CHARLES UNIVERSITY

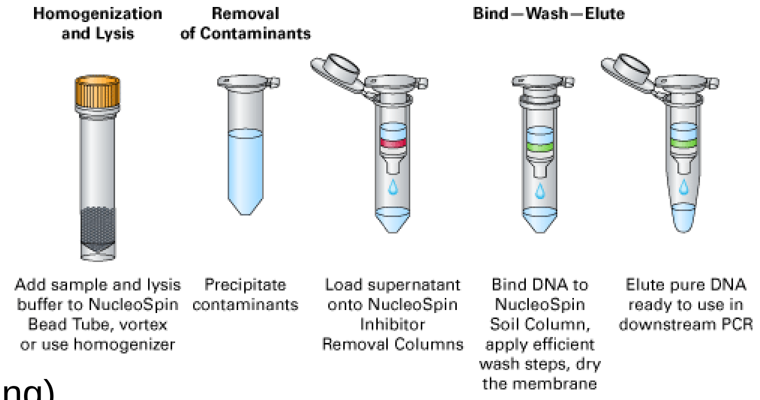
# Bioinformatics and Microbiome Analysis MB140P94

# Amplicon data

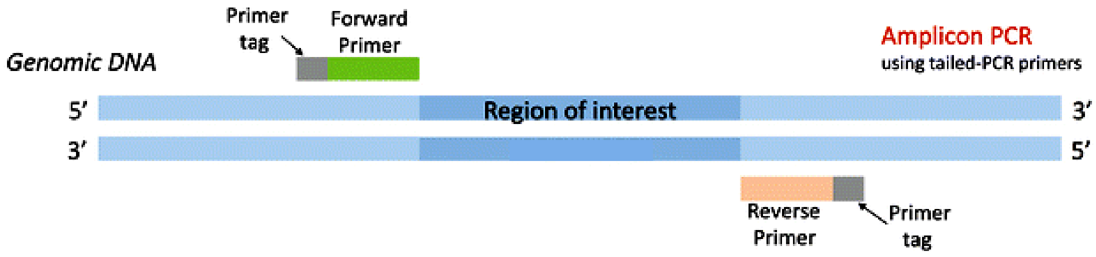
Tomáš Větrovský, Iñaki Odriozola and Petr Baldrian
Laboratory of Environmental Microbiology
Institute of Microbiology of the CAS

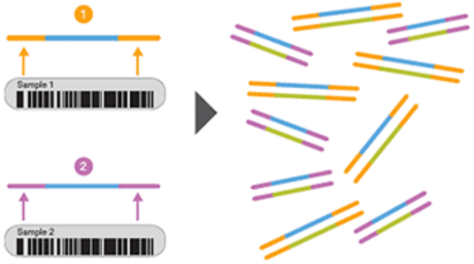
# Illumina amplicon sequence library preparation

1. DNA isolation

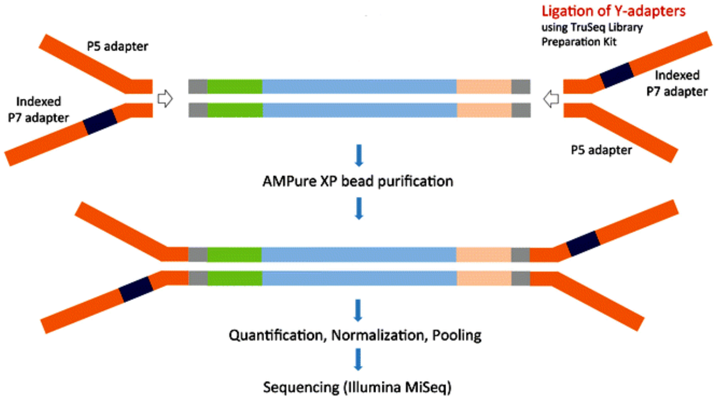

Homogenization and Lysis
Removal of Contaminants
Bind—Wash—Elute
Add sample and lysis buffer to NucleoSpin Bead Tube, vortex or use homogenizer
Precipitate contaminants
Load supernatant onto NucleoSpin Inhibitor Removal Columns
Bind DNA to NucleoSpin Soil Column, apply efficient wash steps, dry the membrane
Elute pure DNA ready to use in downstream PCR

2. PCR with barcoded primers (multiplexing)





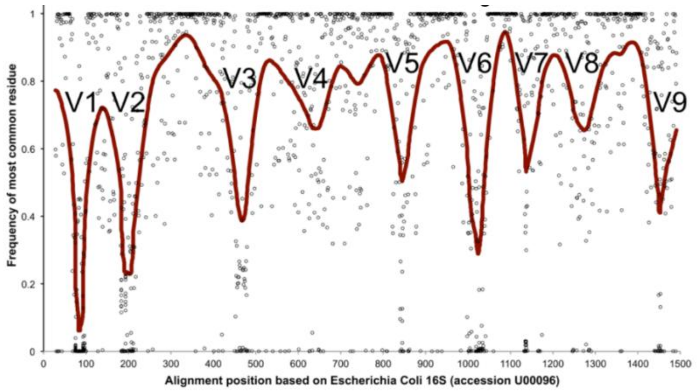
3. Ligation of sequencing adapters – to attach short oligonucleotides (60bp) to your DNA fragments, these oligonucleotides are used to attach to the sequencing flow cell and they are also used as barcode of library


Ligation of Y-adapters
using TruSeq Library Preparation Kit
P5 adapter
Indexed P7 adapter
Indexed P7 adapter
P5 adapter
AMPure XP bead purification
Quantification, Normalization, Pooling
Sequencing (Illumina MiSeq)

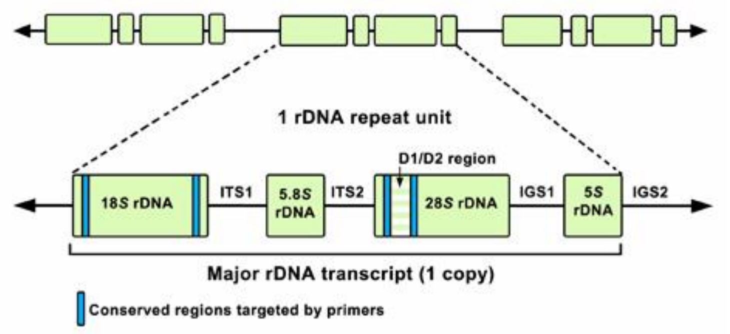
4. Quantification of the library by qPCR – to quantify of the exact amount of ligated fragments

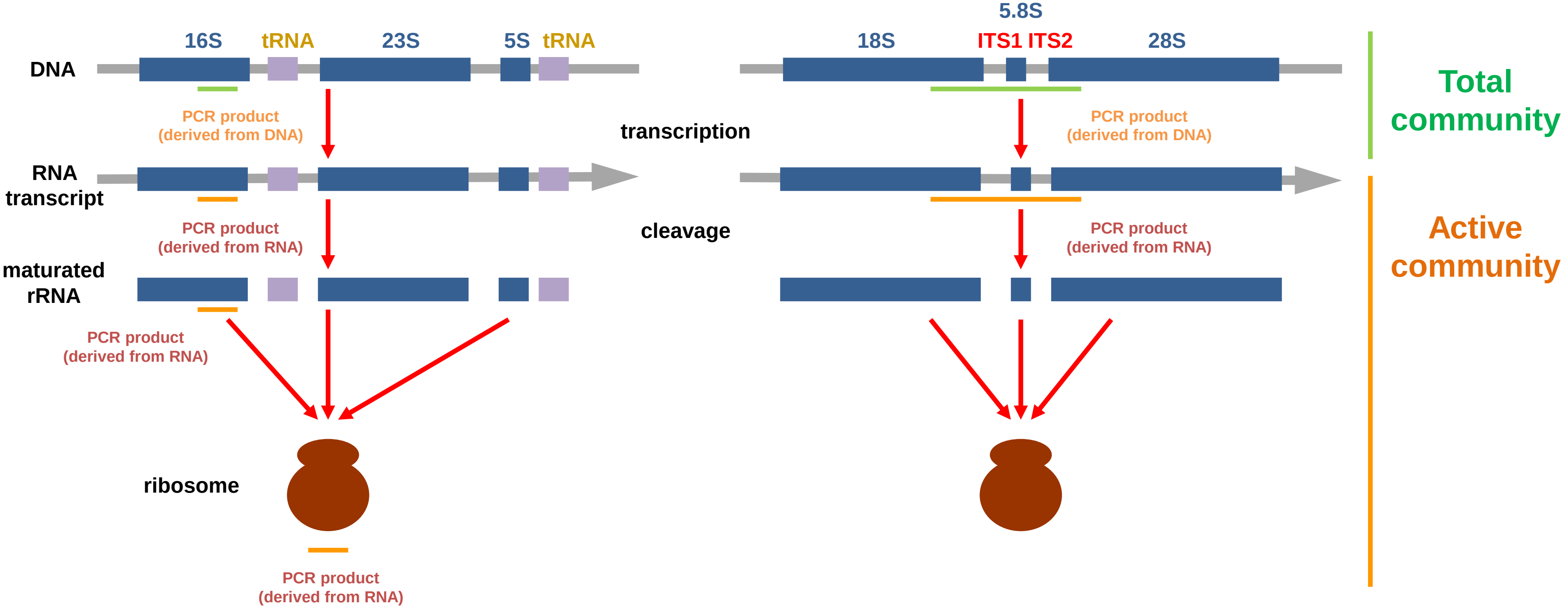
# Most used marker genes

## Bacterial 16S ribosomal RNA gene



## Fungal internal transcribed spacer (ITS)





16S tRNA 23S 5S tRNA

DNA

PCR product
(derived from DNA)

RNA transcript

PCR product
(derived from RNA)

matured rRNA

PCR product
(derived from RNA)

ribosome

PCR product
(derived from RNA)

transcription

cleavage

5.8S

18S ITS1 ITS2 28S

PCR product
(derived from DNA)

PCR product
(derived from RNA)

Total community

Active community

1-15 copies of rDNA per genome

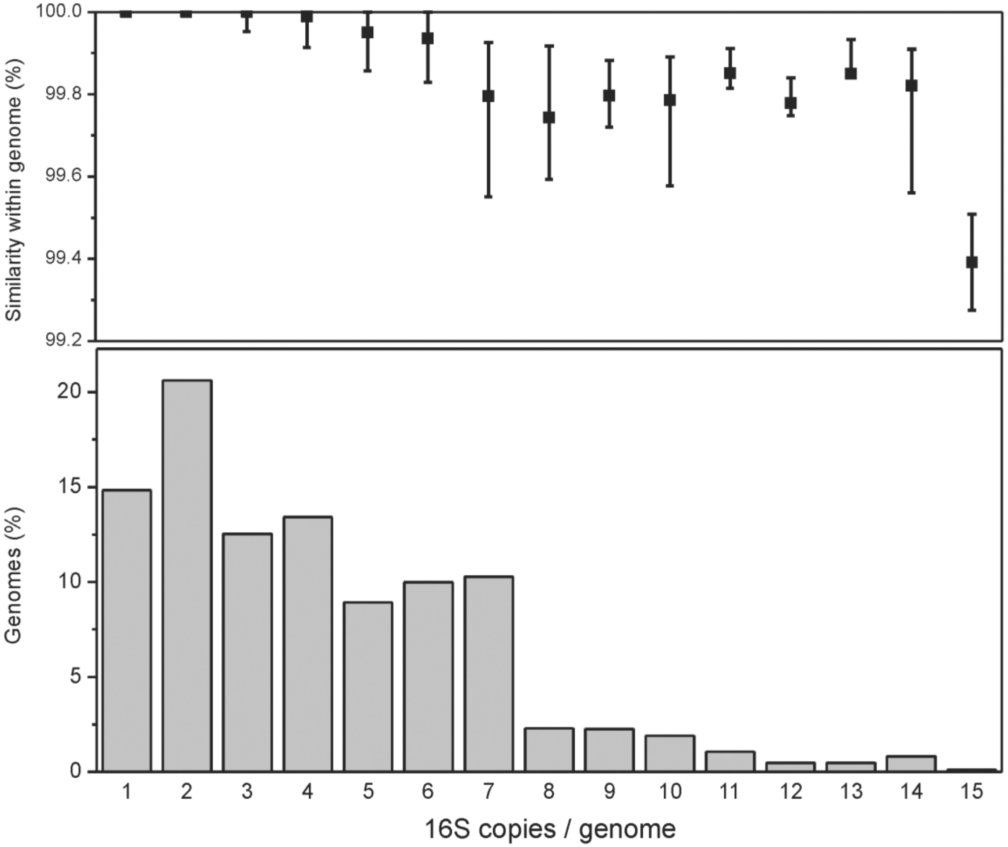
X0–X00? copies of ITS per genome

# 16S rDNA gene vs. alternative (low-copy) markers

**Pros: highly populated reference databases**

**Cons: muticopy nature of bacterial 16S rDNA gene**

- possibility of high intragenomic variability - diversity over estimation (number of OTUs)
- relative abundance estimation is skew -> normalisation by 16S copy number of closest taxon



16S rRNA within-genome similarity and copy numbers in bacterial genomes.

Upper panel: the similarity of genomes with various copy numbers: the values indicated represent the first, the second and the third quartile.

Lower panel: distribution of 16S rRNA copy numbers per genome in 1,690 sequenced bacterial genomes.

T. Větrovský & P. Baldrian - PloS one, 2013

# Sequencing platforms for amplicon sequencing (most used)

**454 Pyrosequencing**

Not supported anymore
(most studies 2009-2012)

Errors in homopolymeric regions

long reads
(up to 700 bp)

illumina

**Illumina**

The most used

Error rate less than 1%

pair-end data
(Illumina)

ion torrent
by life technologies

**IonTorrent**

Very chaep sequencing

Lot of errors

Medium read size
200-600 bp

PACIFIC BIOSCIENCES

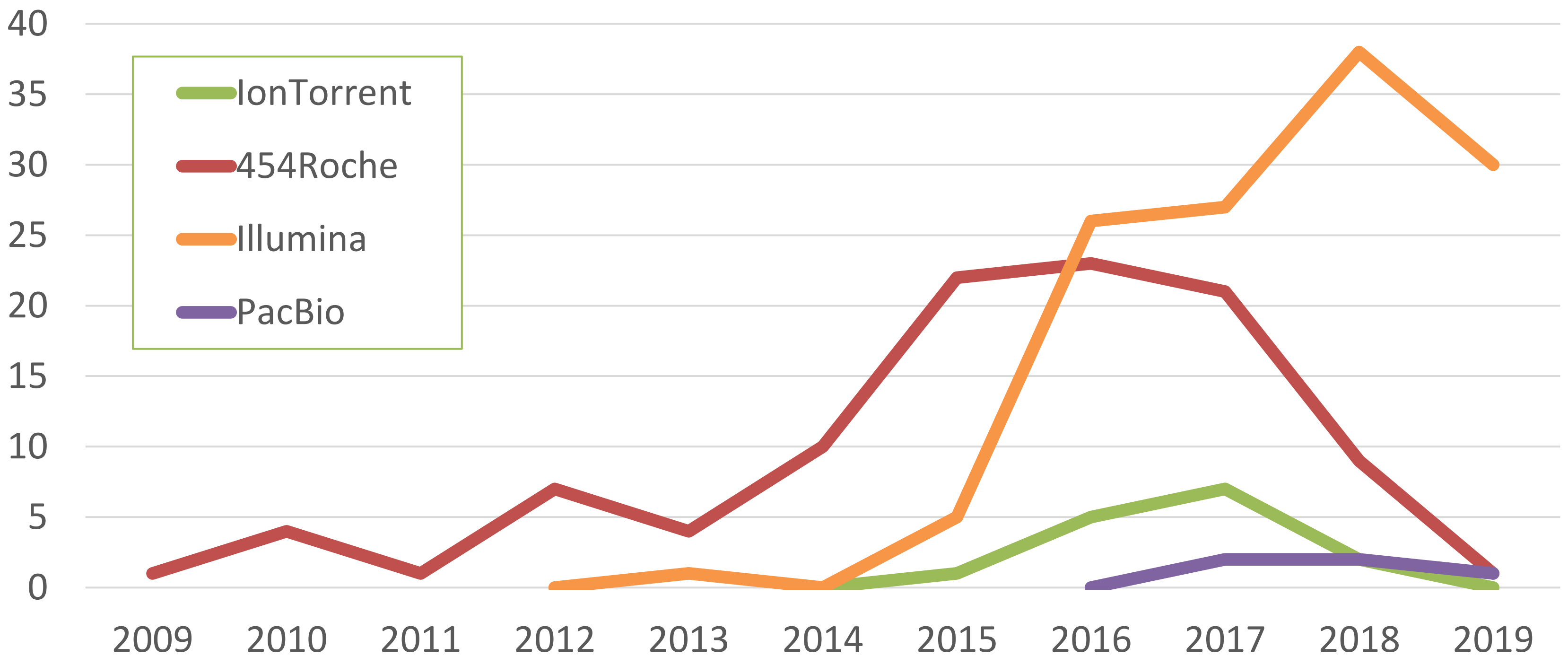
**PacBio**

Still rare

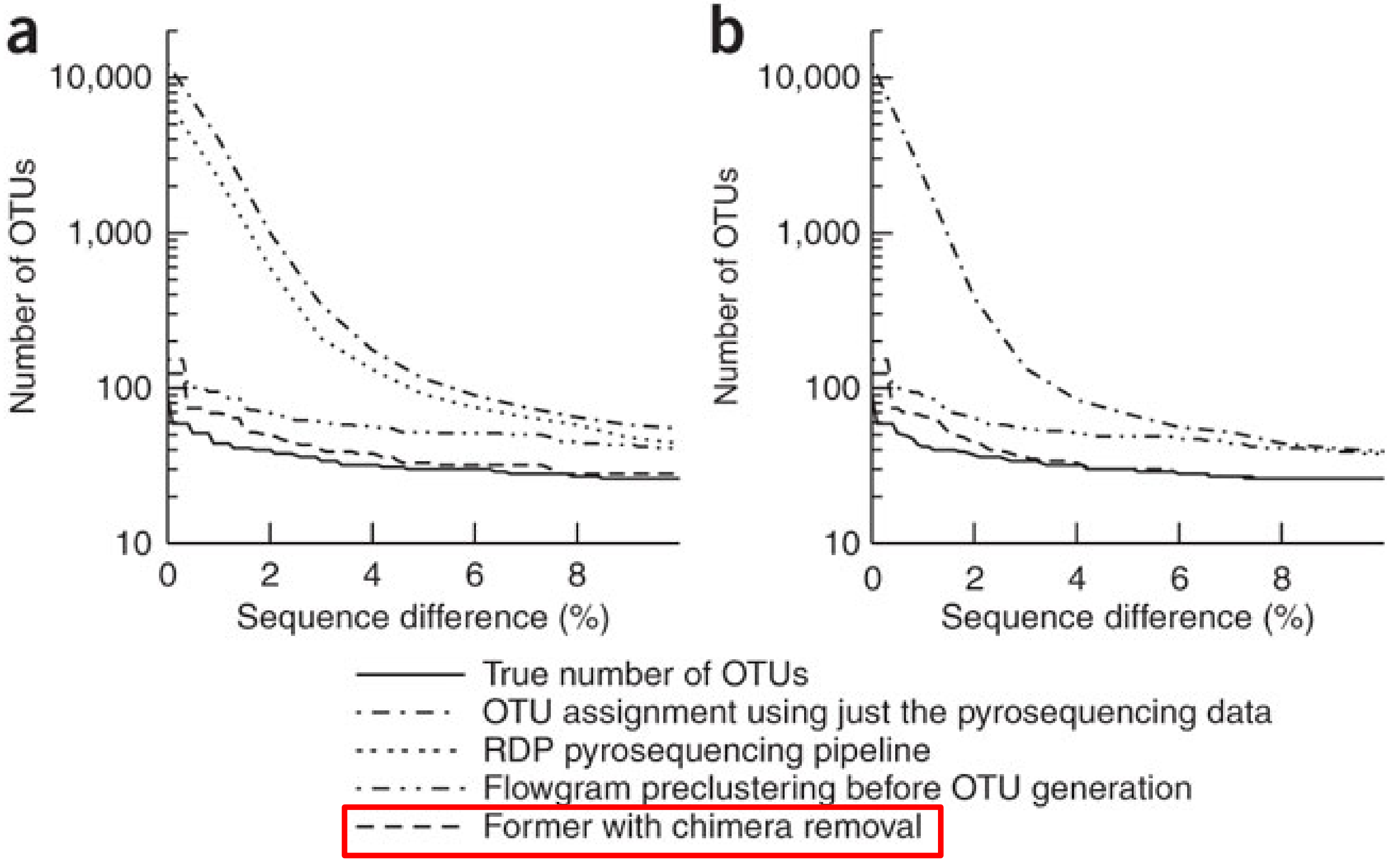
Repeated sequncing of the same region

extra long reads
(up to 10.000 bp)

# Sequencing platforms for amplicon sequencing (based on GlobalFungi data sources)

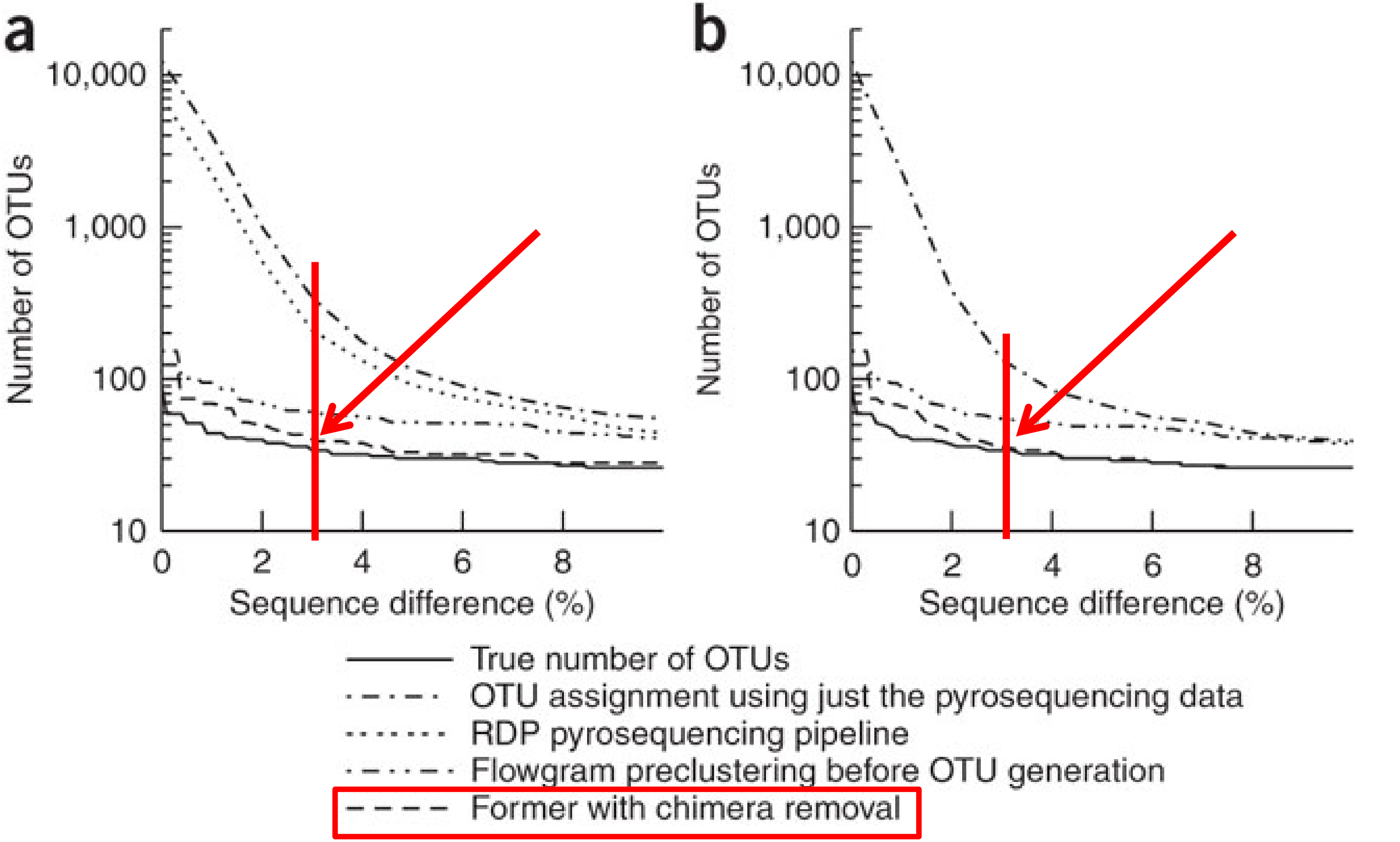


# Searching for true number of taxa (OTUs) in the data



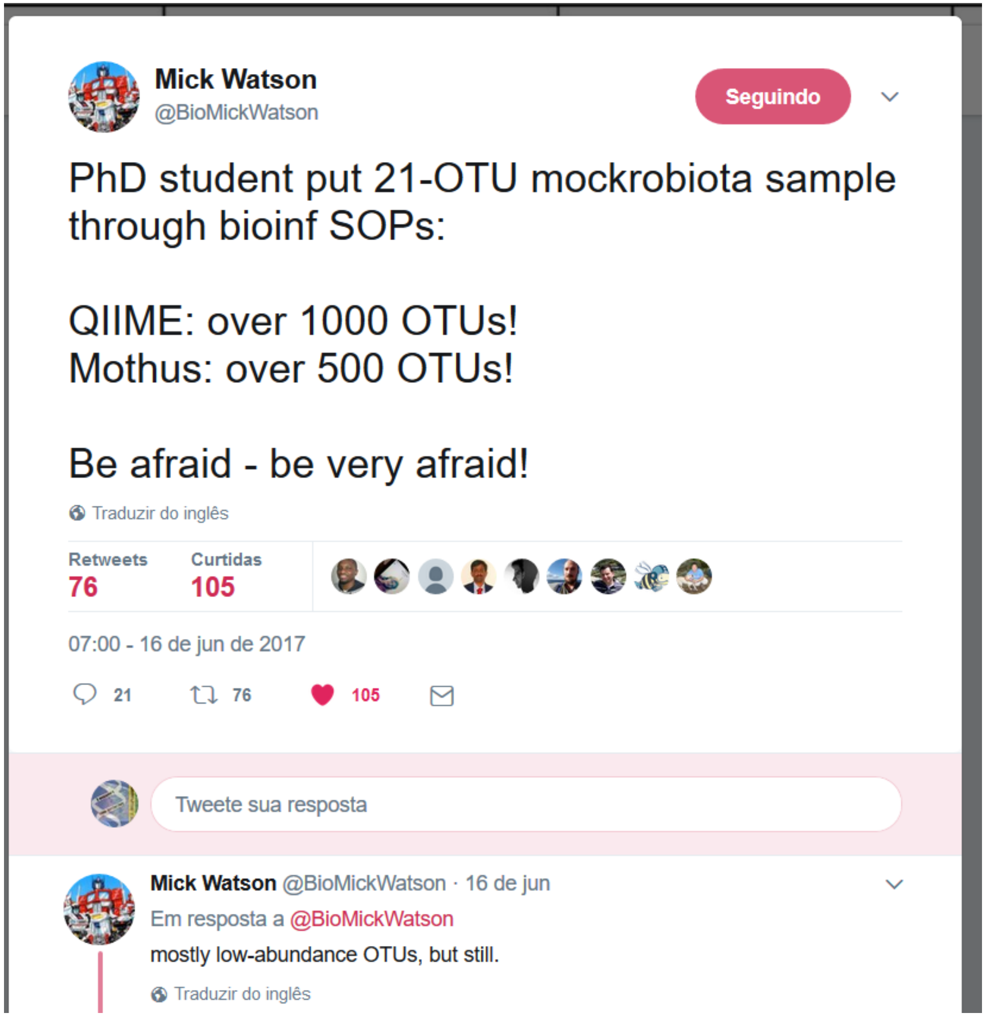
Quince, Christopher, et al. "Accurate determination of microbial diversity from 454 pyrosequencing data." *Nature methods* 6.9 (2009): 639.

# Searching for true number of taxa (OTUs) in the data



Quince, Christopher, et al. "Accurate determination of microbial diversity from 454 pyrosequencing data." *Nature methods* 6.9 (2009): 639.

This is still a big concern to all microbial community analysis pipelines!



**Mick Watson**
@BioMickWatson

PhD student put 21-OTU mockrobiota sample through bioinf SOPs:

QIIME: over 1000 OTUs!
Mothus: over 500 OTUs!

Be afraid - be very afraid!

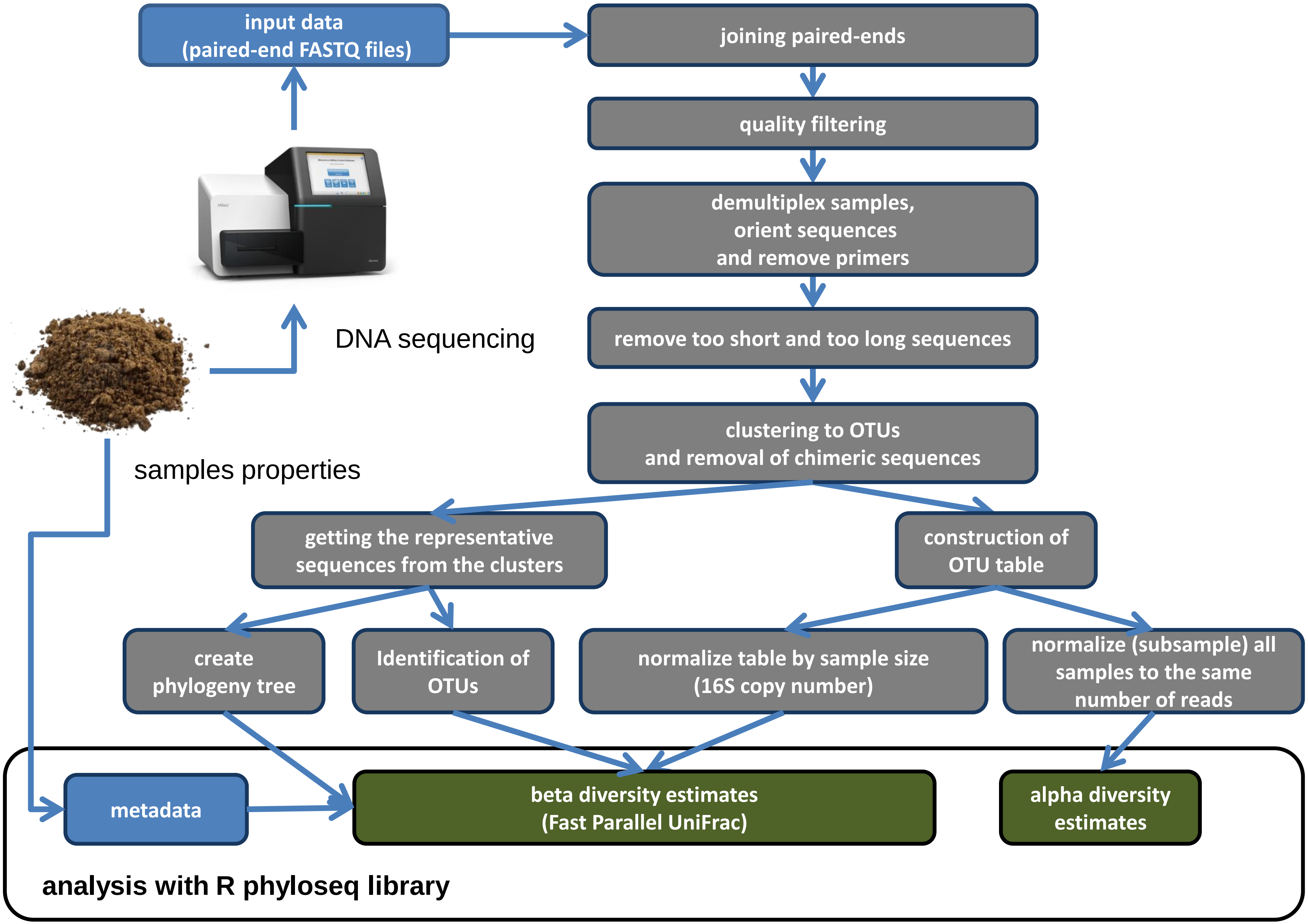
Retweets 76 Curtidas 105

07:00 - 16 de jun de 2017

21 76 105

**Mick Watson** @BioMickWatson · 16 de jun
Em resposta a @BioMickWatson
mostly low-abundance OTUs, but still.

# Amplicons pipeline workflow



input data
(paired-end FASTQ files)

DNA sequencing

samples properties

joining paired-ends

quality filtering

demultiplex samples,
orient sequences
and remove primers

remove too short and too long sequences

clustering to OTUs
and removal of chimeric sequences

getting the representative
sequences from the clusters

construction of
OTU table

create
phylogeny tree

Identification of
OTUs

normalize table by sample size
(16S copy number)

normalize (subsample) all
samples to the same
number of reads

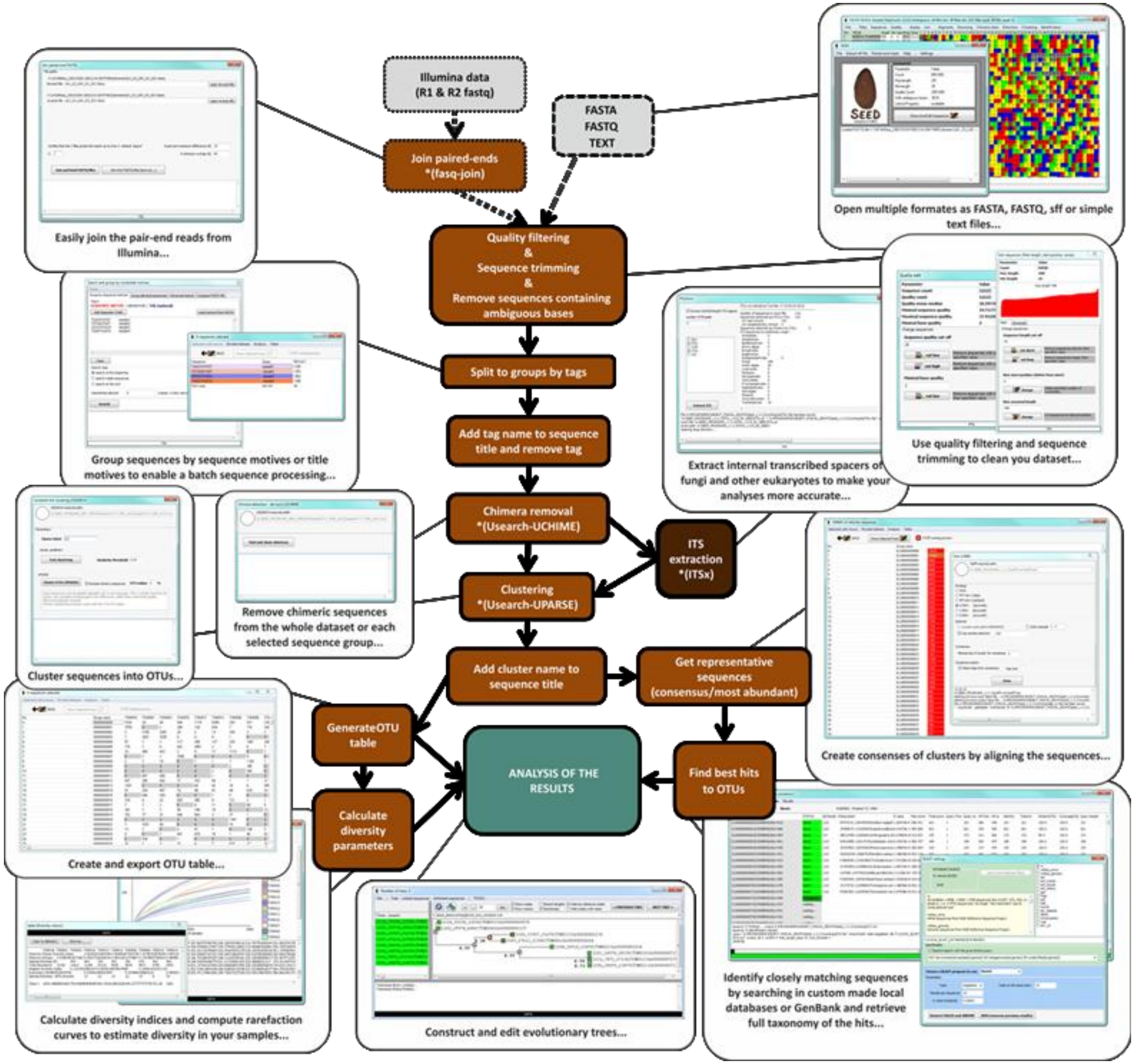
metadata

beta diversity estimates
(Fast Parallel UniFrac)

alpha diversity
estimates

analysis with R phyloseq library

# GUI based alternative for Windows (http://www.biomed.cas.cz/mbu/lbwrf/seed/)


Illumina data (R1 & R2 fastq)
FASTA FASTQ TEXT
Join paired-ends *(fasq-join)
Easily join the pair-end reads from Illumina...
Open multiple formates as FASTA, FASTQ, sff or simple text files...
SEED
Quality filtering & Sequence trimming & Remove sequences containing ambiguous bases
Split to groups by tags
Group sequences by sequence motives or title motives to enable a batch sequence processing...
Add tag name to sequence title and remove tag
Extract internal transcribed spacers of fungi and other eukaryotes to make your analyses more accurate...
Use quality filtering and sequence trimming to clean you dataset...
Chimera removal *(Usearch-UCHIME)
ITS extraction *(ITSx)
Remove chimeric sequences from the whole dataset or each selected sequence group...
Cluster sequences into OTUs...
Clustering *(Usearch-UPARSE)
Add cluster name to sequence title
Get representative sequences (consensus/most abundant)
Create consenses of clusters by aligning the sequences...
GenerateOTU table
Create and export OTU table...
ANALYSIS OF THE RESULTS
Find best hits to OTUs
Calculate diversity parameters
Calculate diversity indices and compute rarefaction curves to estimate diversity in your samples...
Construct and edit evolutionary trees...
Identify closely matching sequences by searching in custom made local databases or GenBank and retrieve full taxonomy of the hits...

Větrovský, Baldrian & Morais (2018) SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. Bioinformatics, bty071, 2018

## SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses

- editing of sequences and their titles
- sorting
- quality trimming
- pair-end joining
- grouping of sequences based on sequence motifs or sequence titles
- batch processing of sequence groups
- denoising
- chimera removal
- ITS extraction
- sequence alignments and clustering
- OTU table construction
- construction of consensus sequences
- creation of local databases for BLAST
- searching either local databases or the whole NCBI
- retrieval of taxonomical classification from the NCBI
- calculation of diversity parameters
- many more...

# SEED is alternative to

https://qiime2.org/

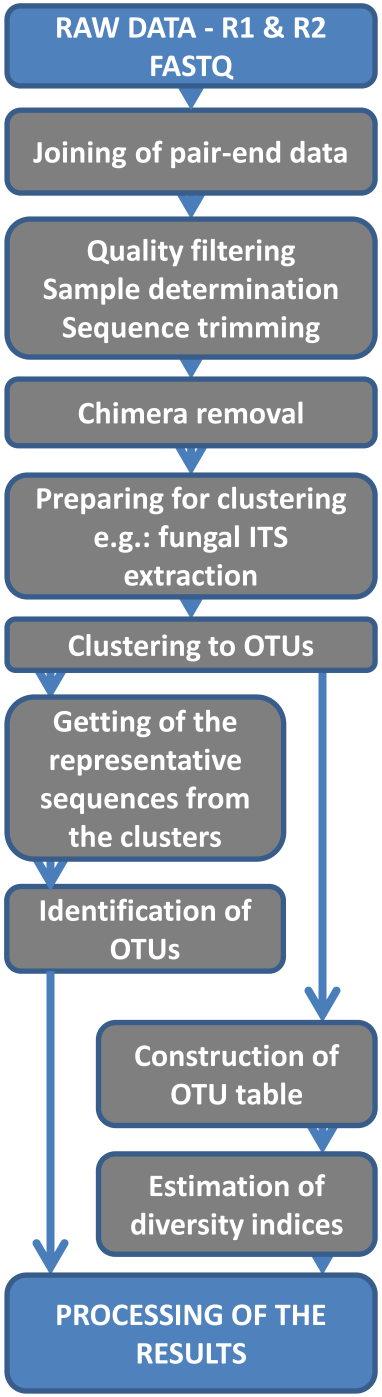
QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed.

**- command line, Unix (Linux) based**

https://mothur.org/

mothur - one of the most widely used tools for analyzing 16S rRNA gene sequence data

**- command line, multiplatform**



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering
e.g.: fungal ITS extraction

Clustering to OTUs

Getting of the representative sequences from the clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

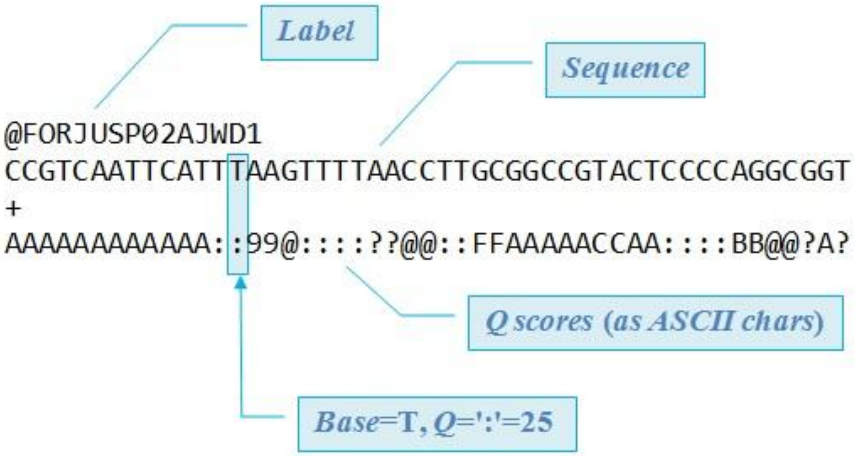
# RAW DATA

## BAC_R1.fastq - first sequence

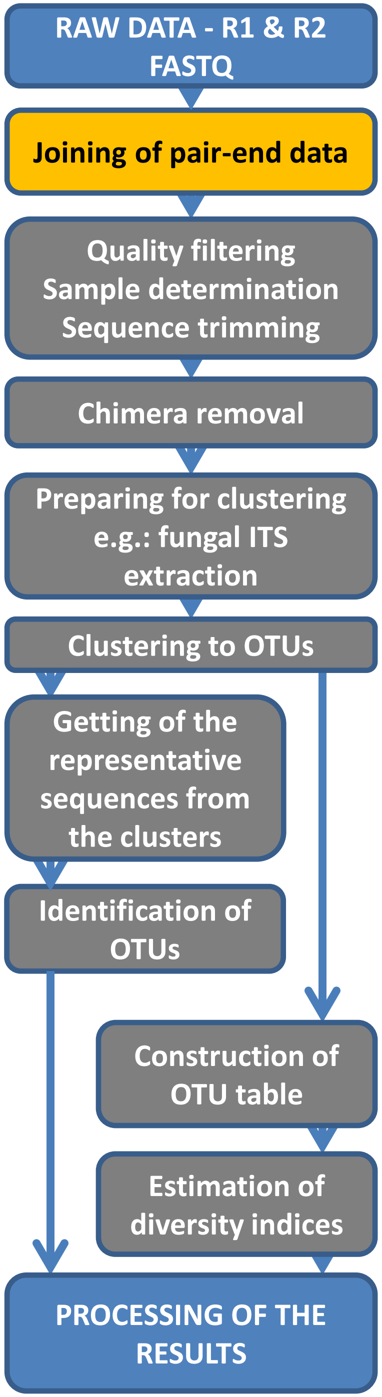
```
@M03794:8:000000000-AJCUU:1:2114:9990:17907 1:N:0:7
AACAGCCGGACTACTGGGGTTTCTAATCCTGTTTGCTCCCCACGCTTTCGTGCCTCAGTGTCAATGACCGTGTAGC
AAGCTGCCTTCGCAATTGGTGTTCTATGTCATATCTAAGCATTTCACCGCTACATGACATATTCCGCTTACCTCCAC
GATATTCAAGACTAATAGTATCAATGGCAGTTCCCAAGTTAAGCTCGGGGATTTCACCACGGACTTACTAGCCCACC
TACGCACCCTTTAAACCCAGT
+
BCCCCFCCCCCCGGGGGGGFGGHHHCHHHHHHHHHHHHHHGG2FGGGGGHGHGGHHHFHHHHHHHHHHHHHGGH
HHHHHHHHHHHHHHHHHGGGGHHHHGHGHHHHHHHHHHFHHHHHHHHHGHHHHHHHGGGGGHHHGHHHHHHHHHHGG
GGHHHHHHHGCGFDHHGGHHHHHHHHHGG>GGGHHG/GHGHHHHHHHFHFHGHHGHHG?DGGGGGGGGGGGGAC
GGGGGGGGGFGGFAAFF;@ADFFFFFFB/FFFFF;
```

## BAC_R2.fastq - first sequence

```
@M03794:8:000000000-AJCUU:1:2114:9990:17907 2:N:0:7
ACGAAGTGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTATCCGGATTCACTGGGTTTAAAGGGTGC
GTAGGTGGGCTAGTAAGTCCGTGGTGAAATCCCCGAGCTTAACTTGGGAACTGCCATTGATACTATTAGTCTTGA
ATATCGTGGAGGTAAGCGGAATATGTCATGTAGCGGTGAAATGCTTAGATATGACATAGAACACCAATTGCGAAG
GCAGCTTGCTACACGGTCATTGACACTG
+
BBBBBBBFFBFDGFFFGGGCGGCGGGHHHGDEDGGGFGGHHHGGGGGFDEE?EEGBGFHHHHGFFGGFFGGFE
FGDFDGFFEGGHHHGFEGFHFHEEFFFHHHHFHGGGFGFHHHFHHHHHHBGHHFBCHFHBGGGHHHHGHHHHH
HGFFHHHGHGD<GGAGHHHGGG@CFHHFCGHH:CCG?AAAGFEFEFGGGBFFFFFFGFFBBFGFFBDE?BBBB.@9
-9..A.B/:AFFFEF.@;AAF//99FFF/
```



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


RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting of the representative sequences from the clusters

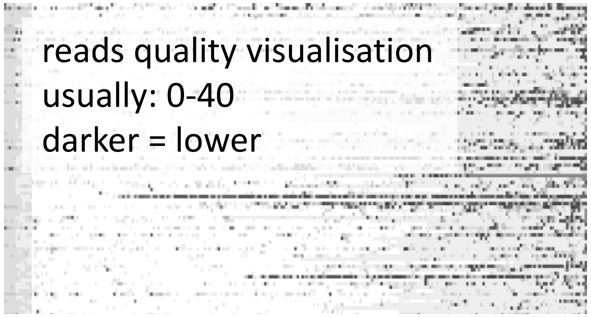
Identification of OTUs

Construction of OTU table

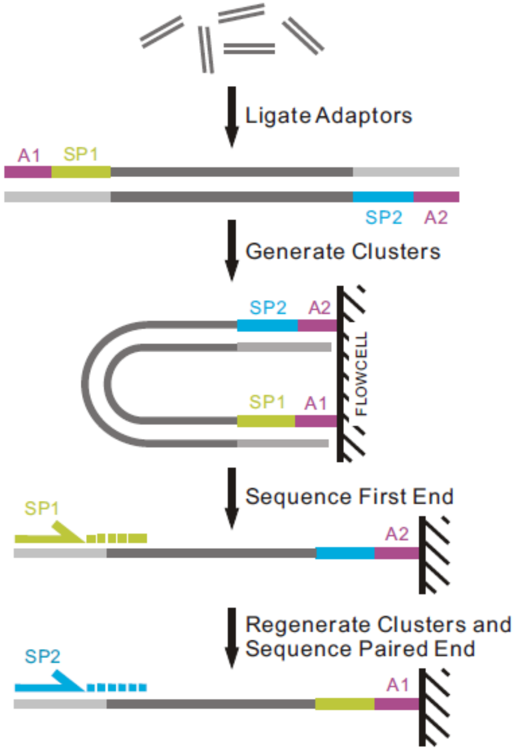
Estimation of diversity indices

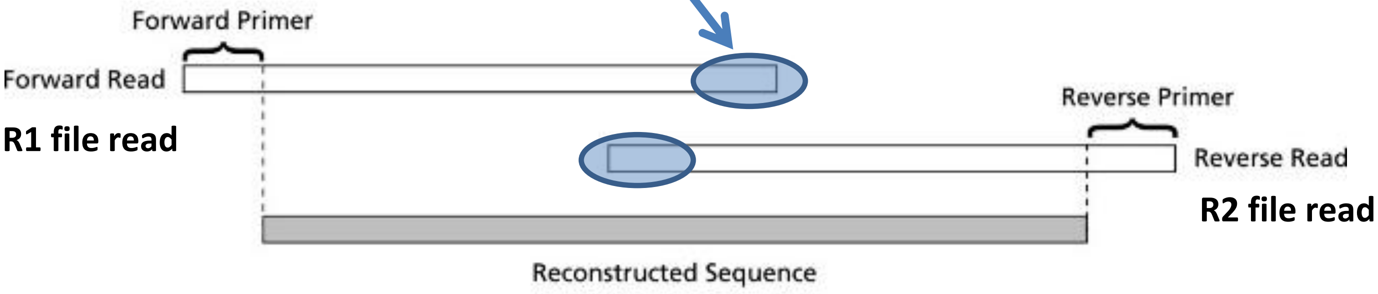
PROCESSING OF THE RESULTS

# Joining of pair-end data (Illumina)


reads quality visualisation
usually: 0-40
darker = lower

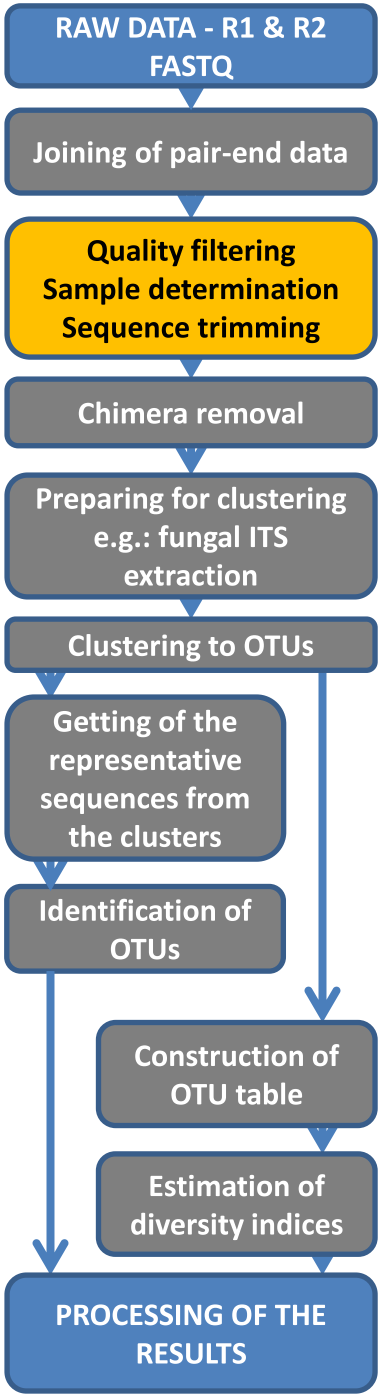
**reads quality is dropping at the ends**





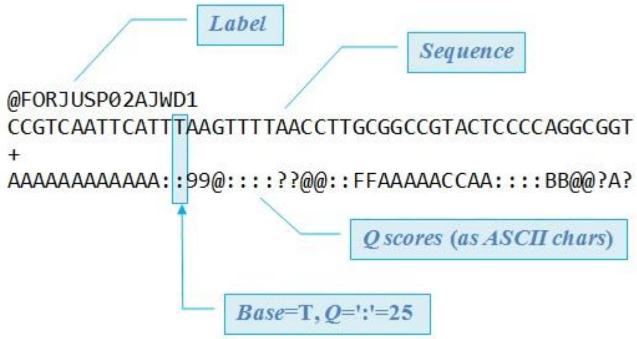
reconstructed sequence are based on bases with higher quality

PROGRAM: **FastqJoin**
set a minimal overlap length and overlap precision


RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering
Sample determination
Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting of the representative sequences from the clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

# Sequence quality and quality filtering

FASTQ format



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

$$Q_{illumina} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right),$$

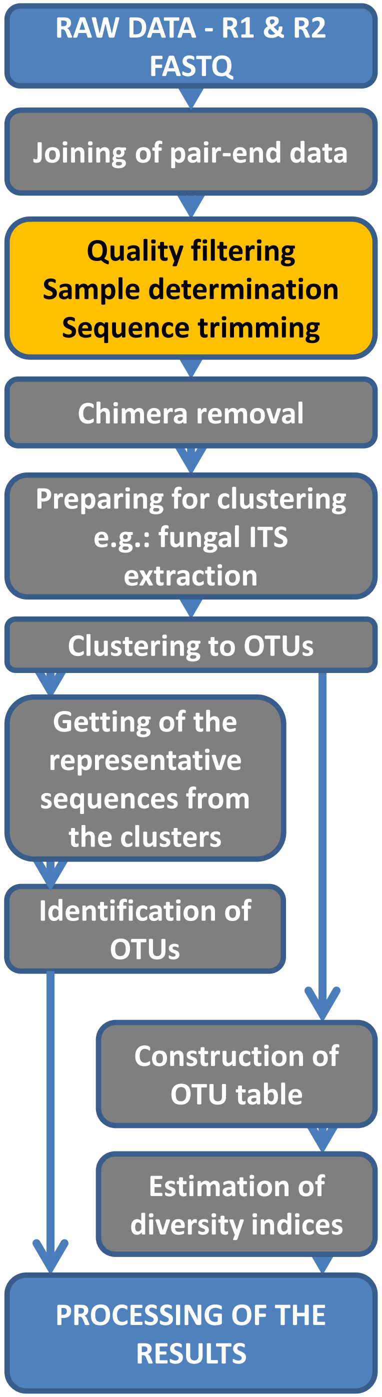
where $P_e$ is the probability of identifying a base incorrectly. For Sanger and other platforms, the formula is as follows [8]:

$$Q_{PHRED} = -10 \times \log_{10}(P_e).$$

$$Q_{illumina} = 10 \times \log_{10}\left(10^{\left\{\frac{Q_{PHRED}}{10}\right\}} + 1\right)$$

**Table 2.** Phred quality scores are logarithmically linked to error probabilities (http://en.wikipedia.org/wiki/Phred_quality_score)

| Phred quality score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.90% |
| 40 | 1 in 10 000 | 99.99% |
| 50 | 1 in 100 000 | 99.999% |
| 60 | 1 in 1 000 000 | 99.9999% |



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

**Quality filtering**
**Sample determination**
**Sequence trimming**

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting of the representative sequences from the clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

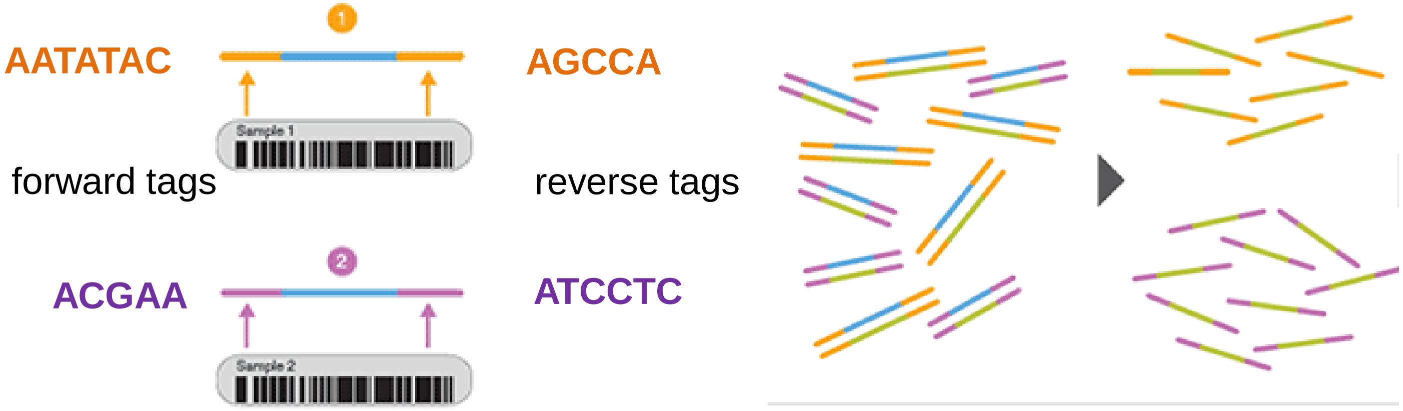
# Multiple samples in one library (Multiplexing)

| name | FWDprimer | REVprimer |
|---|---|---|
| SAMPLE001 | 515F_T103 | 806R_T007 |
| SAMPLE002 | 515F_T002 | 806R_T052 |

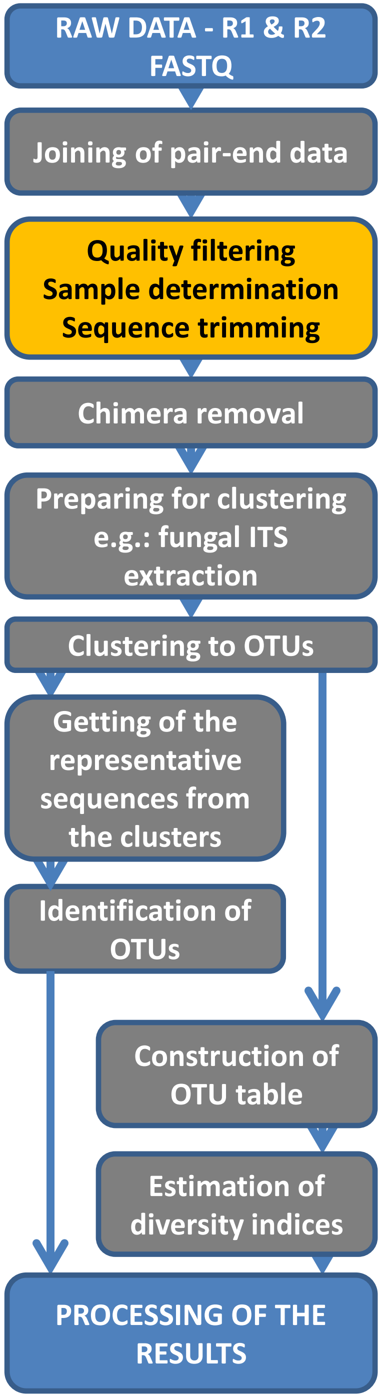
spacer is not presented in native sequences, it is used to prevent overestimation of any taxa

**TAG SPACER ORIGINAL PRIMER**

| | |
|---|---|
| 515F_T103 | AATATACGTGTGCCAGCMGCCGCGGTAA |
| 515F_T002 | ACGAAGTGTGCCAGCMGCCGCGGTAA |
| 806R_T007 | AGCCACCGGACTACHVGGGTWTCTAAT |
| 806R_T052 | ATCCTCCCGGACTACHVGGGTWTCTAAT |

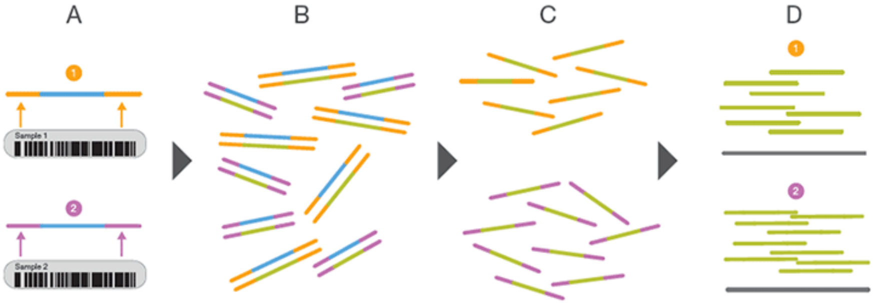


# Sample determination (de-multiplexing)


RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering
Sample determination
Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting of the representative sequences from the clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

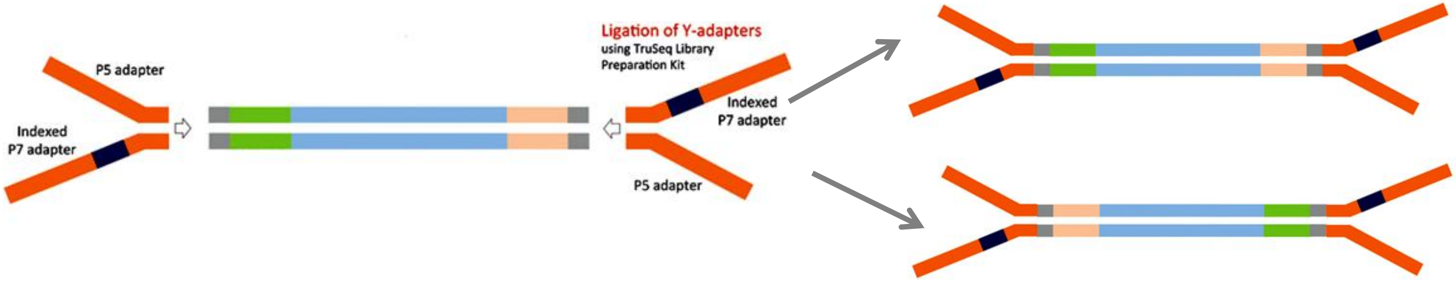
## Demultiplex samples

put sample names to sequence titles



>**SAMPLE034**|M03794:8:000000000-AJCUU:1:2114:9990:17907
CCTGTTTGCTCCCCACGCTTTC
GTGCCTCAGTGTCAATGACCGT
GTAGCAAGCTGCA…

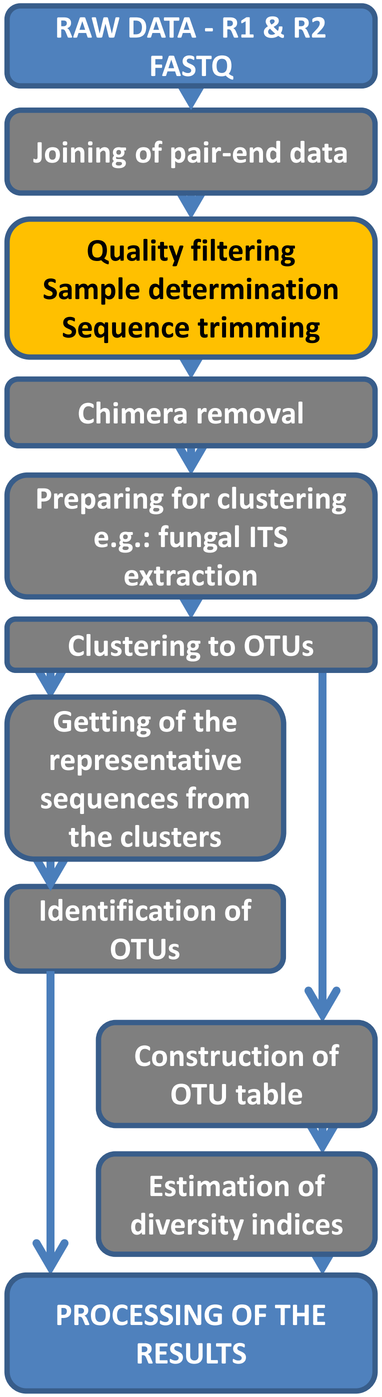
## Orient sequences



**cca 50 % of the reads are reverse complement oriented due to ligation of library adapters**

## Remove primers

since primer sequences are not native to the sample, they need to be removed before clustering to OTUs


RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

**Quality filtering**
**Sample determination**
**Sequence trimming**

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting of the representative sequences from the clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

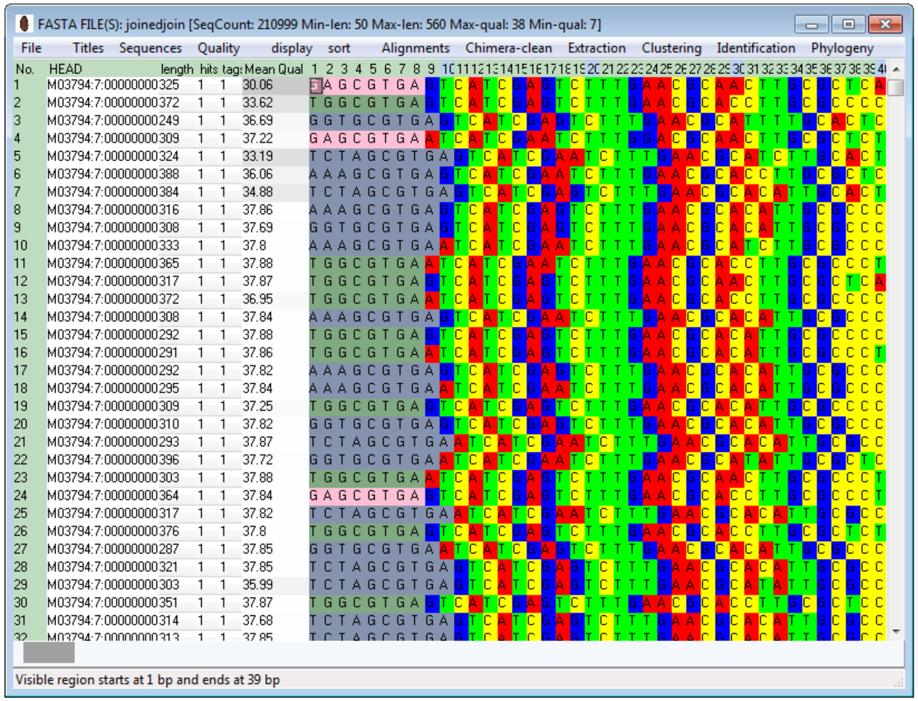
# Sample determination (de-multiplexing) and removing barcodes and primers

| | |
|---|---|
| GAGCGTGA | gITS7_T02 |
| TGGCGTGA | gITS7_T03 |
| GGTGCGTGA | gITS7_T06 |
| AAAGCGTGA | gITS7_T08 |
| TCTAGCGTGA | gITS7_T10 |

search for the sample barcodes at the beginning of reads

| Sequence | Query | RESULT |
|---|---|---|
| GAGCGTGA | gITS7_T02 | 14687 |
| TGGCGTGA | gITS7_T03 | 22568 |
| GGTGCGTGA | gITS7_T06 | 16835 |
| AAAGCGTGA | gITS7_T08 | 19258 |
| TCTAGCGTGA | gITS7_T10 | 20585 |

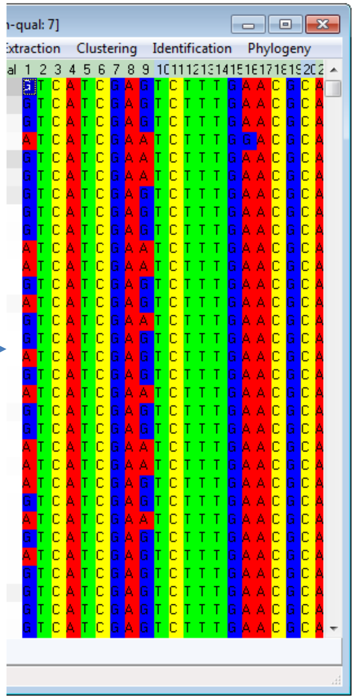
remove barcode after putting the name of sample to sequences header…



FASTA FILE(S): joinedjoin [SeqCount: 210999 Min-len: 50 Max-len: 560 Max-qual: 38 Min-qual: 7]

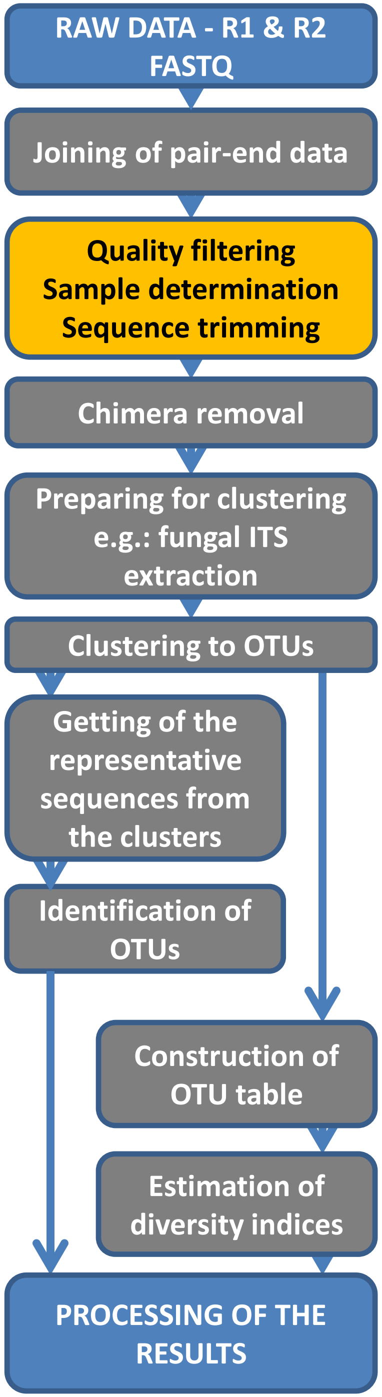
| No. | HEAD | length | hits | tags | Mean Qual | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 |
|---|---|---|---|---|---|---|
| 1 | M03794:7:00000000 | 325 | 1 | 1 | 30.06 | GAGCGTGAGTCATCGAGTCTTTGAACGCAACTTGCGCTCA |
| 2 | M03794:7:00000000 | 372 | 1 | 1 | 33.62 | TGGCGTGAGTCATCGAGTCTTTGAACGCACCTTGCGCCCC |
| 3 | M03794:7:00000000 | 249 | 1 | 1 | 36.69 | GGTGCGTGAGTCATCGAGTCTTTGAACGCATTTTGCACTC |
| 4 | M03794:7:00000000 | 309 | 1 | 1 | 37.22 | GAGCGTGAATCATCGAATCTTTGGACGCAACTTGCGCTCT |
| 5 | M03794:7:00000000 | 324 | 1 | 1 | 33.19 | TCTAGCGTGAGTCATCGAATCTTTGAACGCATCTTGCACT |
| 6 | M03794:7:00000000 | 388 | 1 | 1 | 36.06 | AAAGCGTGAGTCATCGAATCTTTGAACGCACCTTGCGCTC |
| 7 | M03794:7:00000000 | 384 | 1 | 1 | 34.88 | TCTAGCGTGAGTCATCGAGTCTTTGAACGCACATTGCACT |
| 8 | M03794:7:00000000 | 316 | 1 | 1 | 37.86 | AAAGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCCC |
| 9 | M03794:7:00000000 | 308 | 1 | 1 | 37.69 | GGTGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCCC |
| 10 | M03794:7:00000000 | 333 | 1 | 1 | 37.8 | AAAGCGTGAATCATCGAATCTTTGAACGCATCTTGCGCCC |
| 11 | M03794:7:00000000 | 365 | 1 | 1 | 37.88 | TGGCGTGAATCATCGAATCTTTGAACGCACCTTGCGCCCT |
| 12 | M03794:7:00000000 | 317 | 1 | 1 | 37.87 | TGGCGTGAGTCATCGAGTCTTTGAACGCAACTTGCGCTCA |
| 13 | M03794:7:00000000 | 372 | 1 | 1 | 36.95 | TGGCGTGAATCATCGAGTCTTTGAACGCACCTTGCGCCCC |
| 14 | M03794:7:00000000 | 308 | 1 | 1 | 37.84 | AAAGCGTGAGTCATCGAATCTTTGAACGCACATTGCGCCC |
| 15 | M03794:7:00000000 | 292 | 1 | 1 | 37.88 | TGGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCCCC |
| 16 | M03794:7:00000000 | 291 | 1 | 1 | 37.86 | TGGCGTGAATCATCGAGTCTTTGAACGCACATTGCGCCCT |
| 17 | M03794:7:00000000 | 292 | 1 | 1 | 37.82 | AAAGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCCC |
| 18 | M03794:7:00000000 | 295 | 1 | 1 | 37.84 | AAAGCGTGAATCATCGAATCTTTGAACGCACATTGCGCCC |
| 19 | M03794:7:00000000 | 309 | 1 | 1 | 37.25 | TGGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCCCC |
| 20 | M03794:7:00000000 | 310 | 1 | 1 | 37.82 | GGTGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCCC |
| 21 | M03794:7:00000000 | 293 | 1 | 1 | 37.87 | TCTAGCGTGAATCATCGAATCTTTGAACGCACATTGCGCC |
| 22 | M03794:7:00000000 | 396 | 1 | 1 | 37.72 | GGTGCGTGAATCATCGAATCTTTGAACGCATATTGCGCTC |
| 23 | M03794:7:00000000 | 303 | 1 | 1 | 37.88 | TGGCGTGAATCATCGAGTCTTTGAACGCAACTTGCGCCCT |
| 24 | M03794:7:00000000 | 364 | 1 | 1 | 37.84 | GAGCGTGAGTCATCGAGTCTTTGAACGCACCTTGCGCCCT |
| 25 | M03794:7:00000000 | 317 | 1 | 1 | 37.82 | TCTAGCGTGAATCATCGAATCTTTGAACGCACATTGCGCC |
| 26 | M03794:7:00000000 | 376 | 1 | 1 | 37.8 | TGGCGTGAGTCATCGAGTCTTTGAACGCACCTTGCGCTCT |
| 27 | M03794:7:00000000 | 287 | 1 | 1 | 37.85 | GGTGCGTGAATCATCGAGTCTTTGAACGCACATTGCGCCC |
| 28 | M03794:7:00000000 | 321 | 1 | 1 | 37.85 | TCTAGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCC |
| 29 | M03794:7:00000000 | 303 | 1 | 1 | 35.99 | TCTAGCGTGAGTCATCGAGTCTTTGAACGCATATTGCGCC |
| 30 | M03794:7:00000000 | 351 | 1 | 1 | 37.87 | TGGCGTGAGTCATCGAGTCTTTGAACGCACCTTGCGCTCC |
| 31 | M03794:7:00000000 | 314 | 1 | 1 | 37.68 | TCTAGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCC |
| 32 | M03794:7:00000000 | 313 | 1 | 1 | 37.85 | TCTAGCGTGAGTCATCGAGTCTTTGAACGCACATTGCGCC |

Visible region starts at 1 bp and ends at 39 bp



-qual: 7]

```
al 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAATCTTTGGACGCA
   GTCATCGAATCTTTGAACGCA
   GTCATCGAATCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAATCTTTGAACGCA
   ATCATCGAATCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAGTCTTTGAACGCA
   GTCATCGAATCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAATCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAATCTTTGAACGCA
   ATCATCGAATCTTTGAACGCA
   ATCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAATCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   ATCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
   GTCATCGAGTCTTTGAACGCA
```



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting of the representative sequences from the clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

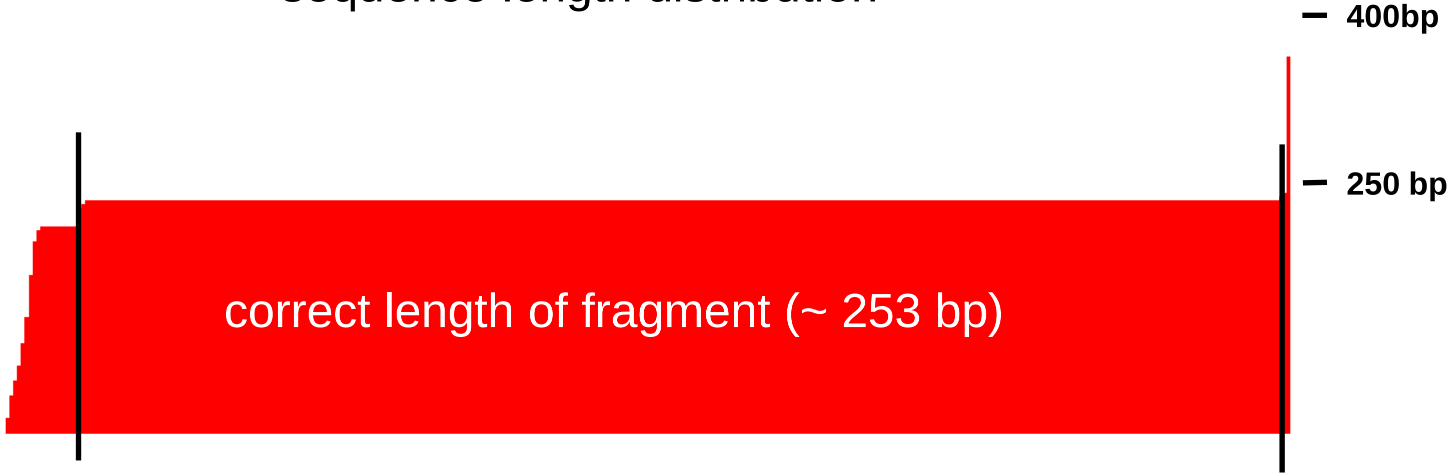
PROCESSING OF THE RESULTS

# Sequence trimming

removing sequences with aberrated length - depends on the marker gene

Too short – nonspecific PCR products/erroneus sequences
Too long - nonspecific PCR products/chimeric sequences

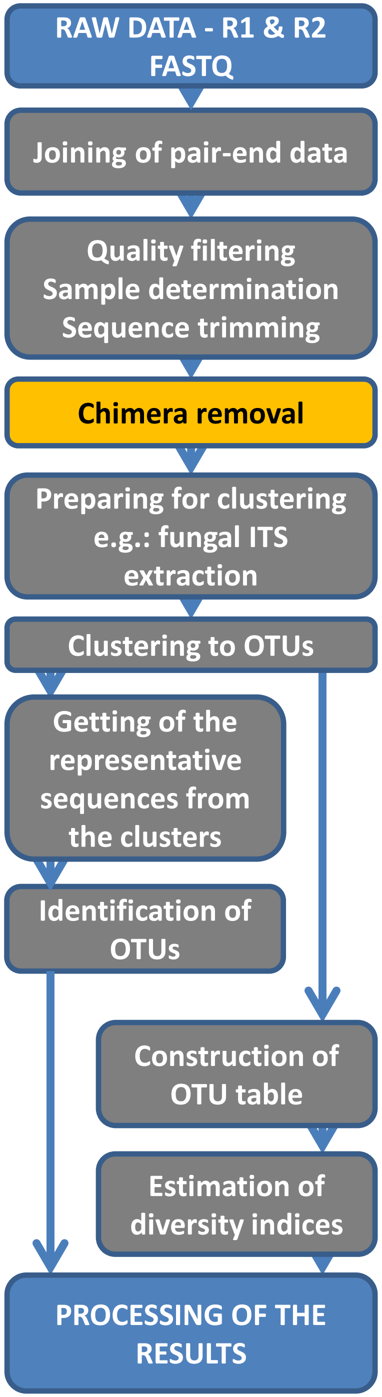
sequence length distribution



chloroplasts
or
mitochondria

probably
chimeric

# Chimera removal



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

**Chimera removal**

Preparing for clustering
e.g.: fungal ITS extraction

Clustering to OTUs

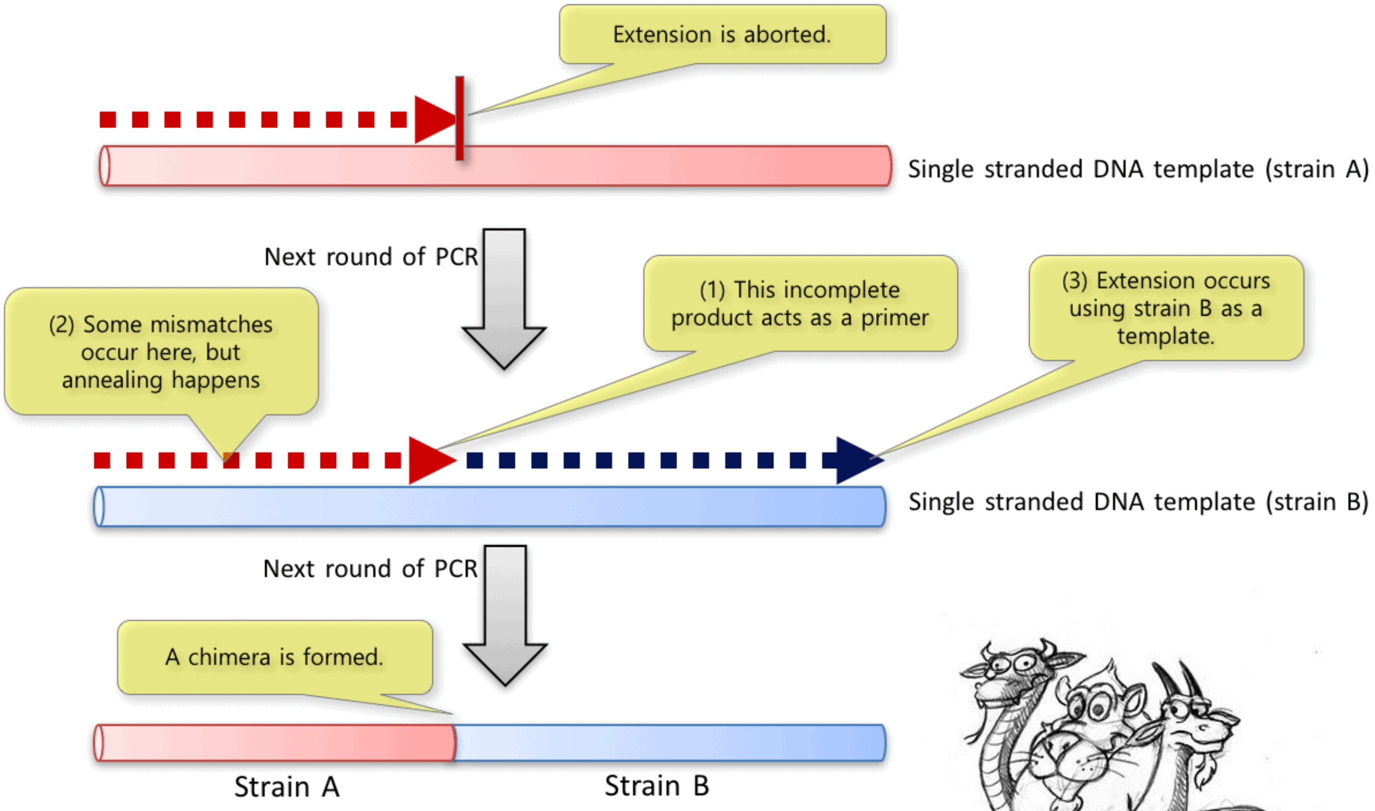
Getting of the representative sequences from the clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS



Extension is aborted.

Single stranded DNA template (strain A)

Next round of PCR

(2) Some mismatches occur here, but annealing happens

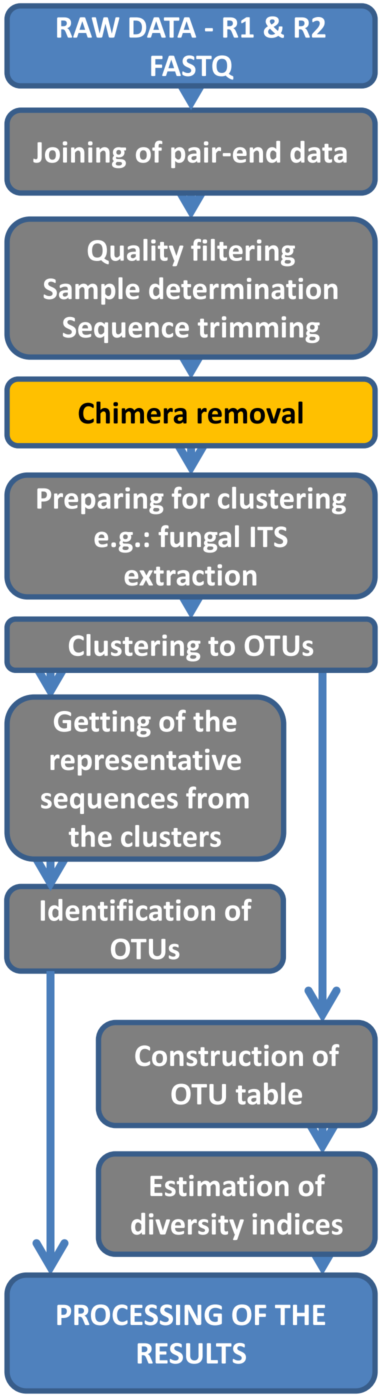
(1) This incomplete product acts as a primer

(3) Extension occurs using strain B as a template.

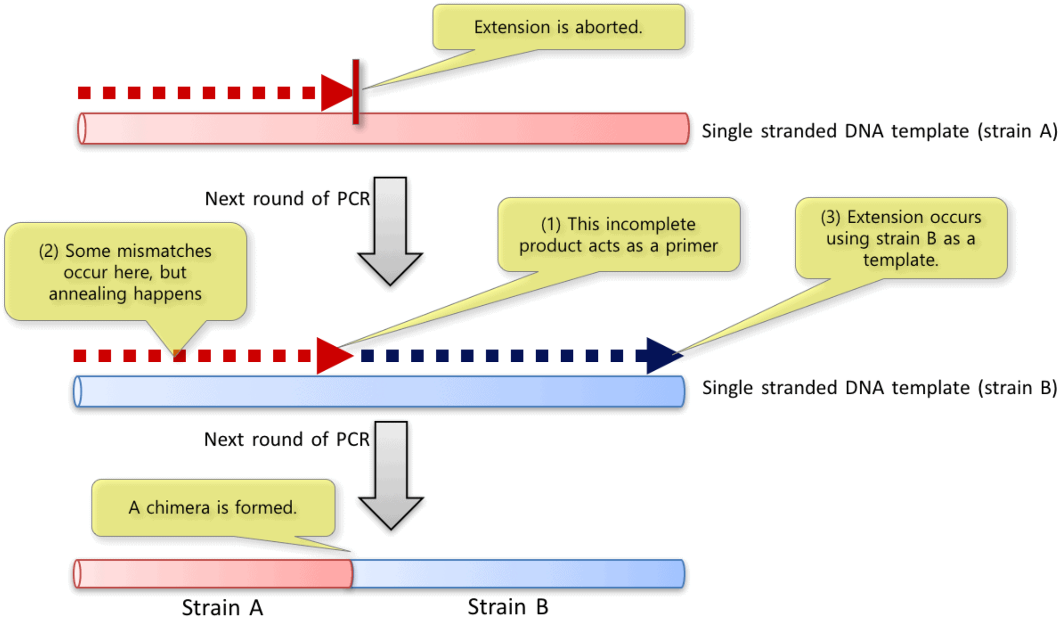
Single stranded DNA template (strain B)

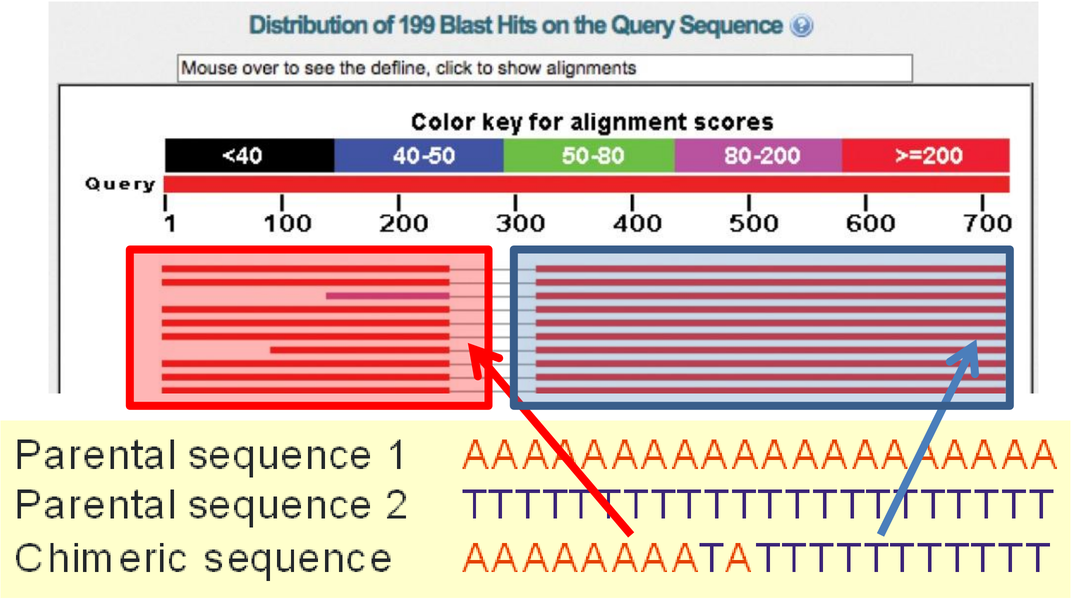
Next round of PCR

A chimera is formed.

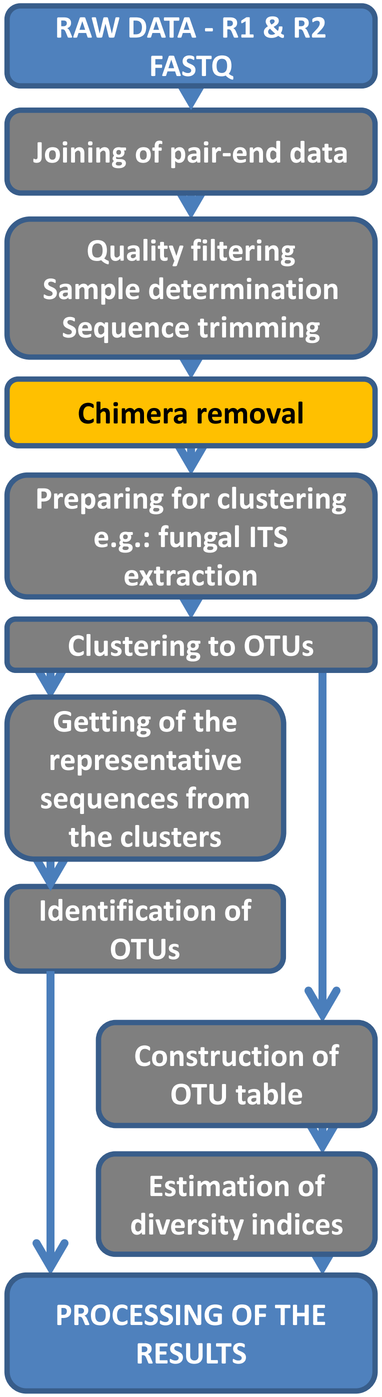

RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering
Sample determination
Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting of the representative sequences from the clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

# Chimera removal


Extension is aborted.
Single stranded DNA template (strain A)
Next round of PCR
(2) Some mismatches occur here, but annealing happens
(1) This incomplete product acts as a primer
(3) Extension occurs using strain B as a template.
Single stranded DNA template (strain B)
Next round of PCR
A chimera is formed.
Strain A
Strain B


Distribution of 199 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600
700

| | |
|---|---|
| Parental sequence 1 | AAAAAAAAAAAAAAAAAAAA |
| Parental sequence 2 | TTTTTTTTTTTTTTTTTTTTTT |
| Chimeric sequence | AAAAAAAATATTTTTTTTTTT |


RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering Sample determination Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting of the representative sequences from the clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

# Chimera removal

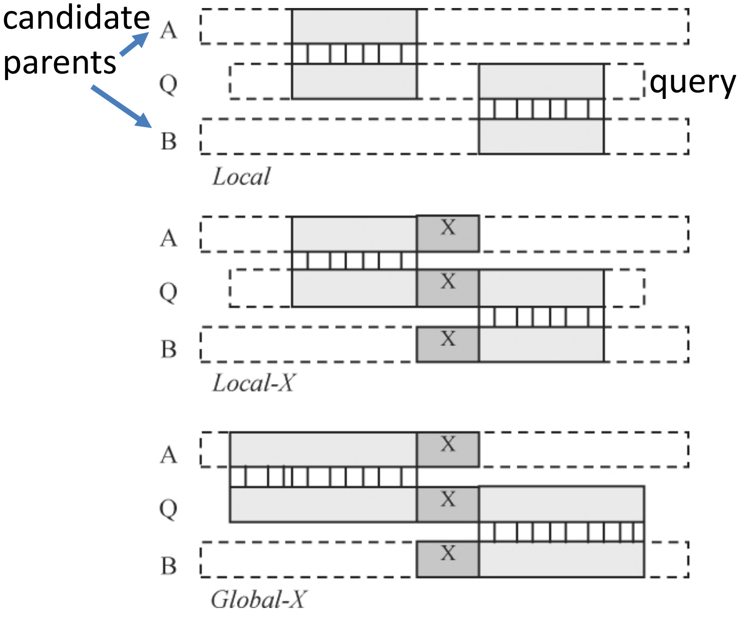
**chimeric sequences are derived from parental sequences -> program search for parents.**

## *de novo* approach:

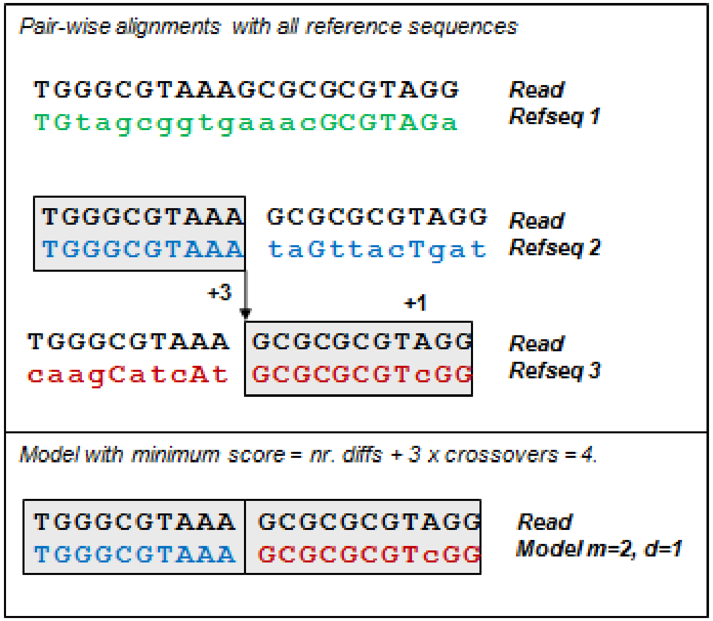
- no reference database
- usually based on measuring of several sequence parts abundances – parents should be more abundant then its offspring (chimeras)

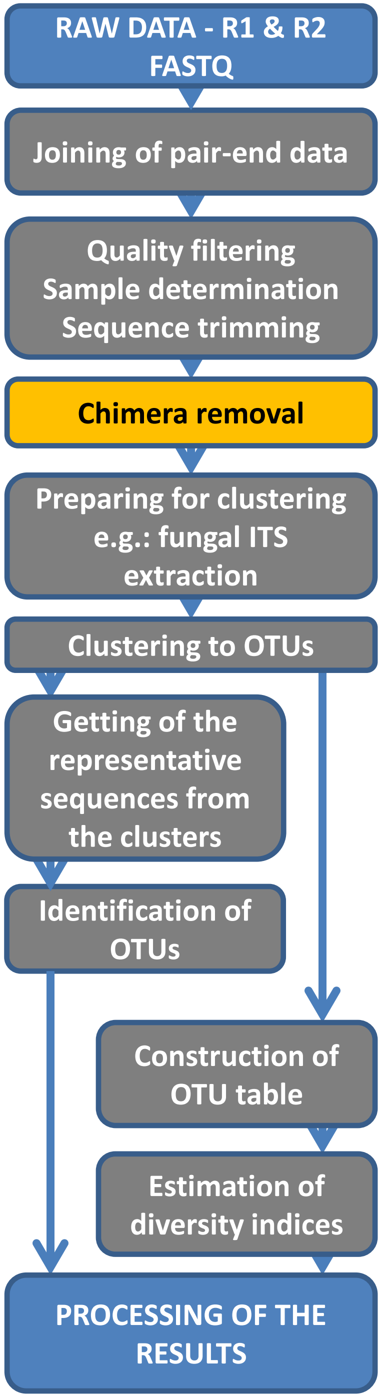
## reference based approach:

- pair-wise alignment with reference sequence database



PROGRAMS: **UCHIME, UPARSE**




RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering
Sample determination
Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting of the representative sequences from the clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

# Chimera removal

**chimeric sequences are derived from parental sequences -> program search for parents.**

## *de novo* approach:

- no reference database
- usually based on measuring of several sequence parts abundances – parents should be more abundant then its offspring (chimeras)

## Problems

- algorithms are not optimal
- computation cost could be high
- problems with highly similar sequences

## reference based approach:

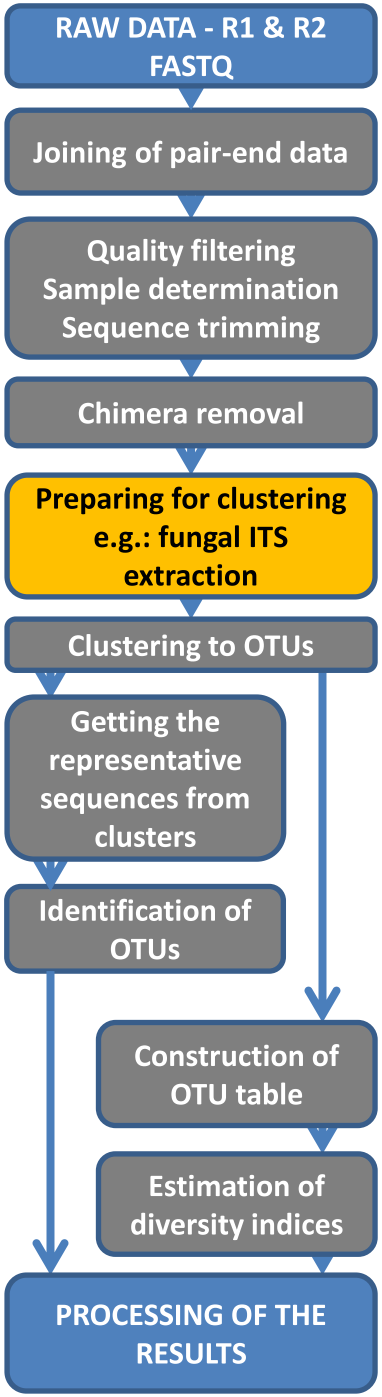
- pair-wise alignment with reference sequence database

## Problems

- no appropriate reference database for environmental samples
- variable quality of reference databases

PROGRAMS: **UCHIME, UPARSE**

# Fungal ITS extraction



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

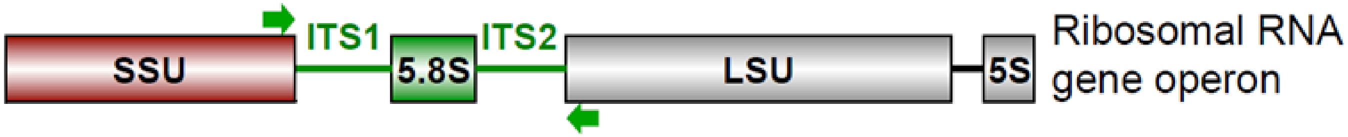
Getting the representative sequences from clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS



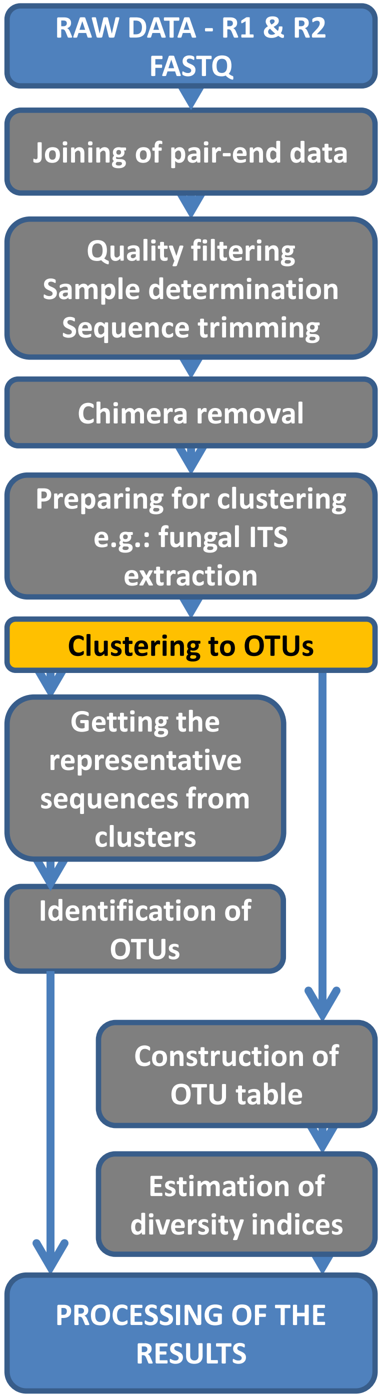
**ITSx**
Improved software for detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for use in environmental sequencing

relies on **HMMER**
searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

using the extracted variable ITS improves resolution when the OTUs are created

**http://microbiology.se/software/itsx/**

# Clustering to OTUs


RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

**OTU (Operational taxonomic unit)** group of similar sequences grouped based on some similarity threshold
usually 97% similarity (16S, ITS) represents Species

## Heuristic

- comparison of each sequence with representative sequence („seed“)
- depends on sequence order

FAST

## Hierarchical

comparison of each sequence with each other (tree contruction)

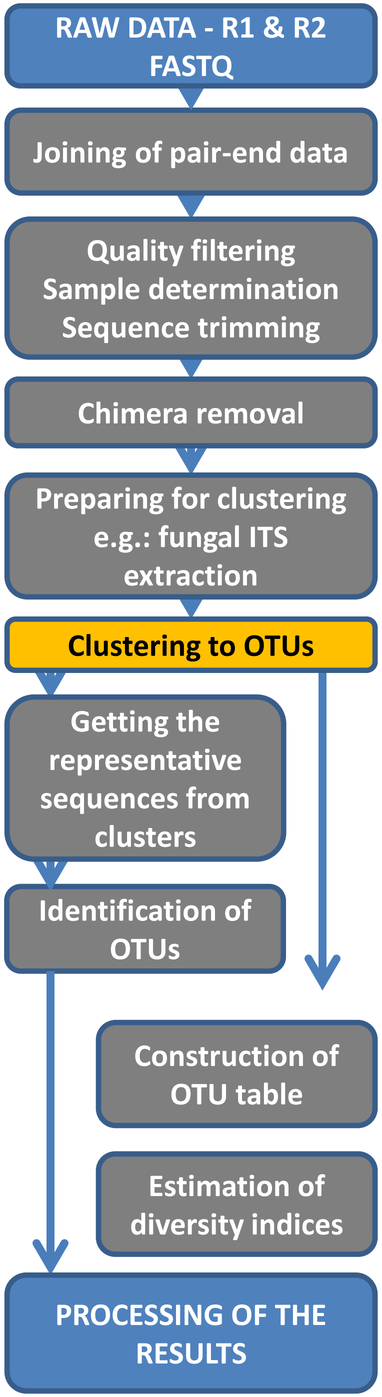
SLOW

## Model based

- probabilistic, iterative
- uses more information than the sequence identity

VERY SLOW

# Clustering to OTUs



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

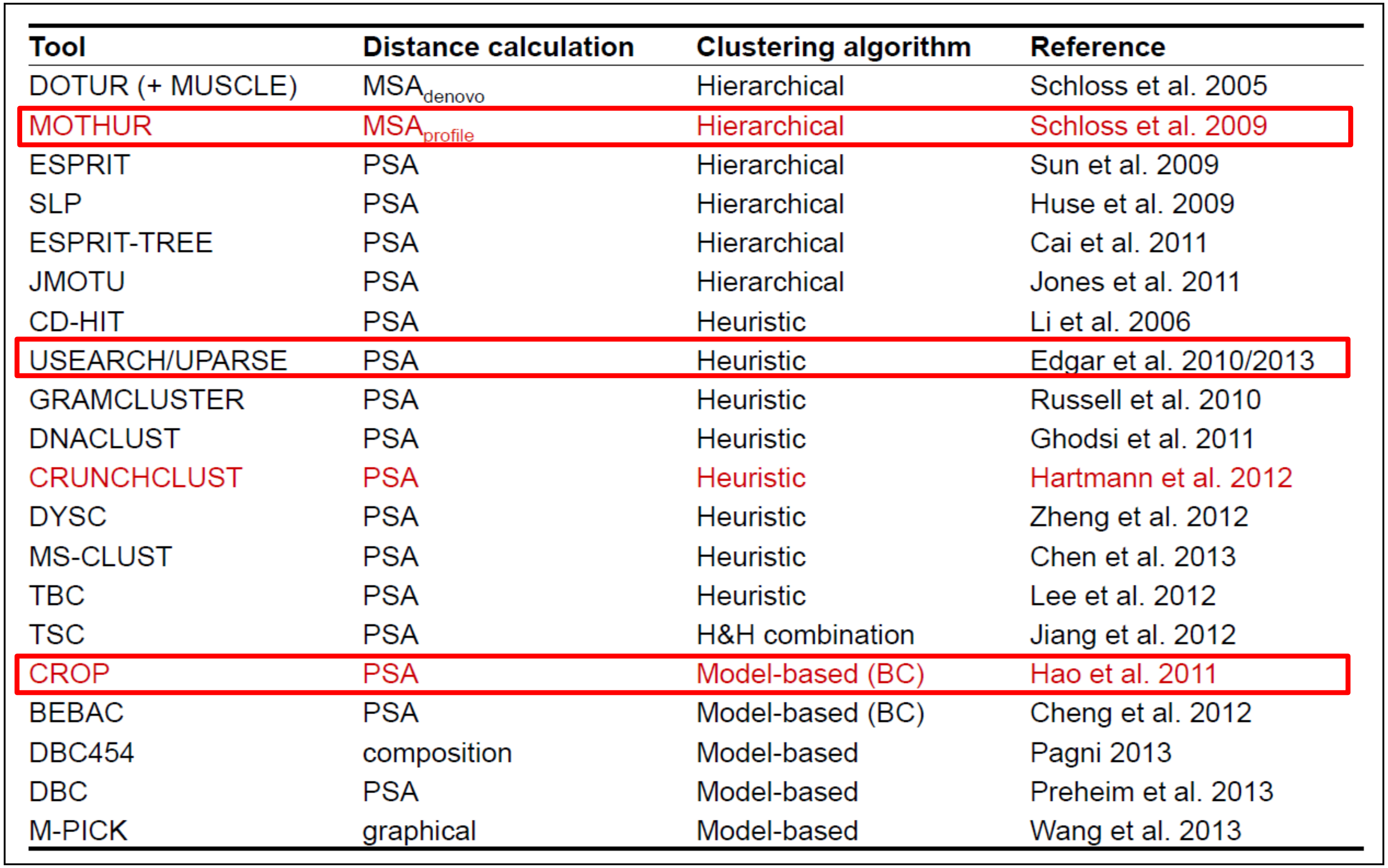
Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

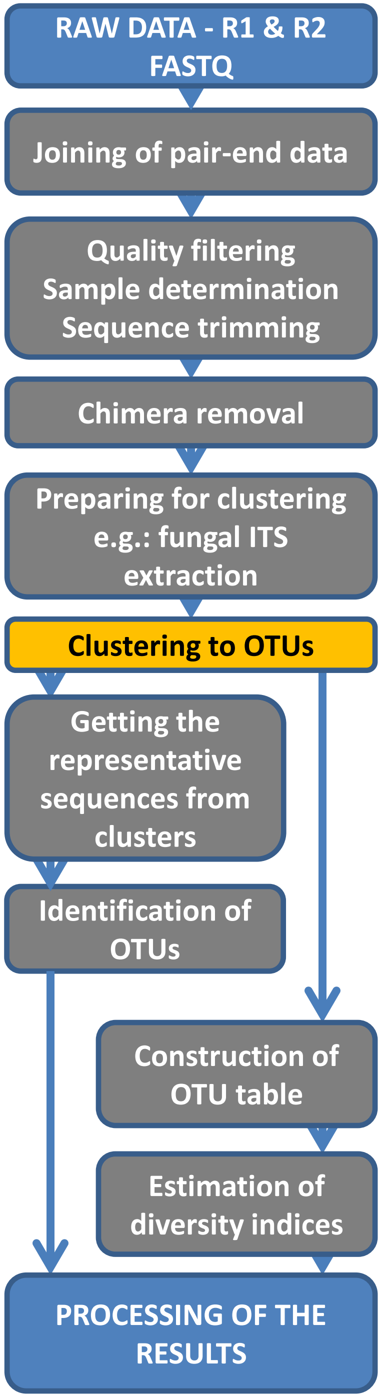
## programs



| Tool | Distance calculation | Clustering algorithm | Reference |
|---|---|---|---|
| DOTUR (+ MUSCLE) | MSA$_{denovo}$ | Hierarchical | Schloss et al. 2005 |
| MOTHUR | MSA$_{profile}$ | Hierarchical | Schloss et al. 2009 |
| ESPRIT | PSA | Hierarchical | Sun et al. 2009 |
| SLP | PSA | Hierarchical | Huse et al. 2009 |
| ESPRIT-TREE | PSA | Hierarchical | Cai et al. 2011 |
| JMOTU | PSA | Hierarchical | Jones et al. 2011 |
| CD-HIT | PSA | Heuristic | Li et al. 2006 |
| USEARCH/UPARSE | PSA | Heuristic | Edgar et al. 2010/2013 |
| GRAMCLUSTER | PSA | Heuristic | Russell et al. 2010 |
| DNACLUST | PSA | Heuristic | Ghodsi et al. 2011 |
| CRUNCHCLUST | PSA | Heuristic | Hartmann et al. 2012 |
| DYSC | PSA | Heuristic | Zheng et al. 2012 |
| MS-CLUST | PSA | Heuristic | Chen et al. 2013 |
| TBC | PSA | Heuristic | Lee et al. 2012 |
| TSC | PSA | H&H combination | Jiang et al. 2012 |
| CROP | PSA | Model-based (BC) | Hao et al. 2011 |
| BEBAC | PSA | Model-based (BC) | Cheng et al. 2012 |
| DBC454 | composition | Model-based | Pagni 2013 |
| DBC | PSA | Model-based | Preheim et al. 2013 |
| M-PICK | graphical | Model-based | Wang et al. 2013 |

PSA - Pairwise Sequence Alignment
MSA - Multiple Sequence Alignment

# Clustering to OTUs (hierarchical)



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

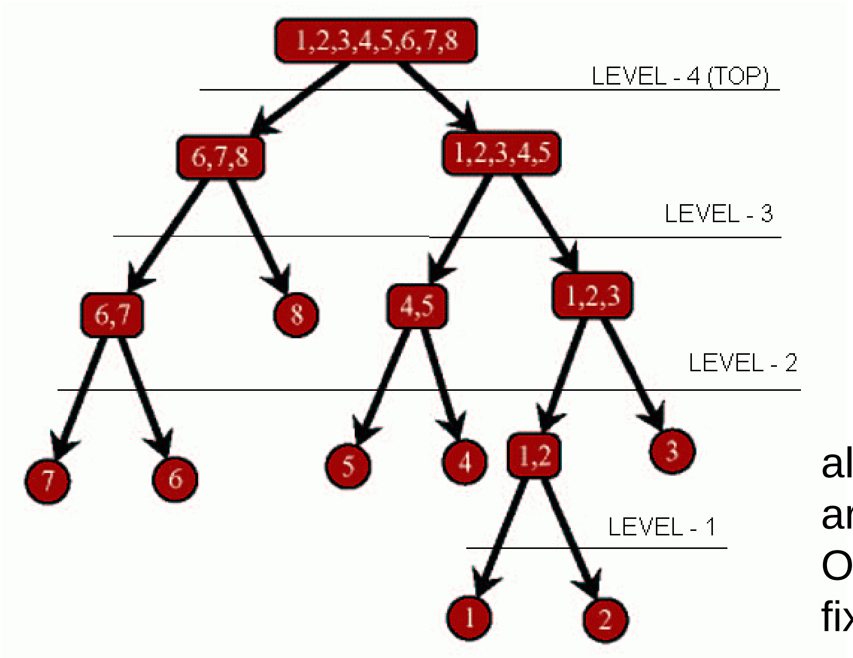
Getting the representative sequences from clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

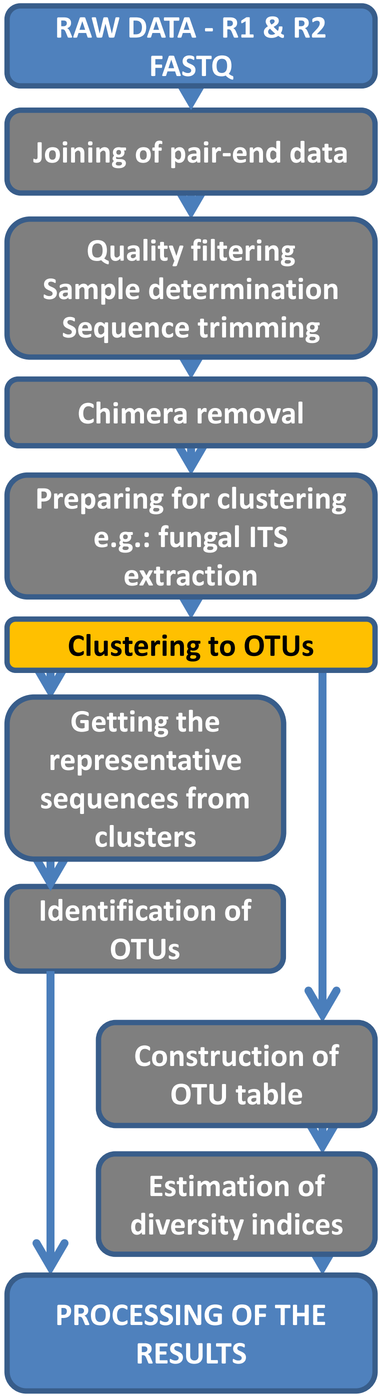


all pairwise comparisons are performed and OTUs are delineated at fixed distance level

**linking method is an important driver of the outcome:**

- **single-linkage clustering** (SL) clusters may be merged together due to single sequences being close to each other, even though many of the sequences in each cluster may be very distant to each other
- **complete-linkage clustering** (CL) tends to find compact clusters of approximately equal diameters. With CL, all objects in a cluster are similar to each other
- **average-linkage clustering** (AL) can be seen as an intermediate between single and complete linkage clustering, resulting in more homogeneous clusters than those obtained by the single-linkage method

# Clustering to OTUs (heuristic - USEARCH)



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

**Clustering to OTUs**

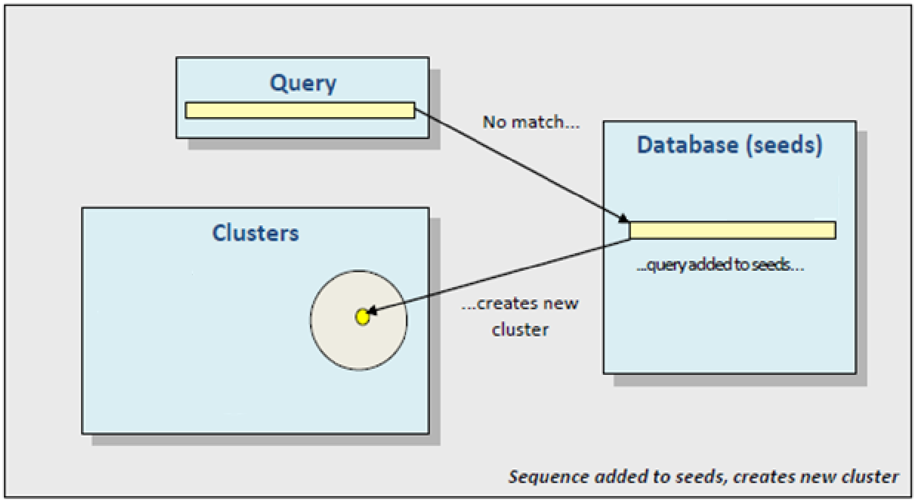
Getting the representative sequences from clusters

Identification of OTUs

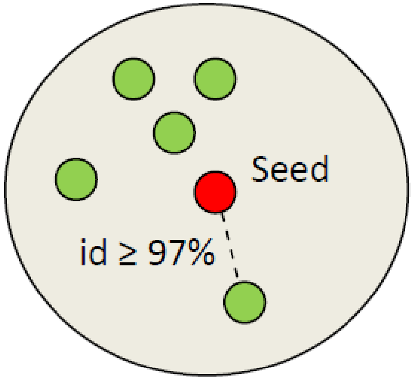
Construction of OTU table

Estimation of diversity indices

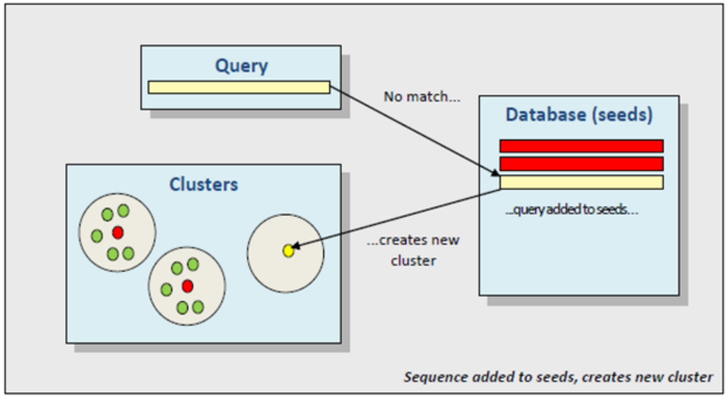
PROCESSING OF THE RESULTS



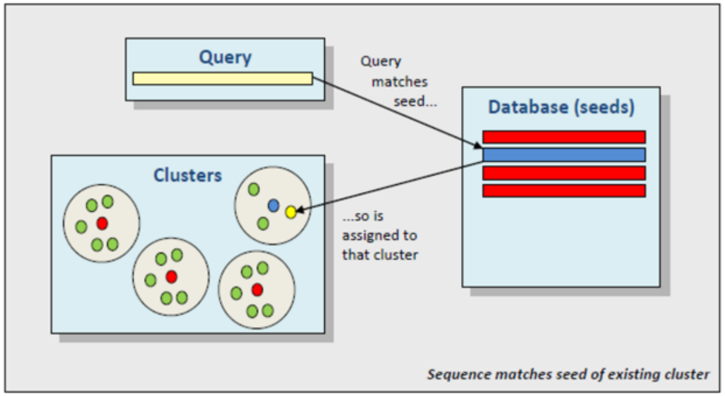
*Sequence added to seeds, creates new cluster*



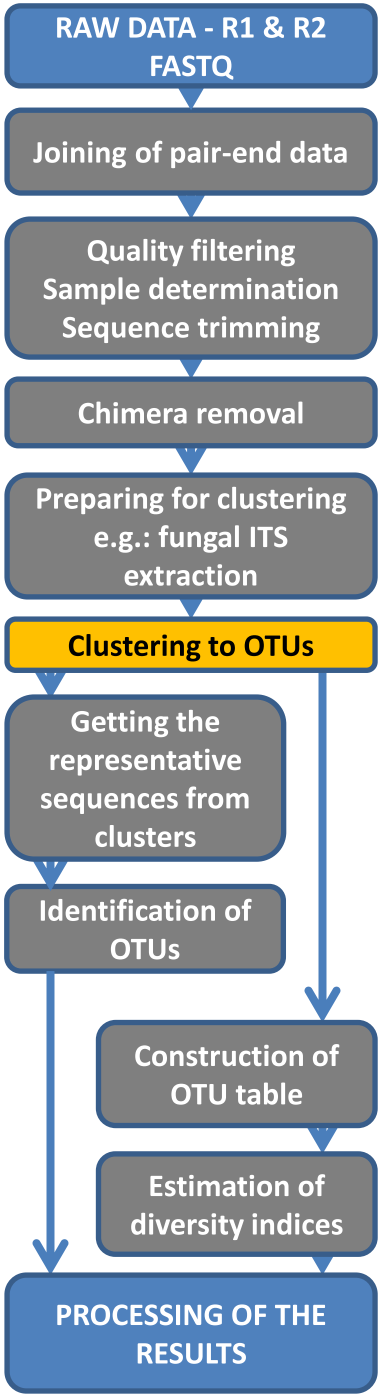
cluster definition



*Sequence added to seeds, creates new cluster*



*Sequence matches seed of existing cluster*



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

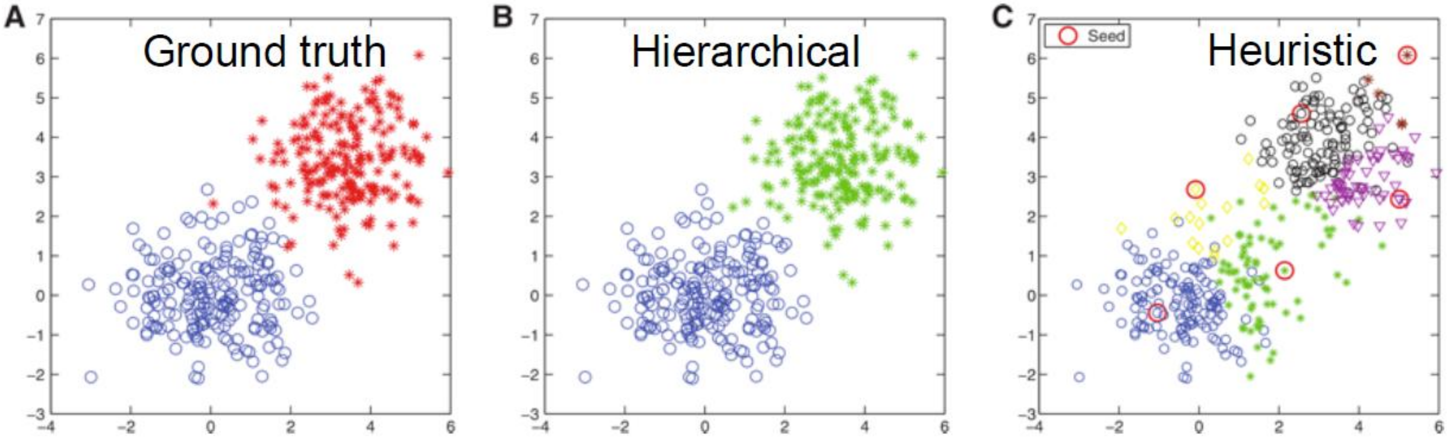
Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

# Clustering to OTUs (hierarchical vs. heuristic)



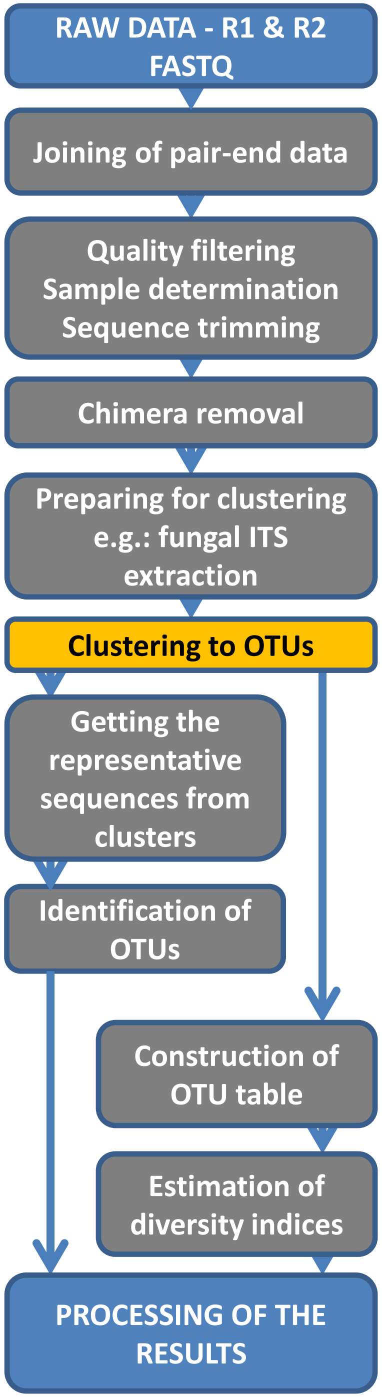
**Hierarchical clustering**

- is able to identify the real clusters (ideally)
- computationally expensive

x

**Heuristic clustering**

- computationally cheap
- often generates artificial clusters (overestimated diversity)

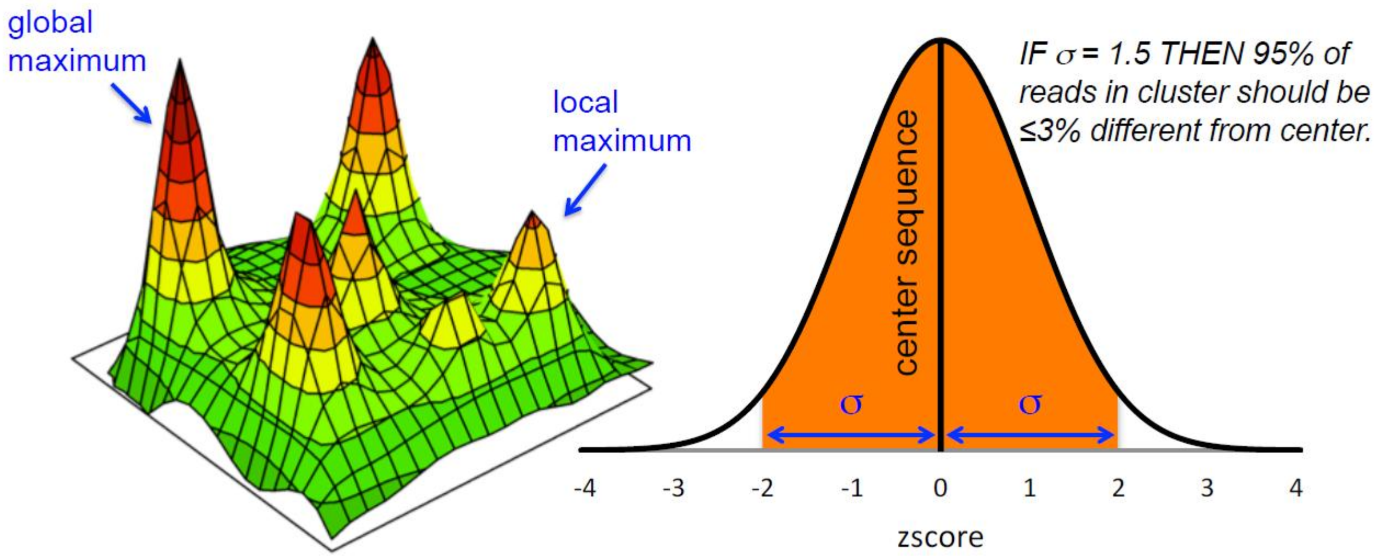

RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering
Sample determination
Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting the representative sequences from clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

# Clustering to OTUs (model based)

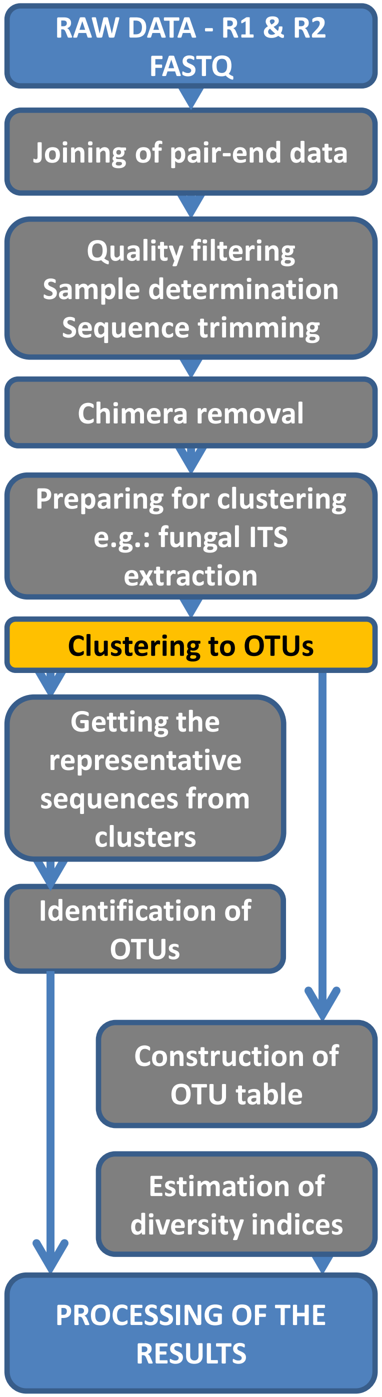
## CROP (FILTER – PSA – BAYESÏAN)

*Hao et al. 2011 (Bioinformatics): "If we consider the sequences as data points in a high-dimensional space […], then the probability that a sequence belongs to a cluster becomes a function of the distance between the sequence and the center."*

**CROP uses a mixture model to find subpopulations among all sequences under the assumption that they are independently drawn from a mixture of Gaussian distributions.**



*IF σ = 1.5 THEN 95% of reads in cluster should be ≤3% different from center.*



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

Identification of OTUs

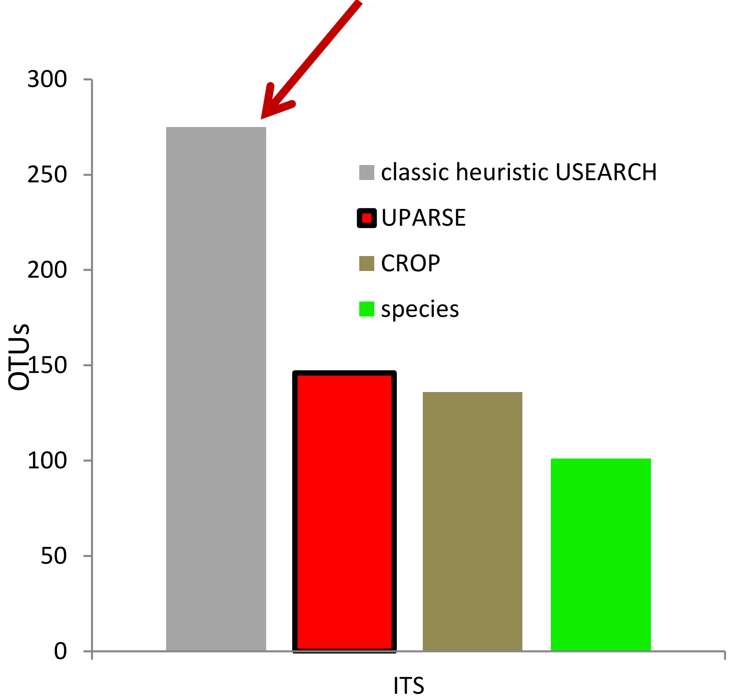
Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

# Clustering to OTUs (comparison)

Problems with overestimation of obtained OTUs…



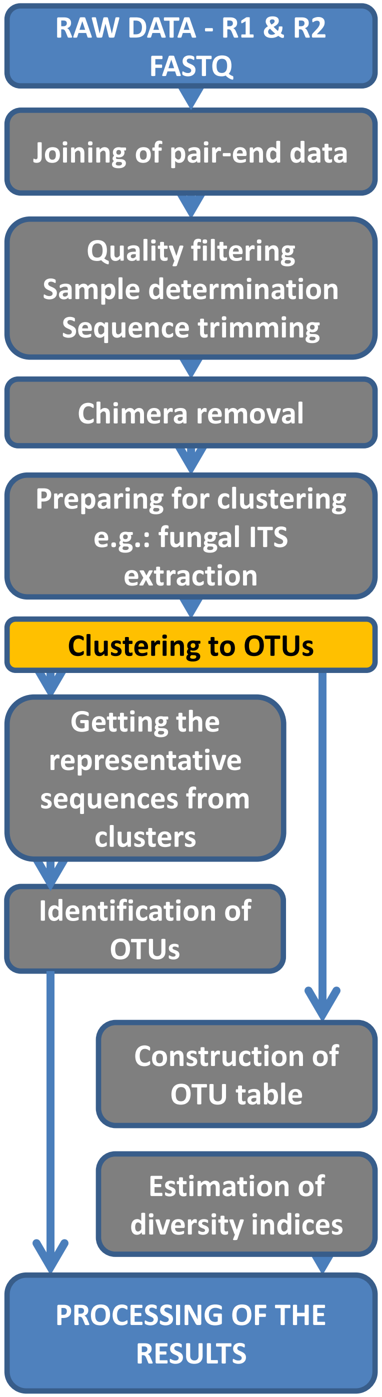
mock fungal community
about 100 fungal species

solution
**UPARSE (USEARCH)** – improved heuristic algorithm which is able to recognize chimeric sequences

**http://drive5.com/uparse/**

Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, *Nature Methods* [Pubmed:23955772, dx.doi.org/10.1038/nmeth.2604].

# Clustering-independent methods



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

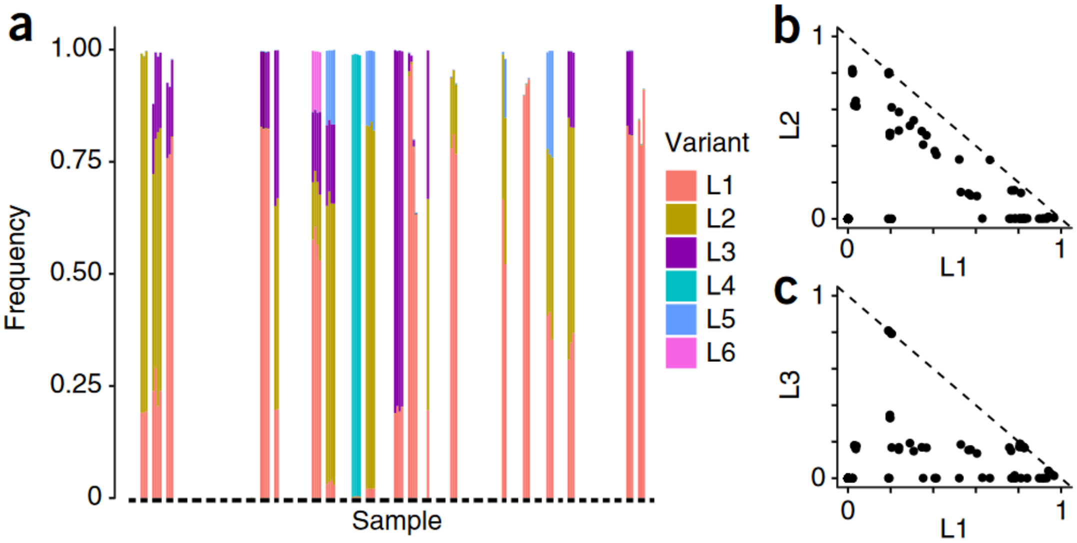
Identification of OTUs

Construction of OTU table

Estimation of diversity indices

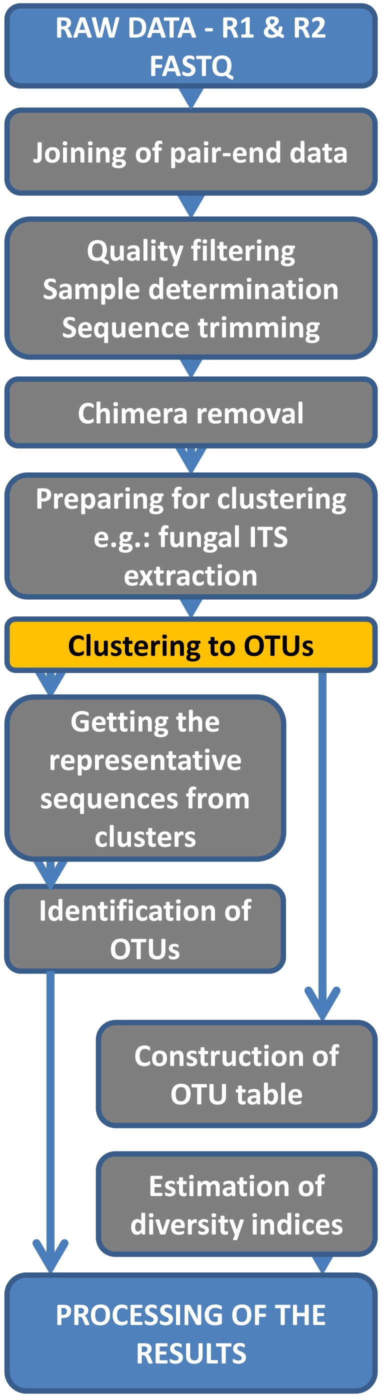
PROCESSING OF THE RESULTS

Callahan, Benjamin J., et al. "**DADA2**: high-resolution sample inference from Illumina amplicon data." *Nature methods* (2016).



L. crispatus sequence variants in the human vaginal community during pregnancy. DADA2 identified six L. crispatus 16S rRNA sequence variants present in multiple samples and a significant fraction of all reads.

Amir, Amnon, et al. "**Deblur** Rapidly Resolves Single-Nucleotide Community Sequence Patterns." *mSystems* 2.2 (2017).



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

Identification of OTUs

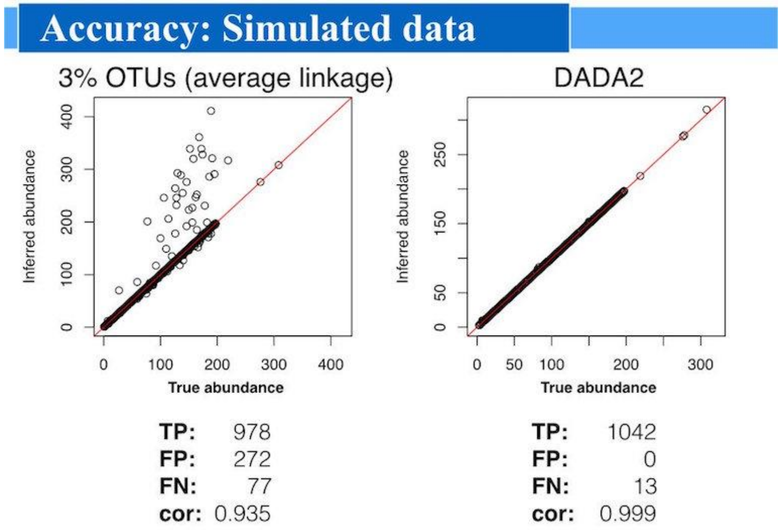
Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

DADA2

Amplicon Sequencing. Exactly. Version 1.18



**Data:** Kopylova, et al. mSystems, 2016.

## Advantages

Resolution: DADA2 infers exact amplicon sequence variants (ASVs) from amplicon data, resolving biological differences of even 1 or 2 nucleotides.

Accuracy: DADA2 reports fewer false positive sequence variants than other methods report false OTUs.

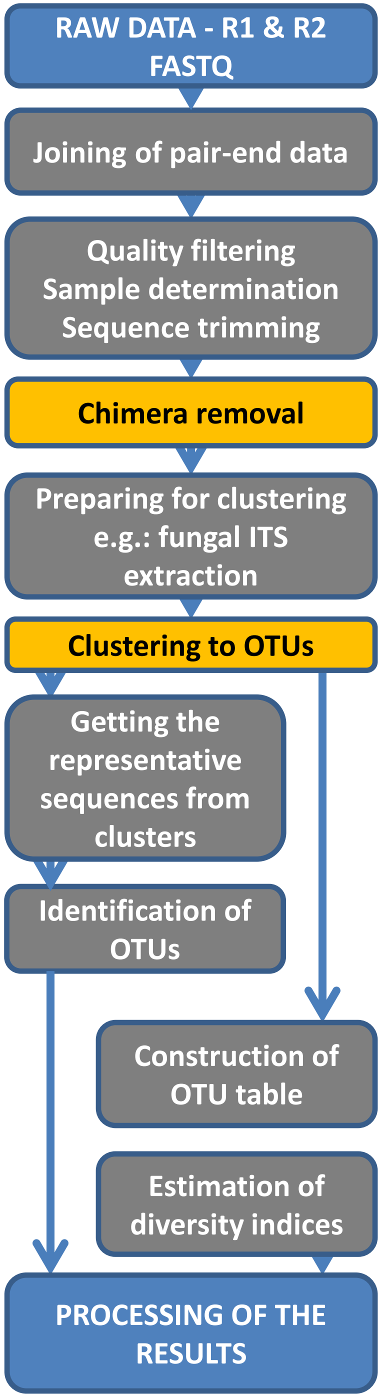
Comparability: The ASVs output by DADA2 can be directly compared between studies, without the need to reprocess the pooled data.

Computational Scaling: The compute time of DADA2 scales linearly sample number, and memory requirements are essentially flat.

## Disadvantage

sequence variants are not representing the real sequences (they are estimated based on the errors modeling)

# UPARSE: Clustering and chimera removal in the same time



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

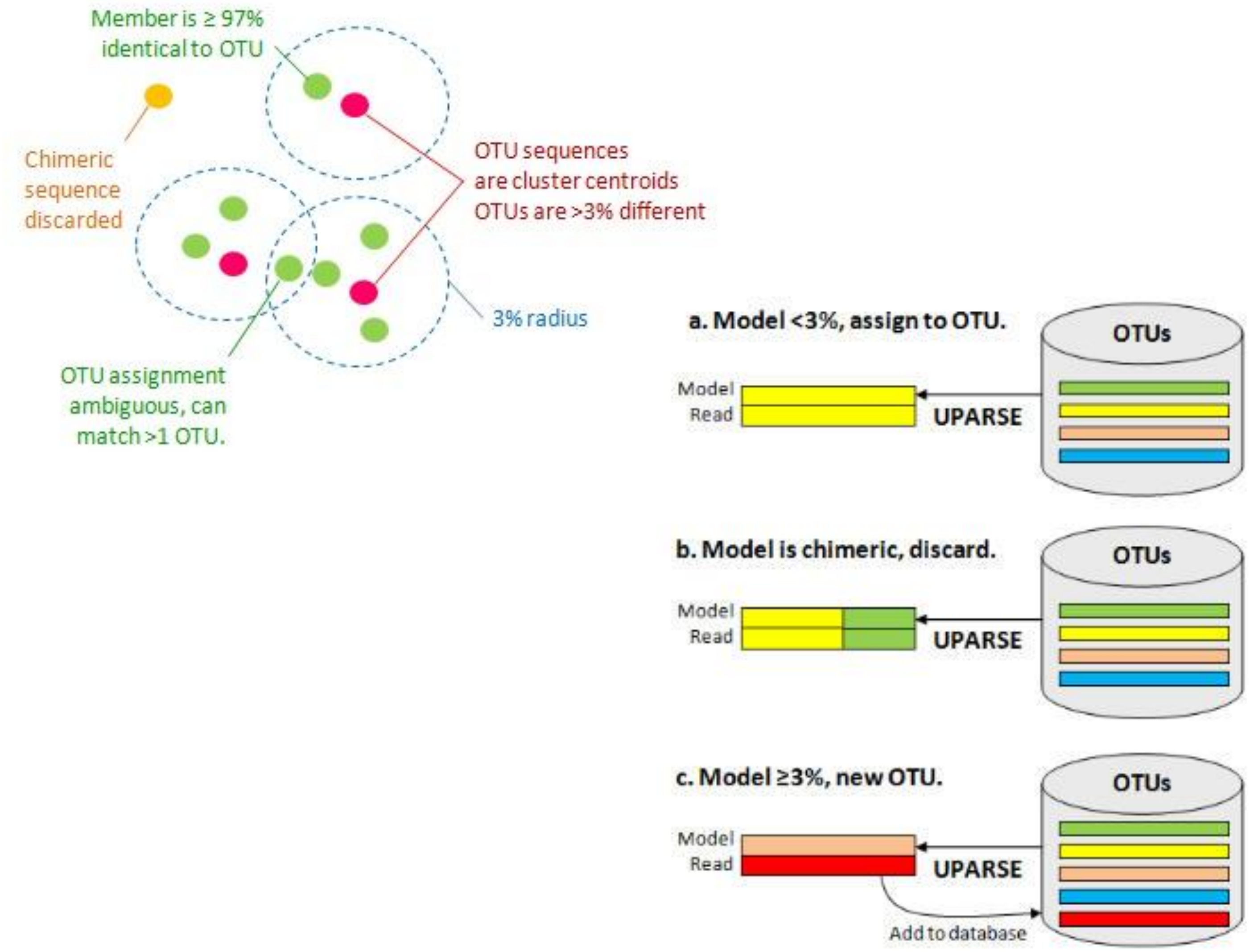
Getting the representative sequences from clusters

Identification of OTUs

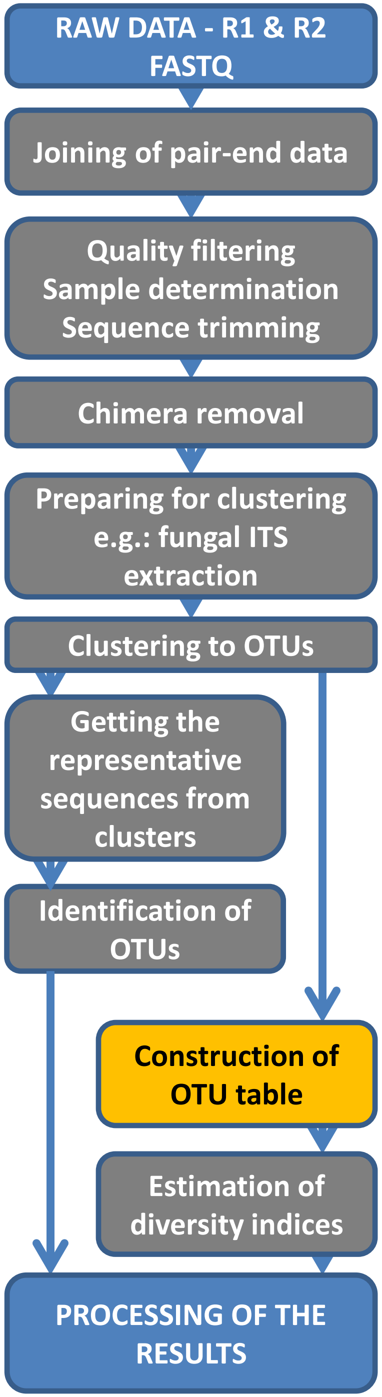
Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS


Member is ≥ 97% identical to OTU
Chimeric sequence discarded
OTU sequences are cluster centroids OTUs are >3% different
3% radius
OTU assignment ambiguous, can match >1 OTU.
a. Model <3%, assign to OTU.
Model
Read
UPARSE
OTUs
b. Model is chimeric, discard.
Model
Read
UPARSE
OTUs
c. Model ≥3%, new OTU.
Model
Read
UPARSE
OTUs
Add to database

Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, *Nature Methods*


RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering Sample determination Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting the representative sequences from clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

# Construction of OTU table

**OTU table** - matrix that gives the number of reads per sample per OTU

| OTU_ID | SAMPLE_1 | SAMPLE_2 | SAMPLE_3 | SAMPLE_4 | SAMPLE_5 | SAMPLE_6 | SAMPLE_7 | SAMPLE_8 |
|---|---|---|---|---|---|---|---|---|
| CL00001 | 249 | 189 | 220 | 311 | 1 | 16 | 68 | 2 |
| CL00002 | 201 | 19 | 169 | 438 | 1 | 8 | 12 | 0 |
| CL00003 | 190 | 39 | 176 | 210 | 0 | 21 | 20 | 1 |
| CL00004 | 183 | 36 | 195 | 177 | 1 | 16 | 16 | 0 |
| CL00005 | 0 | 26 | 2 | 35 | 20 | 164 | 4 | 116 |
| CL00006 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| CL00007 | 133 | 71 | 125 | 89 | 0 | 3 | 26 | 0 |
| CL00008 | 106 | 42 | 96 | 158 | 0 | 10 | 14 | 0 |
| CL00009 | 95 | 46 | 108 | 134 | 2 | 7 | 24 | 0 |
| CL00010 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| CL00011 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

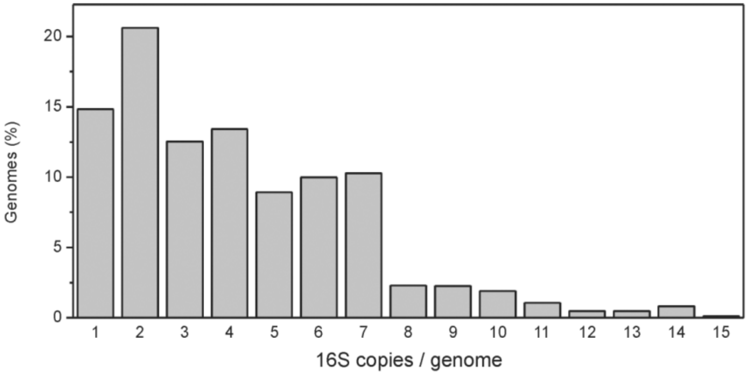
**OTU frequency does not correlate with species frequency**

This means, for example, that the most abundant OTU does not have to be the most abundant species – especially because of multi-copy nature of target genes as 16S and ITS

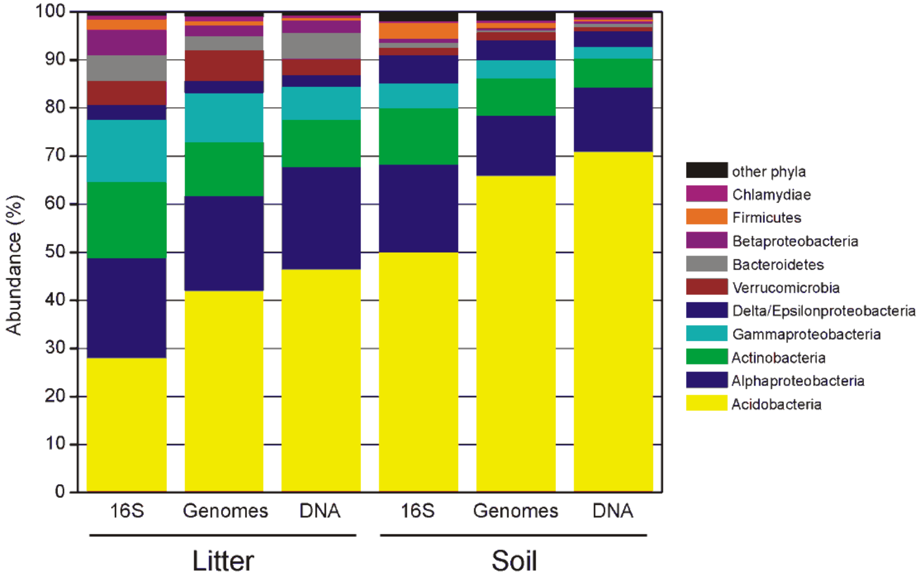
**Singleton counts are especially suspect**

- many OTU table entries are often singletons (have value 1) for smaller OTUs because the total count is distributed over several samples
- Small counts are more likely to be spurious, especially singletons, either because the OTU itself is spurious (e.g., an undetected chimera), or because of cross-talk

# Normalize OTU table by 16S copy number



**rrnDB**
A searchable database documenting variation in ribosomal RNA operons (rrn) in Bacteria and Archaea. Find information such as the 16S gene copy number of an organism by looking up its name under the NCBI or RDP taxonomy or by full-text search of rrnDB's records.

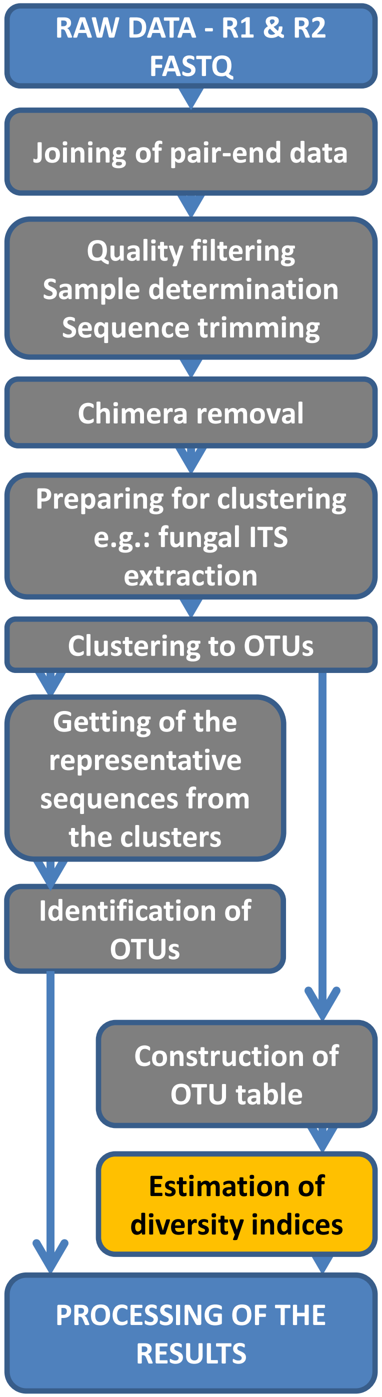


**Abundance of bacterial 16S rRNA sequences, genomes and DNA in forest litter and soil.**

Relative abundance of bacterial 16S rRNA sequences in the amplicon pool from Picea abies litter and soil (Baldrian et al., 2012), and estimates of the relative abundance of bacterial genomes and DNA. The estimates were calculated using the values of 16S rRNA copy numbers and genome sizes of the closest hits to each bacterial OTU.

T. Větrovský & P. Baldrian - PloS one, 2013 & Stoddard et al. (2015) https://rrndb.umms.med.umich.edu/

# Estimation of diversity indices


RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

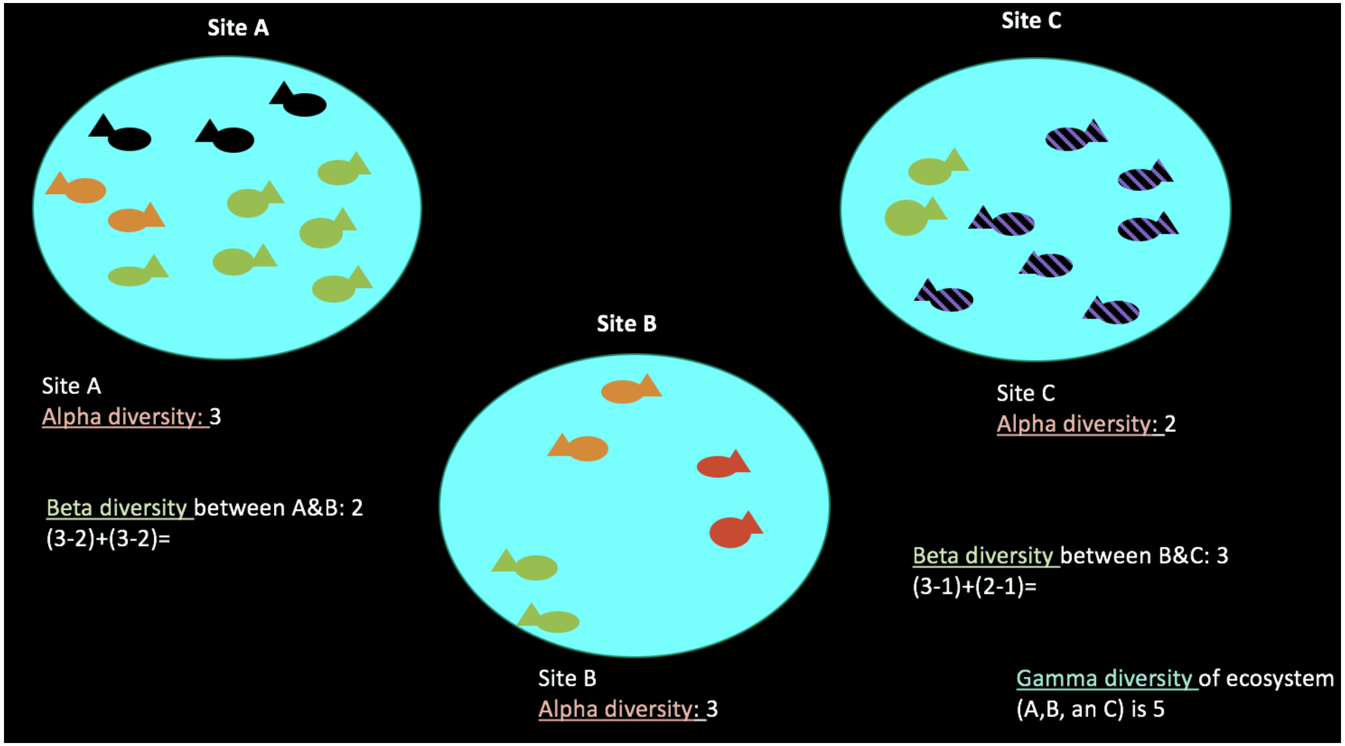
Getting of the representative sequences from the clusters

Identification of OTUs

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS



- **Alpha-diversity**: diversity of organisms in one sample / environment
  - **Shannon index**
  - **Chao1**
  - **Observed OTUs (Richness)**

- **Beta-diversity**: differences in diversities across samples or environments
  - **UniFrac** (Lozupone et al, AEM, 2005) (phylogenetic)
  - **Bray-Curtis** dissimilarity measure (OTU abundance)
  - **Jaccard** similarity coefficient (OTU presence/absence)

# Alpha diversity

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

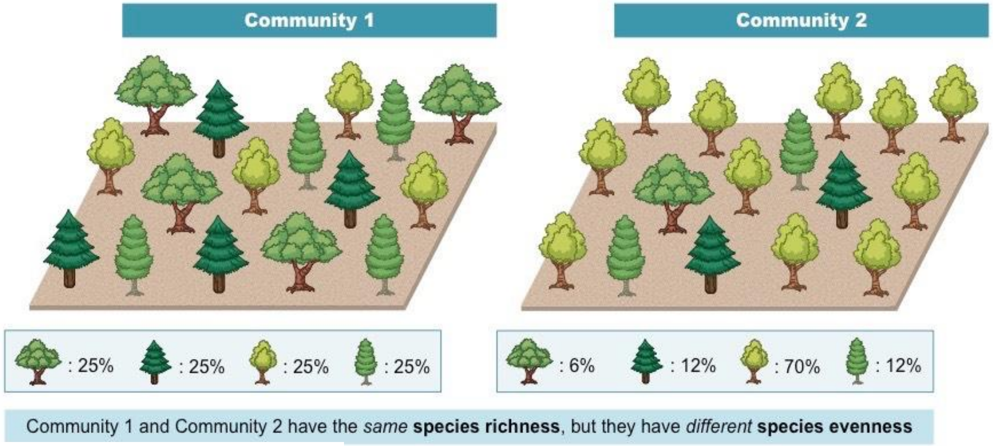
## Shannon index (Shannon entropy)

Then the Shannon entropy quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset.

*$p_i$* – proportion of the population made up of species i
*$S$* – number of species in sample

## Species evenness

Species evenness refers to how close in numbers each species in an environment is. Mathematically it is defined as a diversity index, a measure of biodiversity which quantifies how equal the community is numerically.



Community 1 and Community 2 have the *same* **species richness**, but they have *different* **species evenness**

$$J' = \frac{H'}{H'_{\max}} \qquad H'_{\max} = -\sum_{i=1}^{S} \frac{1}{S} \ln \frac{1}{S} = \ln S$$

## Chao1 index

Estimate diversity from abundance data (importance of rare OTUs)

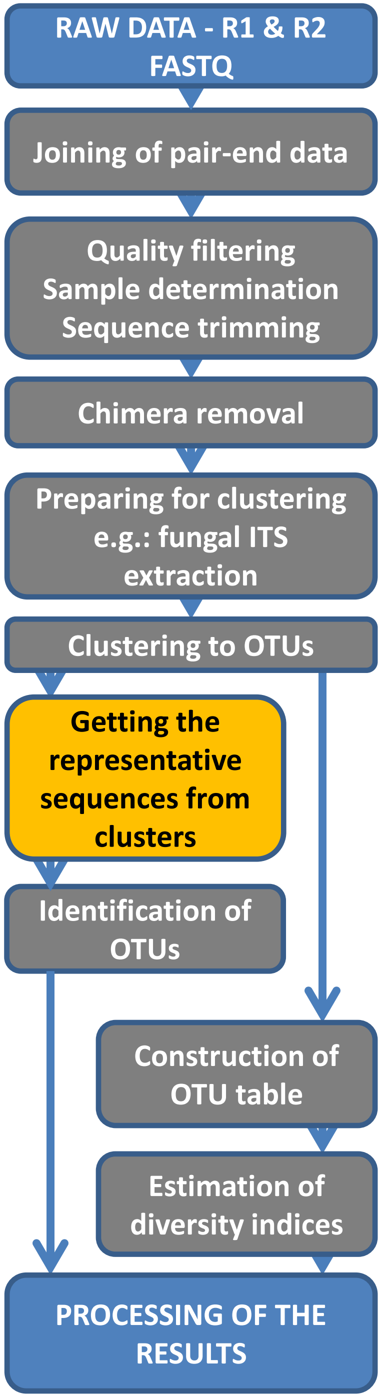
$$S_{est} = S_{obs} + \left(\frac{{f_1}^2}{2f_2}\right)$$

where $S_{obs}$ is the number of species in the sample, $f_1$ is the number of singletons and $f_2$ is the number of doubletons.

Shannon, C. E. (1948) A mathematical theory of communication.
The Bell System Technical Journal, 27, 379–423 and 623–656.
*Chao*, A.; Shen, T-J. (2003)

https://palaeo-electronica.org/2011_1/238/estimate.htm

# Getting of the representative sequences from the clusters



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

**Getting the representative sequences from clusters**

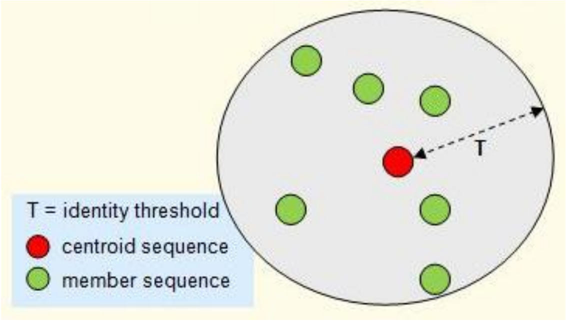
Identification of OTUs

Construction of OTU table

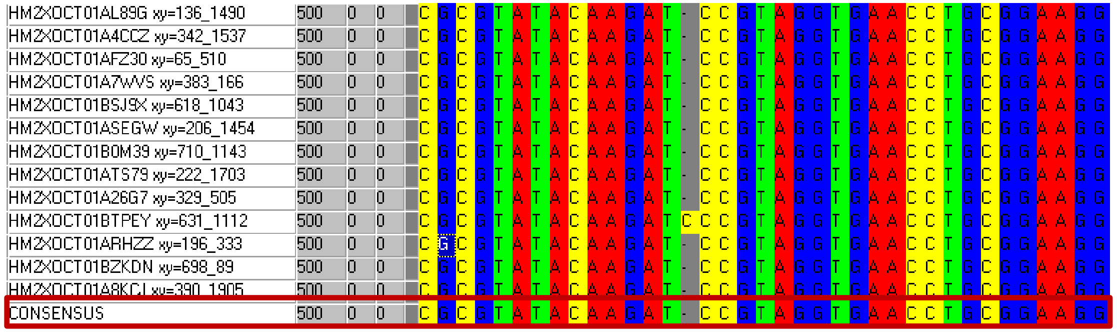
Estimation of diversity indices

PROCESSING OF THE RESULTS

## centroid

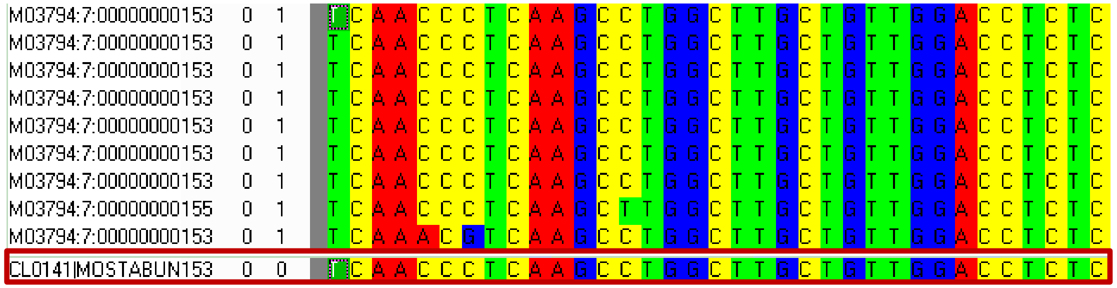


## consensus

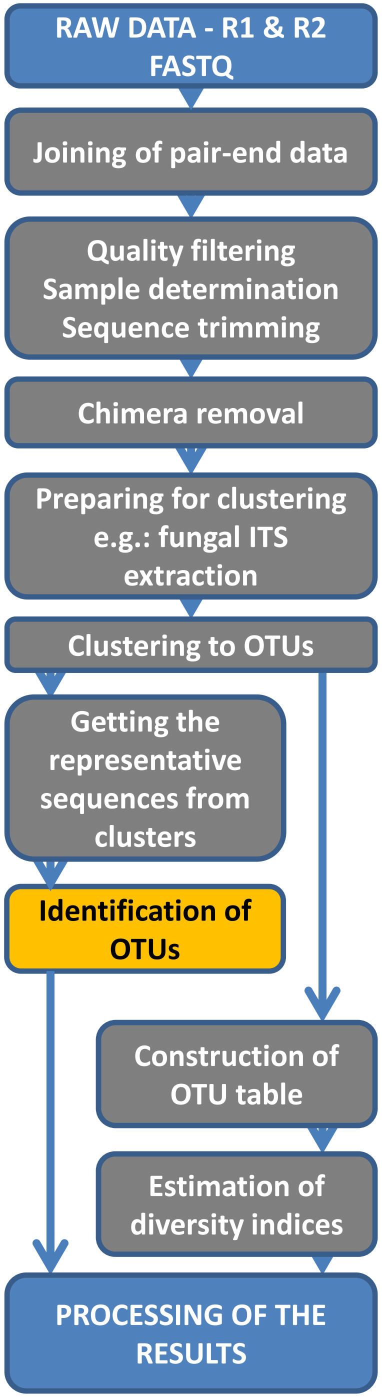


```
HM2XOCT01AL89G xy=136_1490   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01A4CCZ xy=342_1537   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01AFZ30 xy=65_510     500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01A7WVS xy=383_166    500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01BSJ9X xy=618_1043   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01ASEGW xy=206_1454   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01B0M39 xy=710_1143   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01ATS79 xy=222_1703   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01A26G7 xy=329_505    500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01BTPEY xy=631_1112   500  0  0  CGCGTATACAAGATCCCGTAGGTGAACCTGCGGAAGG
HM2XOCT01ARHZZ xy=196_333    500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01BZKDN xy=698_89     500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
HM2XOCT01A8KCJ xy=390_1905   500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
CONSENSUS                    500  0  0  CGCGTATACAAGAT-CCGTAGGTGAACCTGCGGAAGG
```

## most abundant



```
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000155   0  1  TCAACCCTCAAGCTTGGCTTGCTGTTGGACCTCTC
M03794:7:00000000153   0  1  TCAAACGTCAAGCCTGGCTTGCTGTTGGACCTCTC
CL0141|MOSTABUN153     0  0  TCAACCCTCAAGCCTGGCTTGCTGTTGGACCTCTC
```



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

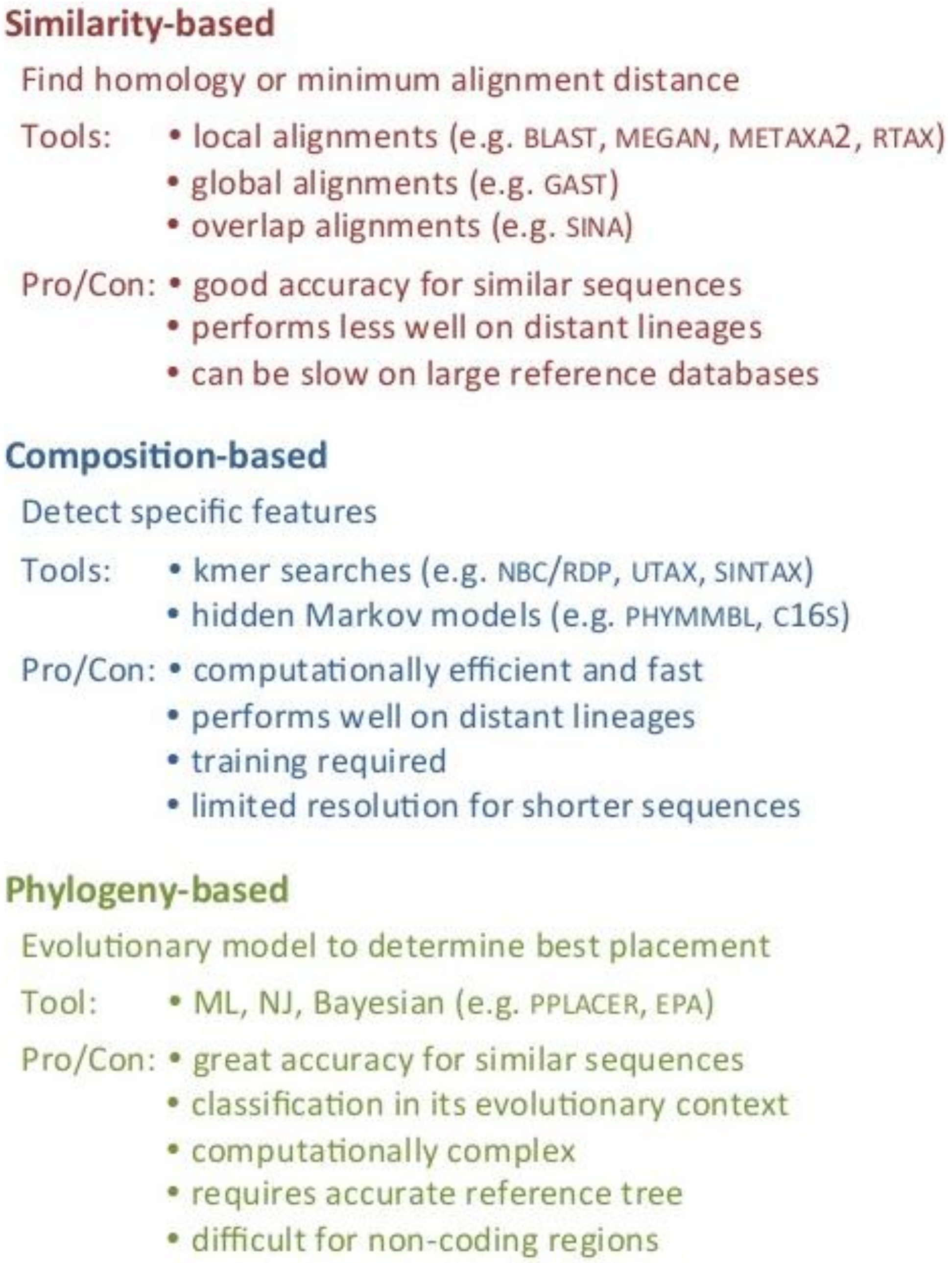
**Identification of OTUs**

Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

# Taxonomic classification of OTUs



## Similarity-based

Find homology or minimum alignment distance

Tools:
- local alignments (e.g. BLAST, MEGAN, METAXA2, RTAX)
- global alignments (e.g. GAST)
- overlap alignments (e.g. SINA)

Pro/Con:
- good accuracy for similar sequences
- performs less well on distant lineages
- can be slow on large reference databases

## Composition-based

Detect specific features

Tools:
- kmer searches (e.g. NBC/RDP, UTAX, SINTAX)
- hidden Markov models (e.g. PHYMMBL, C16S)

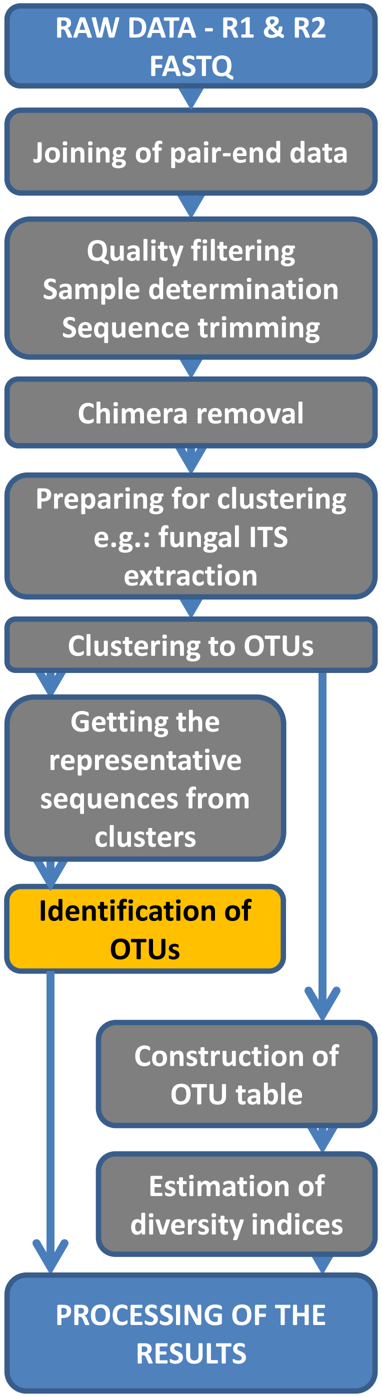
Pro/Con:
- computationally efficient and fast
- performs well on distant lineages
- training required
- limited resolution for shorter sequences

## Phylogeny-based

Evolutionary model to determine best placement

Tool:
- ML, NJ, Bayesian (e.g. PPLACER, EPA)

Pro/Con:
- great accuracy for similar sequences
- classification in its evolutionary context
- computationally complex
- requires accurate reference tree
- difficult for non-coding regions


RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering Sample determination Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting the representative sequences from clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

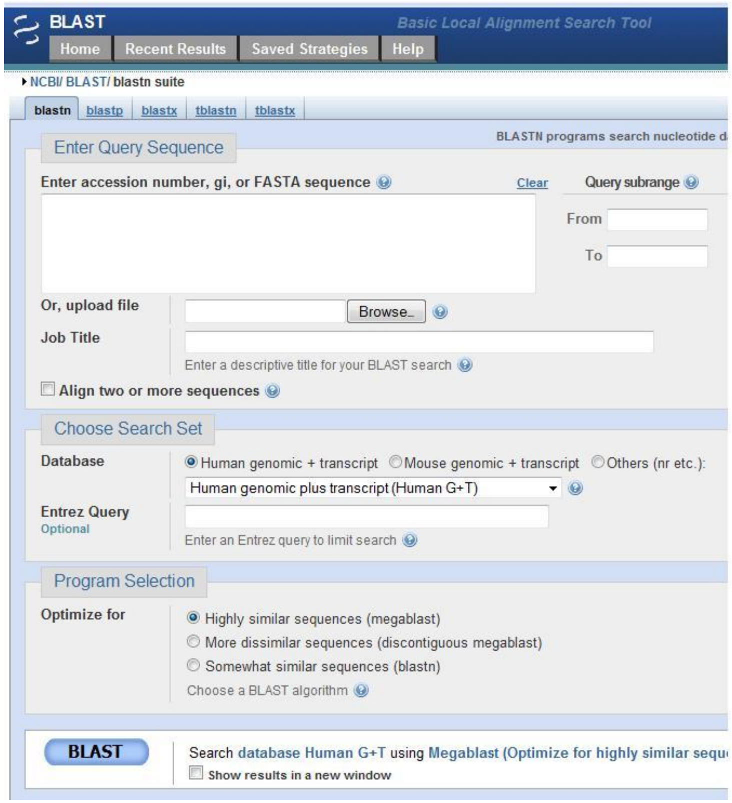
# Identification of OTUs

NCBI

All genes

**GenBank - *genetic sequence database, an annotated collection of all publicly available DNA sequence***

- largest ☺
- many errors ☹



BLASTN programs search nucleotide d

Enter Query Sequence

Enter accession number, gi, or FASTA sequence    Clear    Query subrange

From

To

Or, upload file    Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database    ● Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript (Human G+T)

Entrez Query Optional

Enter an Entrez query to limit search

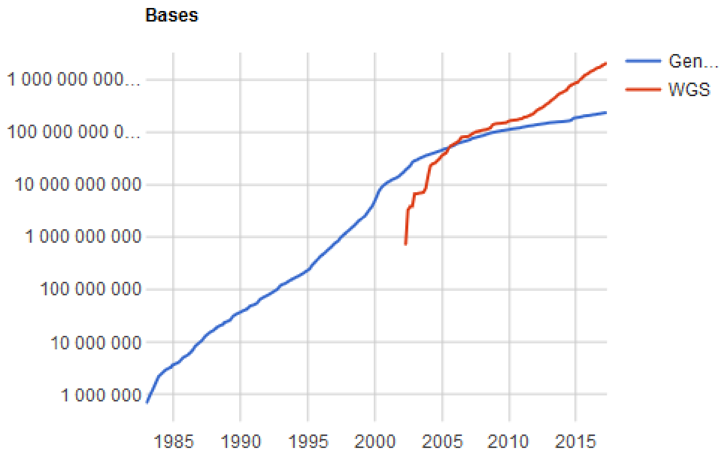
Program Selection

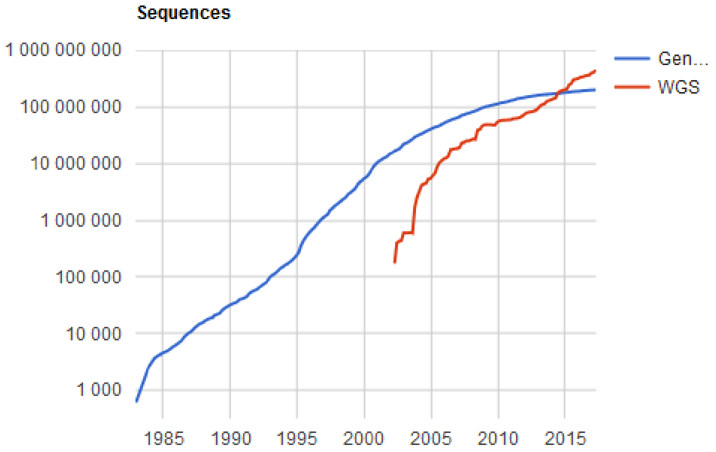
Optimize for
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST    Search database Human G+T using Megablast (Optimize for highly similar sequ

□ Show results in a new window



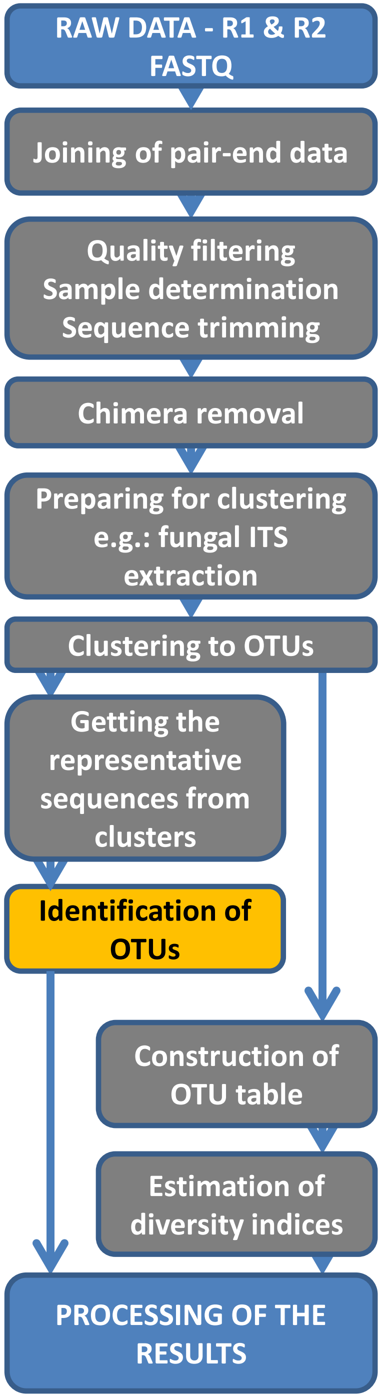


**https://www.ncbi.nlm.nih.gov/genbank/**

Gen – GenBank

WGS – whole genome sequences

# Identification of OTUs


RAW DATA - R1 & R2 FASTQ
Joining of pair-end data
Quality filtering Sample determination Sequence trimming
Chimera removal
Preparing for clustering e.g.: fungal ITS extraction
Clustering to OTUs
Getting the representative sequences from clusters
Identification of OTUs
Construction of OTU table
Estimation of diversity indices
PROCESSING OF THE RESULTS

rdp

Identification of bacteria

**RDP – *Ribosomal Database Project*** provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community

- Hb Hierarchy Browser
- Cl Classifier
- Pm Probe Match
- Fg FunGene
- Al Aligner
- Tb Tree Builder
- Os RDP Open Source
- Tu Tutorials
- Mg MlxS GoogleSheets
- Lc Library Compare
- Sm Sequence Match
- Rp RDPipeline

**https://rdp.cme.msu.edu/**

silva
high quality ribosomal RNA databases

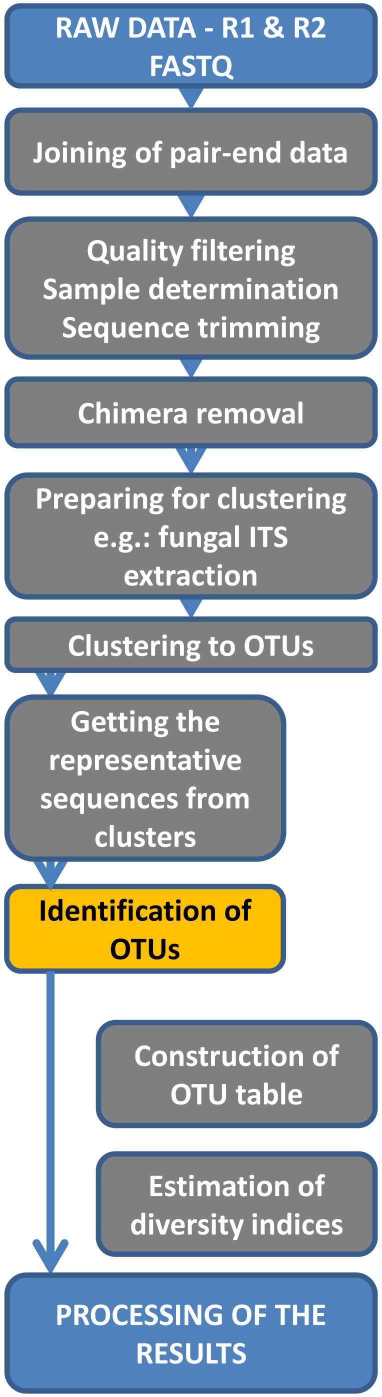
**https://www.arb-silva.de/**

**SILVA** - provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).

GG GREENGENES
The 16S rRNA Gene Database and Tools

**http://greengenes.secondgenome.com/**

# Identification of OTUs


RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

Identification of OTUs

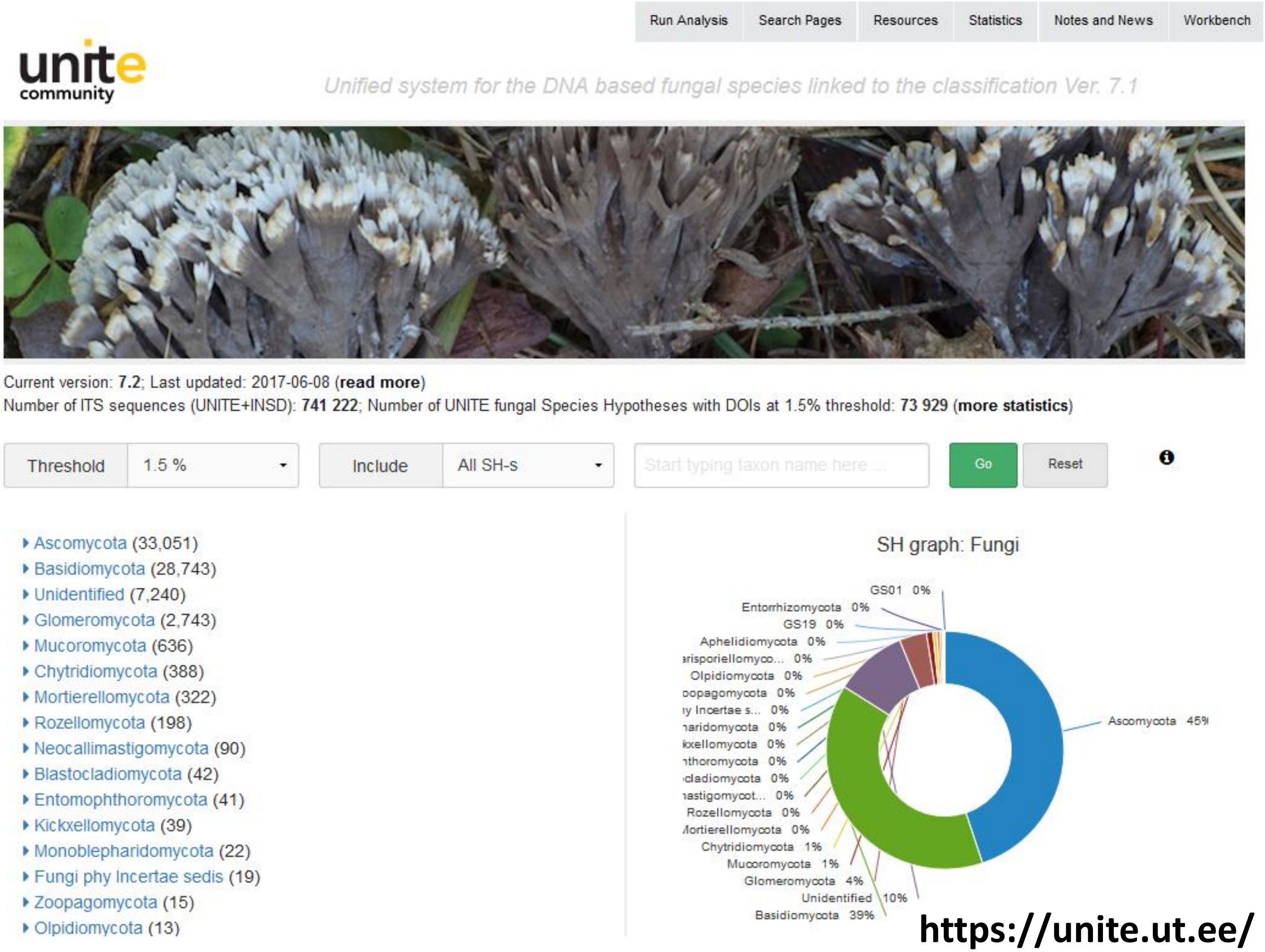
Construction of OTU table

Estimation of diversity indices

PROCESSING OF THE RESULTS

Identification of fungi

**UNITE - *Unified system for the DNA based fungal species linked to the classification***



Unified system for the DNA based fungal species linked to the classification Ver. 7.1

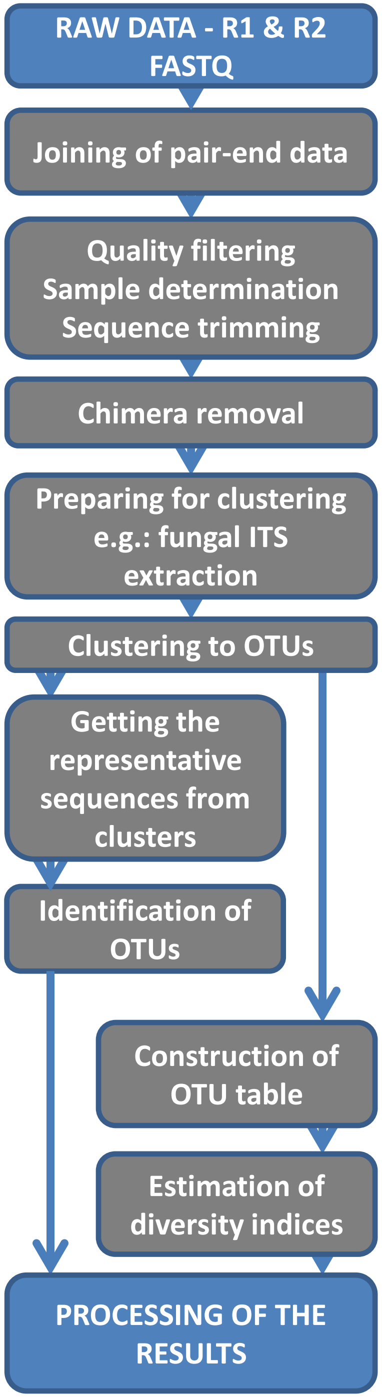
Current version: 7.2; Last updated: 2017-06-08 (**read more**)
Number of ITS sequences (UNITE+INSD): **741 222**; Number of UNITE fungal Species Hypotheses with DOIs at 1.5% threshold: **73 929** (**more statistics**)

Threshold: 1.5 %
Include: All SH-s

- Ascomycota (33,051)
- Basidiomycota (28,743)
- Unidentified (7,240)
- Glomeromycota (2,743)
- Mucoromycota (636)
- Chytridiomycota (388)
- Mortierellomycota (322)
- Rozellomycota (198)
- Neocallimastigomycota (90)
- Blastocladiomycota (42)
- Entomophthoromycota (41)
- Kickxellomycota (39)
- Monoblepharidomycota (22)
- Fungi phy Incertae sedis (19)
- Zoopagomycota (15)
- Olpidiomycota (13)

SH graph: Fungi

**https://unite.ut.ee/**



RAW DATA - R1 & R2 FASTQ

Joining of pair-end data

Quality filtering
Sample determination
Sequence trimming

Chimera removal

Preparing for clustering
e.g.: fungal ITS extraction

Clustering to OTUs

Getting the representative sequences from clusters

Identification of OTUs

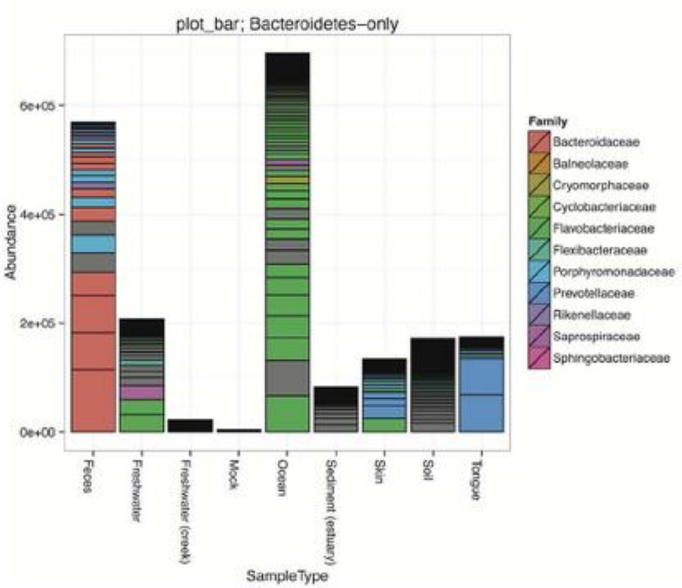
Construction of OTU table

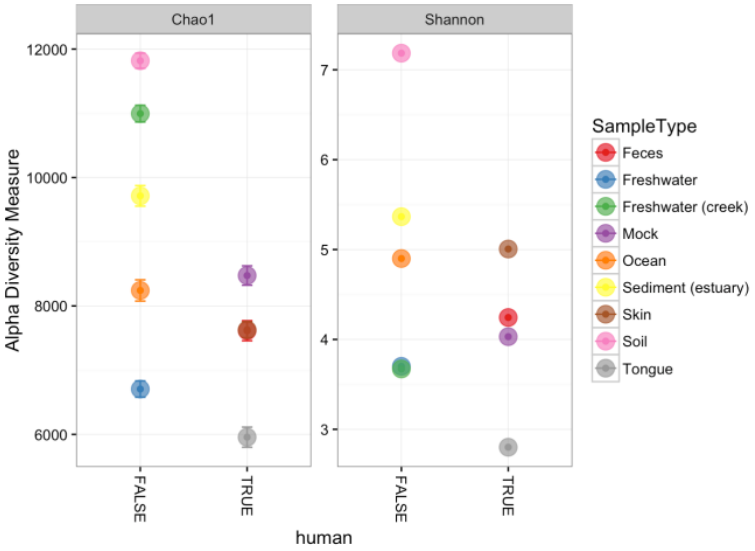
Estimation of diversity indices

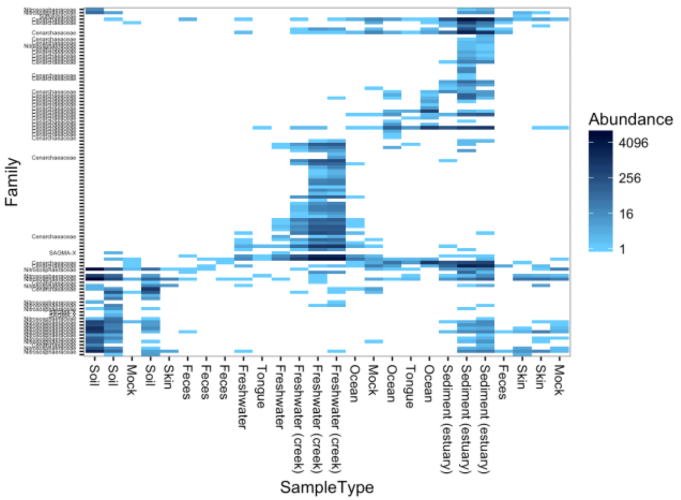
PROCESSING OF THE RESULTS

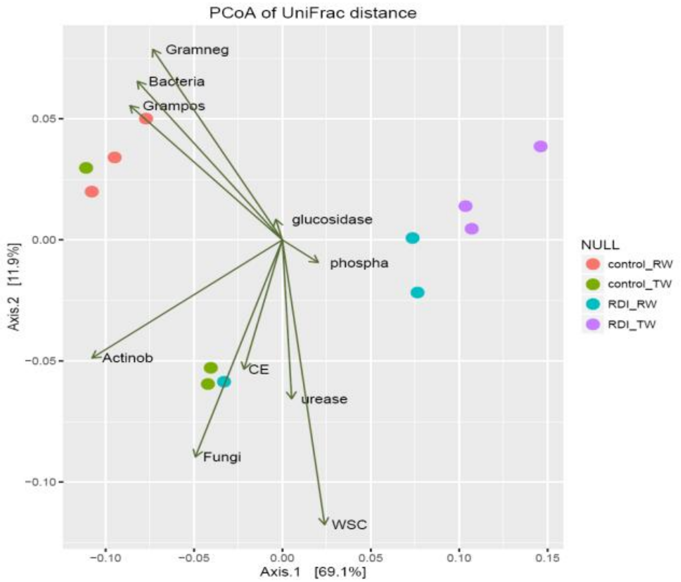
# PROCESSING OF THE RESULTS

## Introduction to multivariate data analysis (Iñaki Class)


plot_bar; Bacteroidetes–only
Abundance
6e+05
4e+05
2e+05
0e+00
Feces
Freshwater
Freshwater (creek)
Mock
Ocean
Sediment (estuary)
Skin
Soil
Tongue
SampleType
Family
Bacteroidaceae
Balneolaceae
Cryomorphaceae
Cyclobacteriaceae
Flavobacteriaceae
Flexibacteraceae
Porphyromonadaceae
Prevotellaceae
Rikenellaceae
Saprospiraceae
Sphingobacteriaceae


Chao1
Shannon
Alpha Diversity Measure
FALSE
TRUE
human
SampleType
Feces
Freshwater
Freshwater (creek)
Mock
Ocean
Sediment (estuary)
Skin
Soil
Tongue


Family
Abundance
Soil
Soil
Mock
Soil
Skin
Feces
Feces
Feces
Freshwater
Tongue
Freshwater
Freshwater
Freshwater (creek)
Freshwater (creek)
Freshwater (creek)
Ocean
Mock
Ocean
Tongue
Ocean
Sediment (estuary)
Sediment (estuary)
Sediment (estuary)
Feces
Skin
Skin
Mock
SampleType


PCoA of UniFrac distance
Gramneg
Bacteria
Grampos
glucosidase
phospha
Actinob
CE
urease
Fungi
WSC
Axis.2 [11.9%]
Axis.1 [69.1%]
NULL
control_RW
control_TW
RDI_RW
RDI_TW