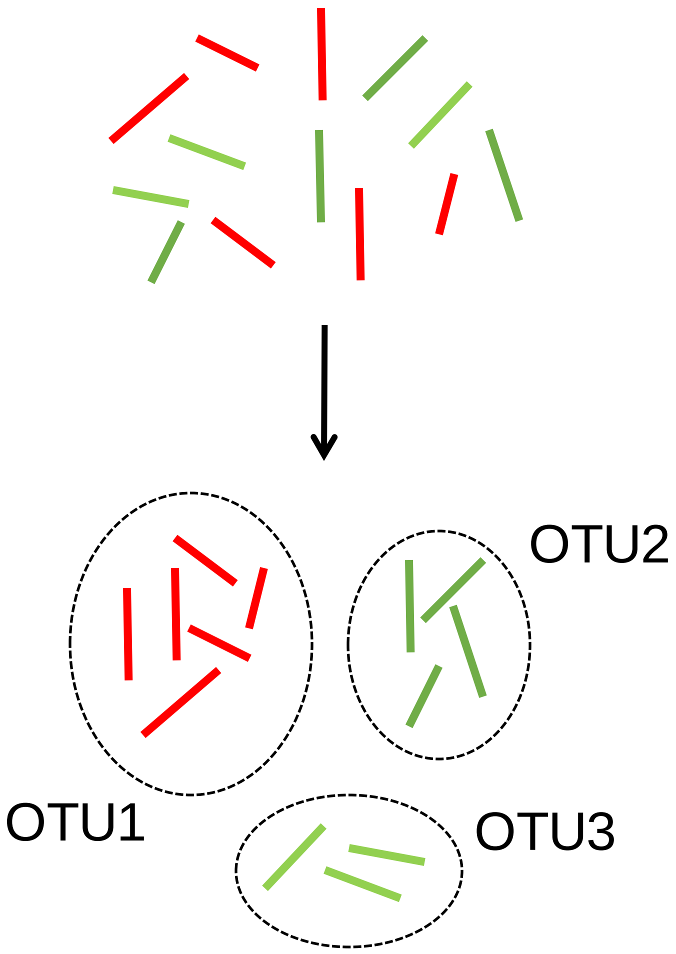
OTU modifications

Positive negatives

Singleton removal

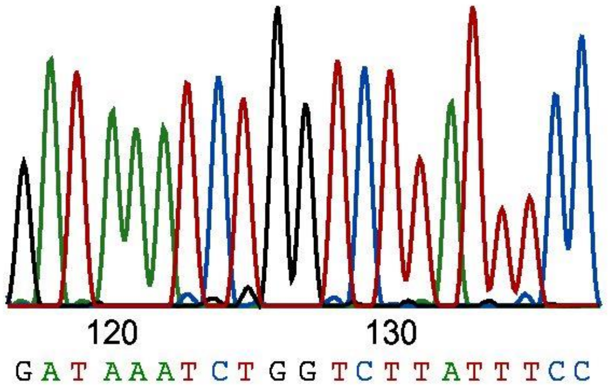
Data transformation (e.g rarification)

More on this later!!

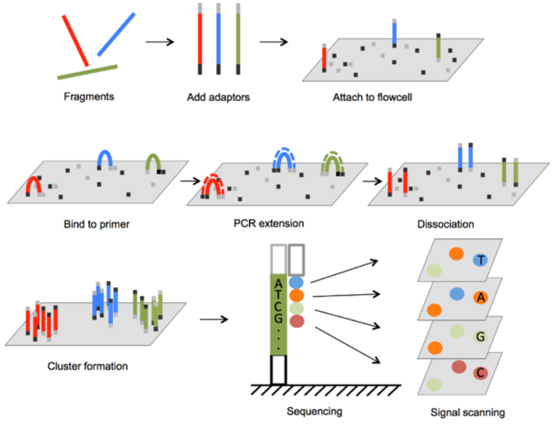


PCR-induced errors

- PCR mutations: polymerase enzymes introduce erroneous nucleotides now and then, even those enzymes with proof-reading activity
  - Dependent on the technology whether these become «visible»
  - In classic (direct) Sanger sequencing such errors become «diluted»
  - In methods where your final sequences are derived from one single DNA template, they become visible and must be corrected for!



Sanger



Illumina

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

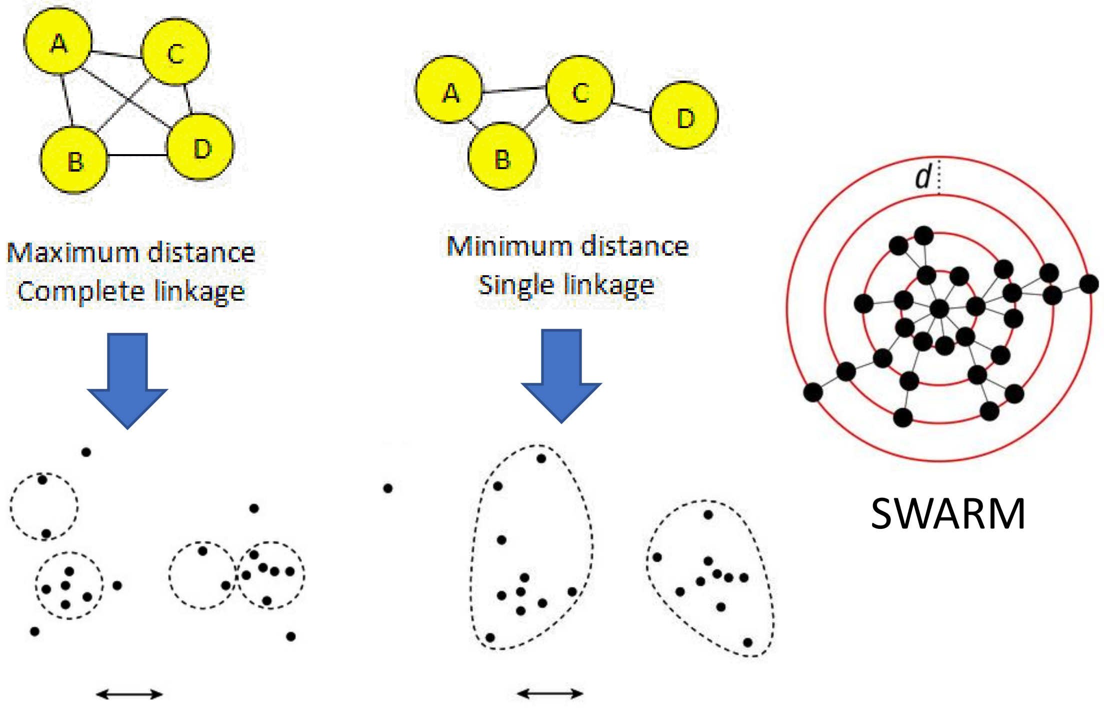
OTU modifications

Positive negatives

Singleton removal

Data transformation

Many different clustering approaches



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

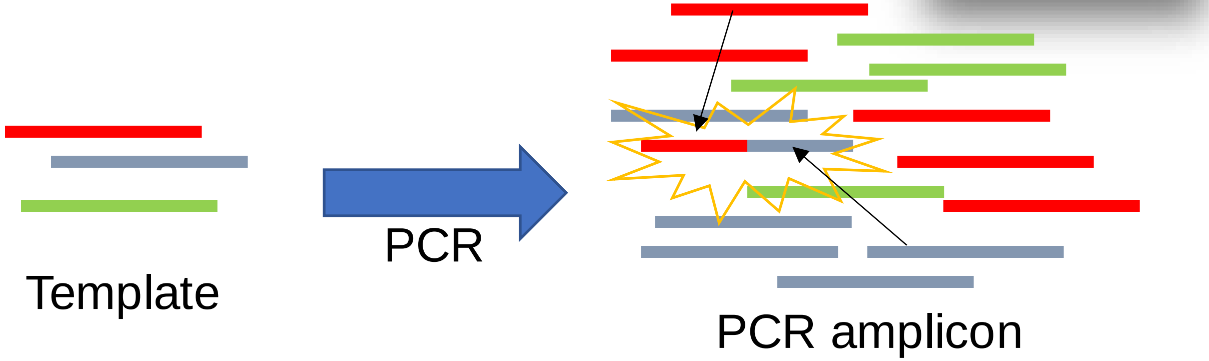
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



De novo versus reference based chimera checking

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

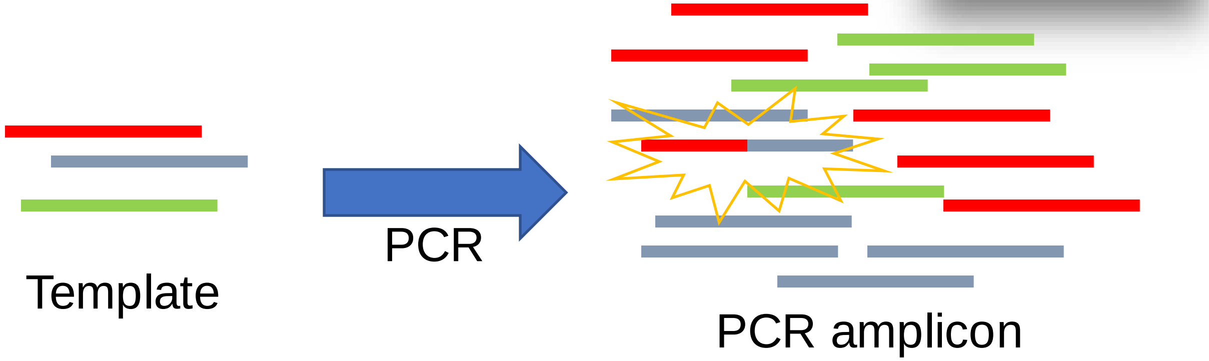
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



De novo ~~versus reference based~~ chimera checking

The level of chimeric sequences depends on how variable the marker is! → Be aware of false positives in the tests

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

PR2

NCBI

EUK REF

unite

silva
high quality ribosomal RNA databases

BOLD SYSTEMS

Why do taxonomic assignment?

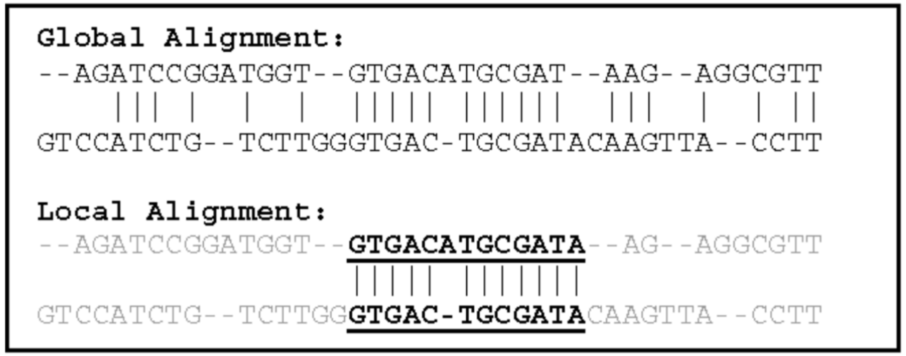
- Not strictly necessary to answer alpha and beta diversity questions
  - Detecting shifts in community composition and genetic diversity doesn’t require taxonomic assignments
- Assigning taxonomy links sequences to a wealth of pre-existing information
  - Linking sequences to species improves interpretation and explanation of patterns in alpha and beta diversity
- Choice of marker can impact taxonomic assignment
  - No marker is perfect
  - Discriminating power varies between markers and taxonomic groups
  - Database quality, availability, and completeness varies between markers

Taxonomic assignment - databases

- Different algorithms for comparing sequences to databases and scoring the results
- Alignment based
  - **BLAST**, vsearch, OBItools
- Phylogenetic based
  - HmmUFOtu, TIPP, DECARD, SAP
- Kmer-based machine learning approaches
  - RDP, UTAX, SINTAX

Taxonomic assignment - Alignment based

- Alignment based
  - **BLAST (Basic Local Alignment Search Tool) , vsearch**
- Local alignments begin by checking a small piece of the query sequence against the reference database and then expanding the match to find areas of high similarity
- Global alignments find the best match in the reference database across the entire length of the query sequence
- Output is typically an alignment score, percent identity, and coverage score



```
Global Alignment:
--AGATCCGGATGGT--GTGACATGCGAT--AAG--AGGCGTT
    ||| |  |  |  ||||| ||||||  |||  |  | ||
GTCCATCTG--TCTTGGGTGAC-TGCGATACAAGTTA--CCTT

Local Alignment:
--AGATCCGGATGGT--GTGACATGCGATA--AG--AGGCGTT
                 ||||| |||||||
GTCCATCTG--TCTTGGGTGAC-TGCGATACAAGTTA--CCTT
```

62% similarity

93% similarity
32% coverage

Taxonomic assignment - databases

- Taxonomic assignment quality is highly dependent on database accuracy and completeness
  - Misidentified sequences create identification errors and low-quality assignments
  - Missing reference sequences reduce the resolution of taxonomic assignments or result in misidentifications
- Taxonomic assignment is only as good as the database!
  - **PR2**: a curated database for protists 18S
  - **Silva**: 18S and 28S, Bacteria, Archaea, and Eukaryota. Comprehensive, but not curated.
  - **UNITE:** ITS. Mostly used for Fungi, other eukaryotes are also included
- Quality of taxonomic assignments can often be improved by creating custom-curated reference databases

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

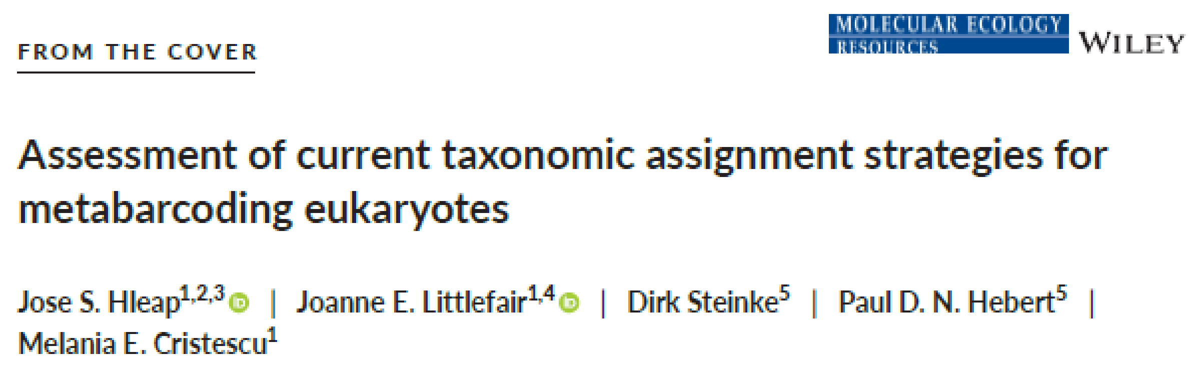
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



FROM THE COVER

MOLECULAR ECOLOGY RESOURCES WILEY

Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes

Jose S. Hleap[1,2,3] | Joanne E. Littlefair[1,4] | Dirk Steinke[5] | Paul D. N. Hebert[5] | Melania E. Cristescu[1]

Hleap et al. 2021. Mol Ecol Resources

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

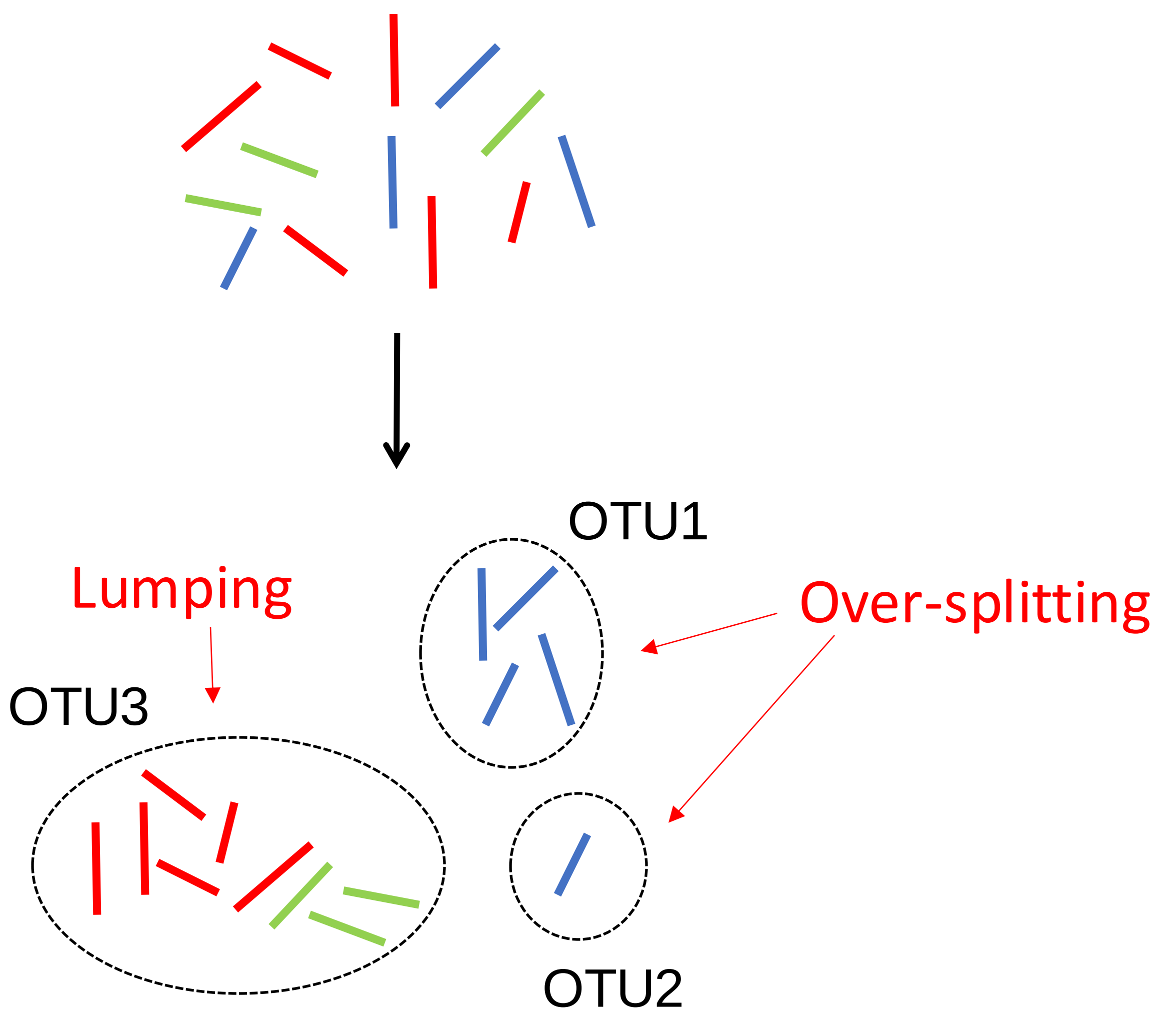
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

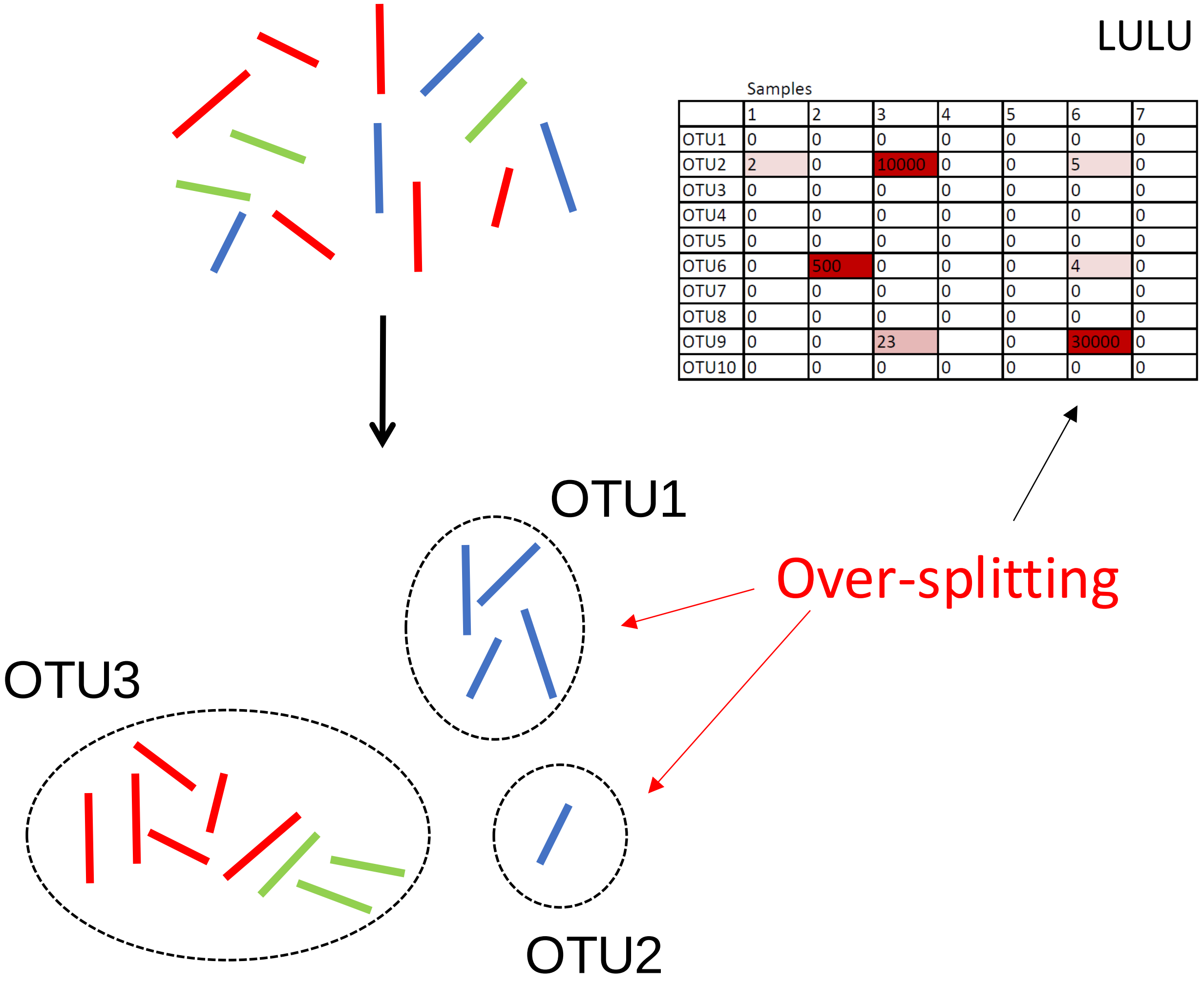
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation



LULU

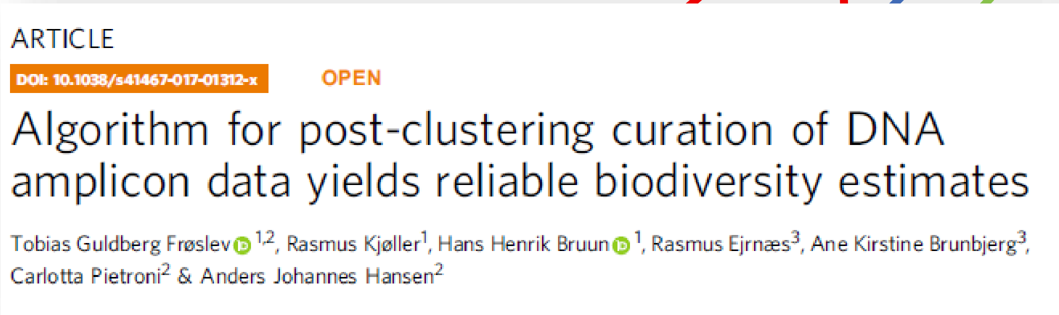
Samples

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	500	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	30000	0
OTU10	0	0	0	0	0	0	0

Bioinformatics – main steps

Quality control

Produce contigs



ARTICLE

DOI: 10.1038/s41467-017-01312-x OPEN

Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates

Tobias Guldberg Frøslev[1,2], Rasmus Kjøller[1], Hans Henrik Bruun[1], Rasmus Ejrnæs[3], Ane Kirstine Brunbjerg[3], Carlotta Pietroni[2] & Anders Johannes Hansen[2]

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

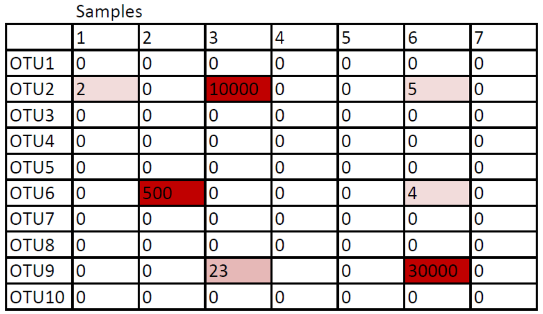
OTU modifications

Positive negatives

Singleton removal

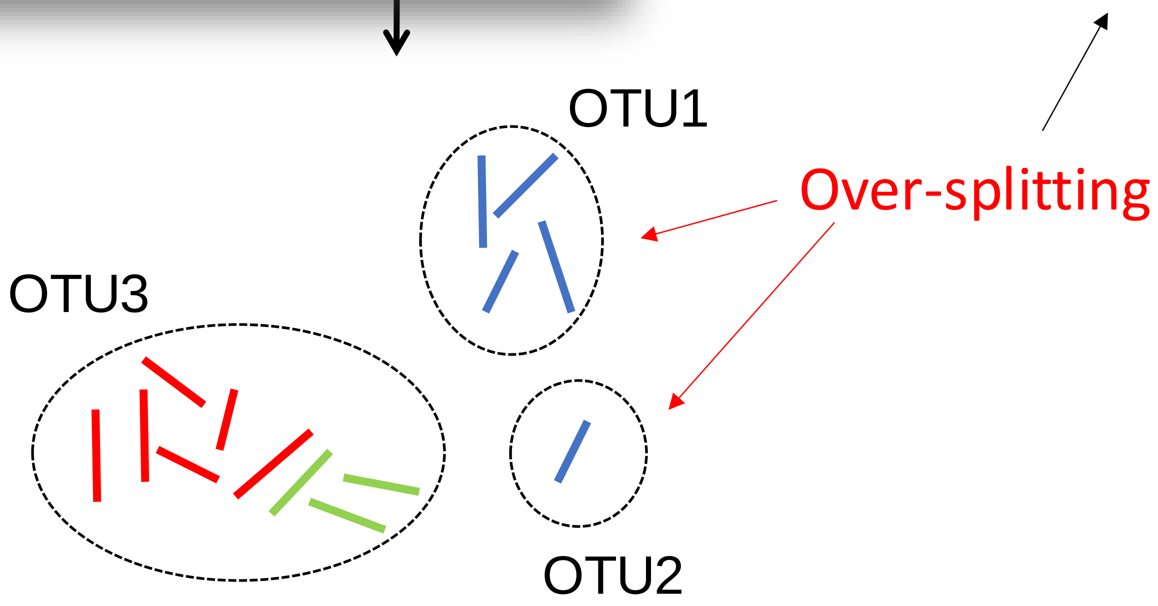
Data transformation

LULU



Samples

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	500	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	30000	0
OTU10	0	0	0	0	0	0	0



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

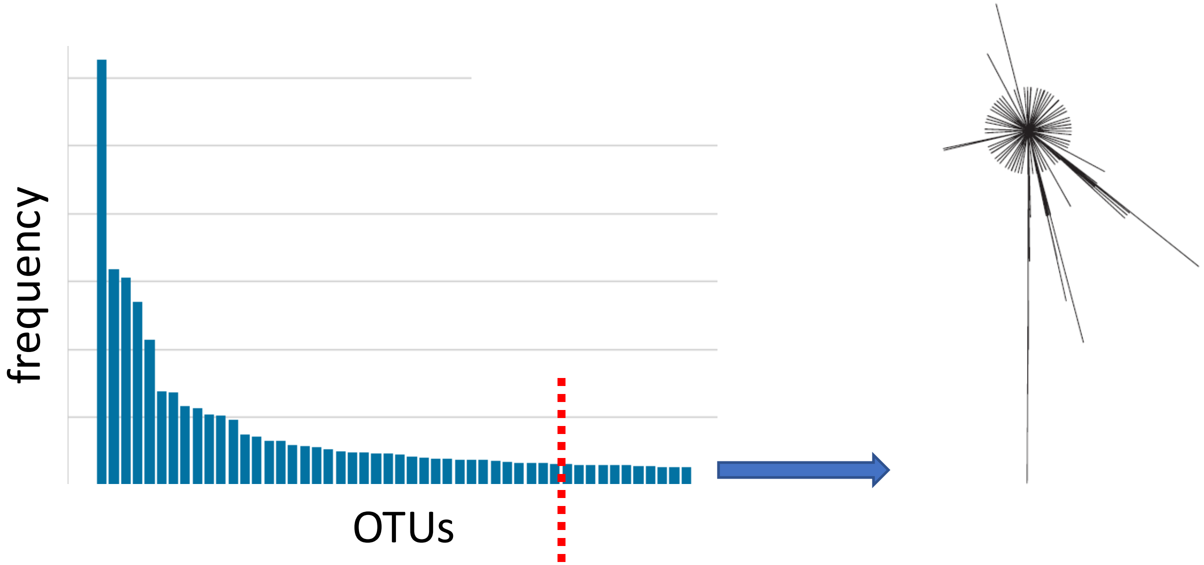
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation



What is a ‘singleton’? → Depends on the sequencing depth and quality of your data. Should also take the study aim into consideration

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

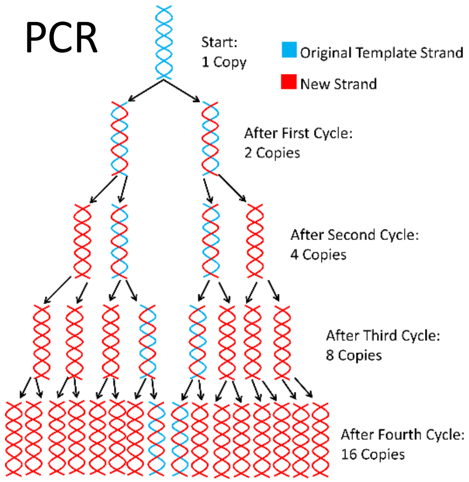
Singleton removal

Data transformation

Samples

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	500	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	30000	0
OTU10	0	0	0	0	0	0	0

PCR



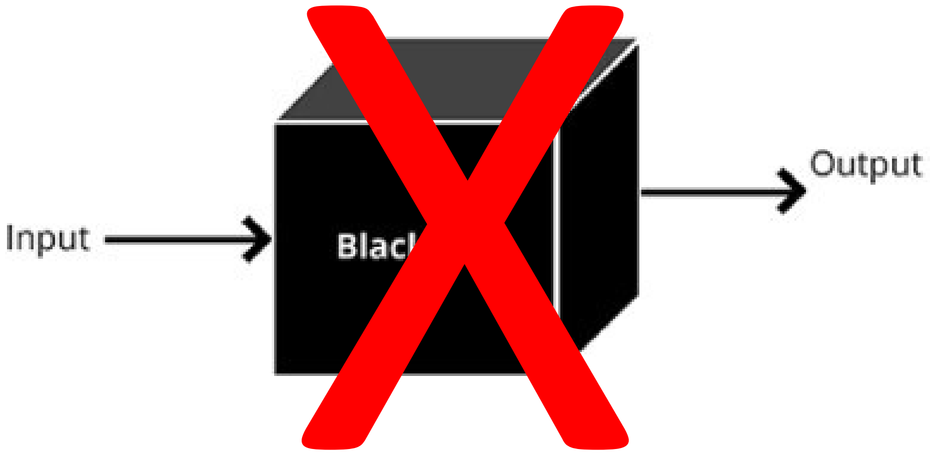
Be careful with resampling and transformations!

Check the effect from various data treatments options on the results!

Depends on the study aims!

Which methods to use???

- Which methods to use? → No general answer – it is context-dependent. You must argue for your choices!



Which methods to use?

- There are pipelines that will help with all the steps
- Qiime, LotuS
- Or you can pick different tools for different parts of the workflow to better suit your data.

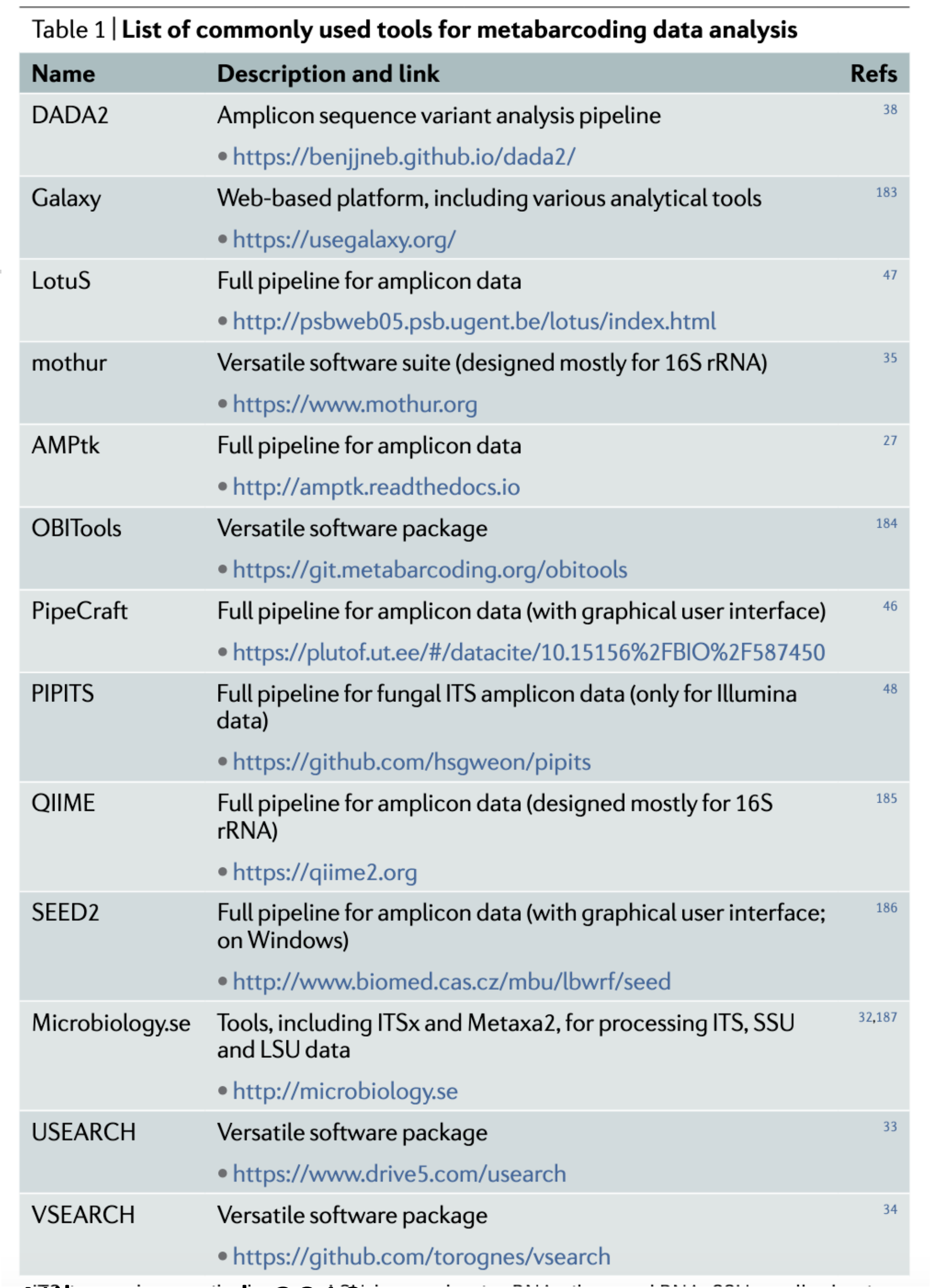


Table 1 | **List of commonly used tools for metabarcoding data analysis**

Name	Description and link	Refs
DADA2	Amplicon sequence variant analysis pipeline • https://benjjneb.github.io/dada2/	38
Galaxy	Web-based platform, including various analytical tools • https://usegalaxy.org/	183
LotuS	Full pipeline for amplicon data • http://psbweb05.psb.ugent.be/lotus/index.html	47
mothur	Versatile software suite (designed mostly for 16S rRNA) • https://www.mothur.org	35
AMPtk	Full pipeline for amplicon data • http://amptk.readthedocs.io	27
OBITools	Versatile software package • https://git.metabarcoding.org/obitools	184
PipeCraft	Full pipeline for amplicon data (with graphical user interface) • https://plutof.ut.ee/#/datacite/10.15156%2FBIO%2F587450	46
PIPITS	Full pipeline for fungal ITS amplicon data (only for Illumina data) • https://github.com/hsgweon/pipits	48
QIIME	Full pipeline for amplicon data (designed mostly for 16S rRNA) • https://qiime2.org	185
SEED2	Full pipeline for amplicon data (with graphical user interface; on Windows) • http://www.biomed.cas.cz/mbu/lbwrf/seed	186
Microbiology.se	Tools, including ITSx and Metaxa2, for processing ITS, SSU and LSU data • http://microbiology.se	32,187
USEARCH	Versatile software package • https://www.drive5.com/usearch	33
VSEARCH	Versatile software package • https://github.com/torognes/vsearch	34

Nilsson et al. 2019

Biases and sources of errors

- **Extraction bias.** no universal DNA or RNA extraction method will work for all organisms and environments.
- **Marker bias**. DNA and RNA markers differ substantially in length, taxonomic resolution, copy number and alignability.
- **Primer bias.** Primers differ in their melting temperature, binding specificity, and coverage of the targeted organisms.
- **PCR bias.** Differential amplification, mistakes in nucleotide incorporation, chimaera formation, keep amplification cycles low
- **Sequencing bias.** Differences depend strongly on the platform and sequencing
- **Bioinformatics biases.** Incorrect classification, false positive or invalid operational taxonomic units (OTUs) formation
- **Poor clustering**. Accumulation of sequencing errors leads to wrongly constructed OTUs, can lead to overestimation of richness. Use Stringent denoising, and removal of rare OTUs
- **Unequal sequencing depth**. Having different numbers of sequences between samples complicates tests of taxonomic diversity because OTU accumulation curves fail to reach a plateau in large composite samples.

Nilsson et al. 2019