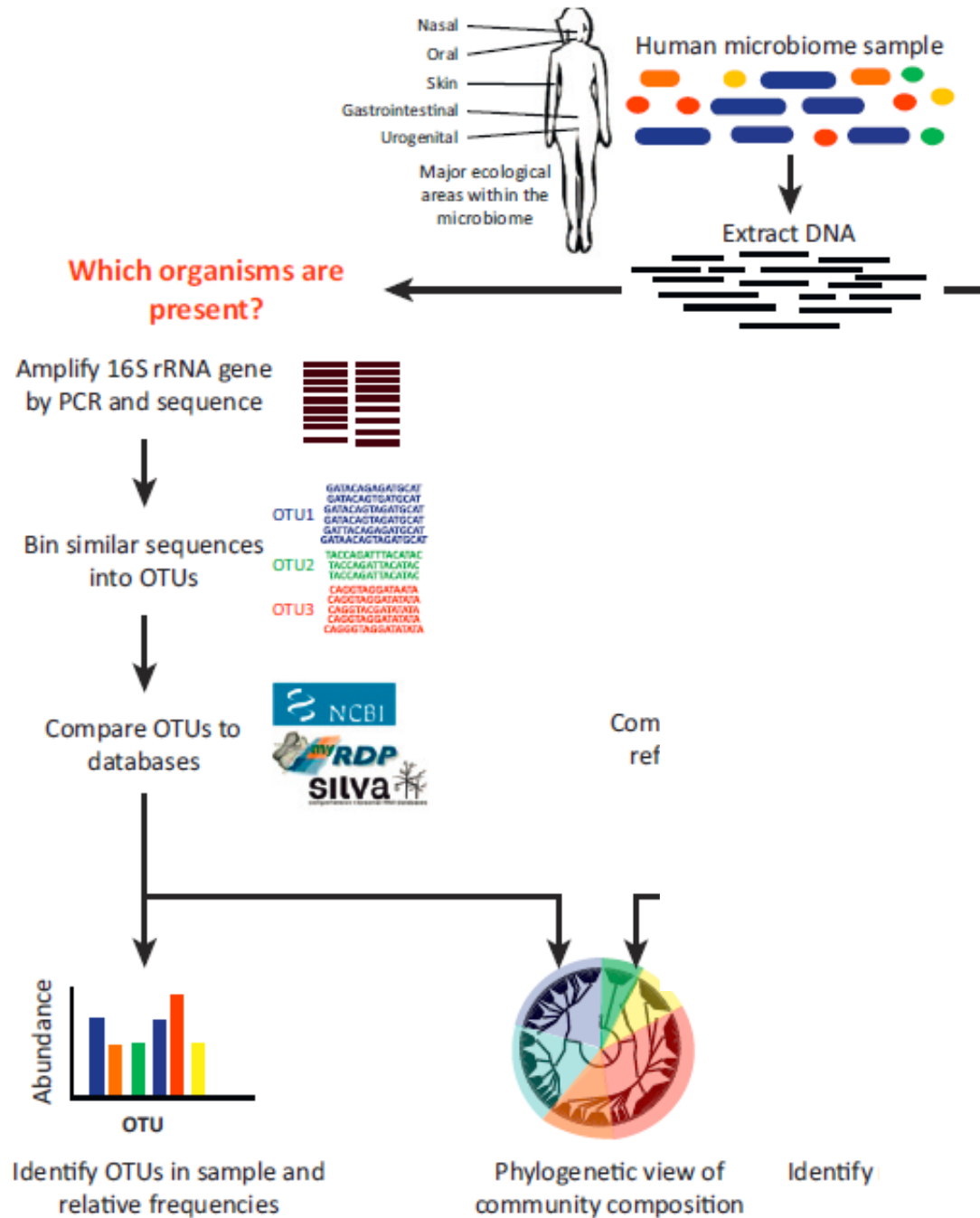


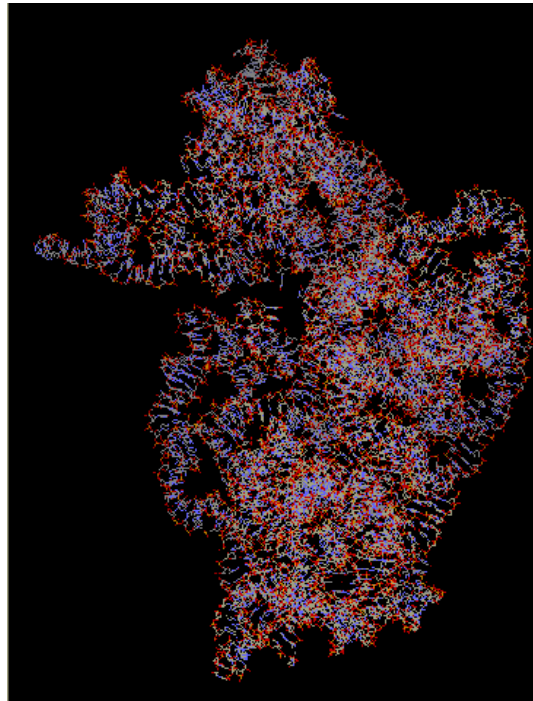
Fundamentals and Overview of bioinformatics pipelines for amplicon sequencing data analysis

Blastocystis COST Action Training School: Blastocystis and the Gut
Microbiome



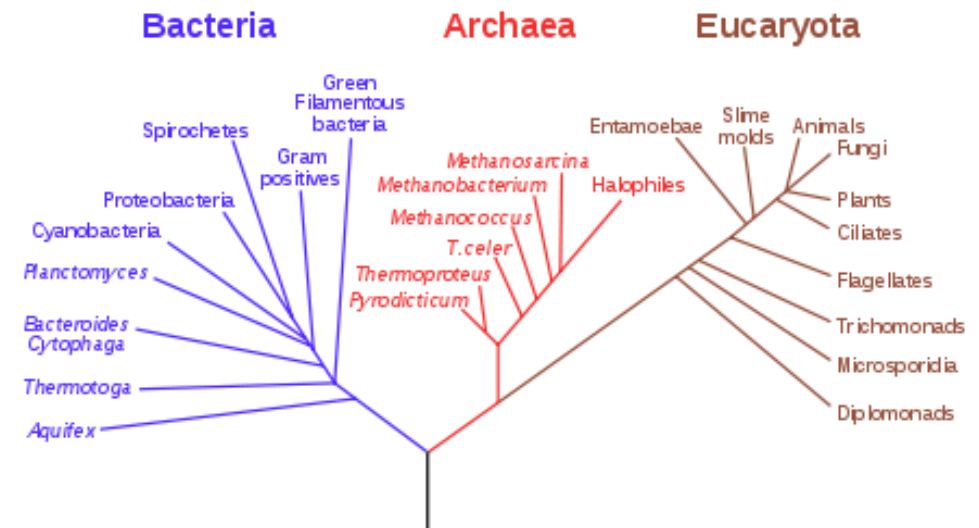
Gut microbiota profiling

Universal marker gene: SSU rRNA gene



<http://www.biochem.umd.edu>

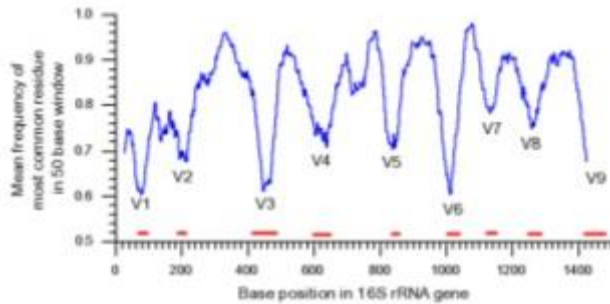
Phylogenetic Tree of Life



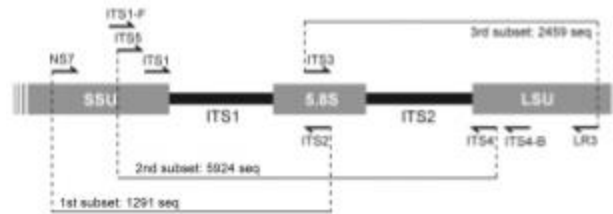
http://en.wikipedia.org/wiki/Carl_Woese

Types of amplicon sequencing

16S rRNA survey of bacterial microbiome



ITS survey of fungal microbiome



- First we need a gene that is universally conserved (also called marker gene)
 - Ribosomal RNA works well
 - SSU (16S, 18S)
 - LSU (23S, 28S)
 - ITS
 - Universally conserved marker
 - E.g. ribosomal proteins
 - Diversity too high to get good primers

Ribosomal marker gene databases

- SILVA – SSU & LSU

- arb-silva.de
- Frequent updates
- Most comprehensive database (SSU & LSU)
- 598,470 (SSU), 96,642(LSU) sequences



- Greengenes – SSU

- greengenes.secondgenome.com
- Last update May 2013
- 1,262,986 sequences



- Unite – ITS

- unite.ut.ee/
- Relatively frequent updates
- 690,548 sequences

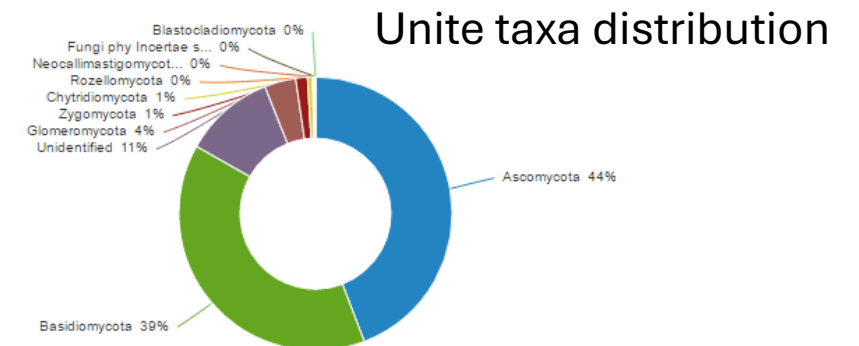


- PR2 – 18S

- specialized on protists

SILVA SSU / LSU 123 - full release

	SSU Parc	SSU Ref	SSU Ref NR	LSU Parc	LSU Ref
Minimal length	300	1200/900	1200/900	300	1900
Quality filtering	basic	strong	strong	basic	strong
Guide Tree	no	no	yes	no	yes
Release date	23.07.15	23.07.15	23.07.15	23.07.15	23.07.15
Aligned rRNA sequences	4,985,791	1,756,783	597,607	563,332	96,642



Error rates

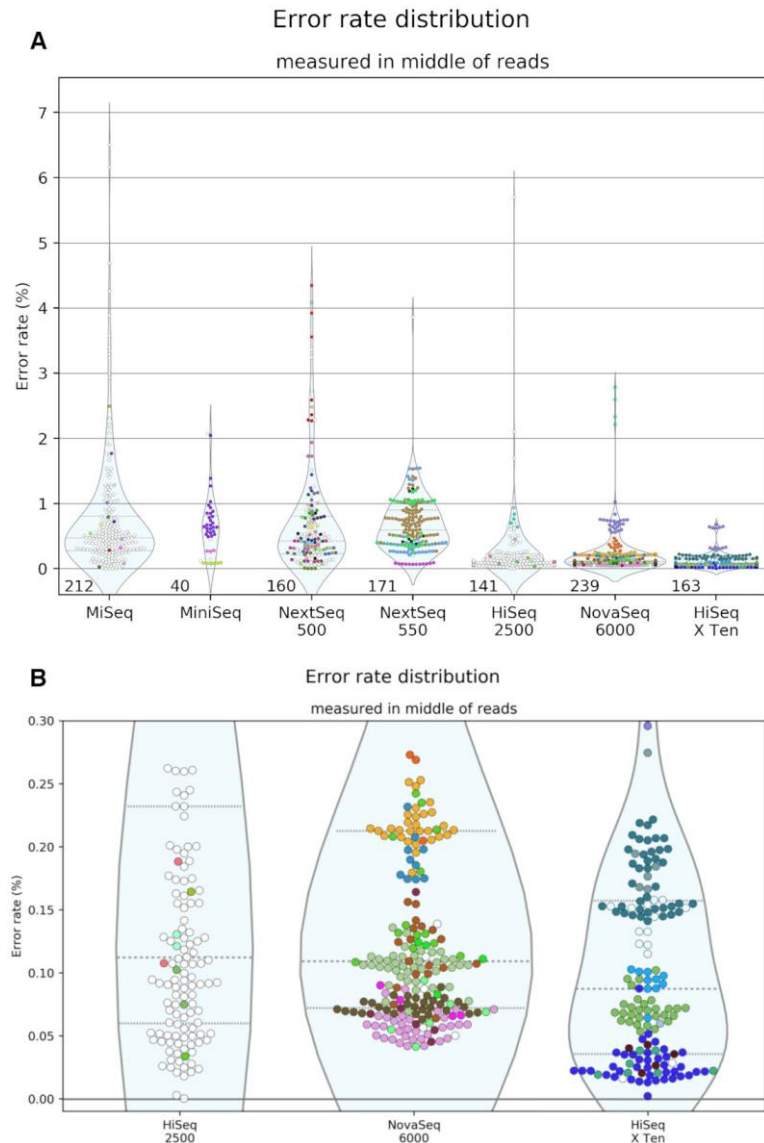


Table 1. Statistics of mappable length and error rates of PacBio and ONT long reads.

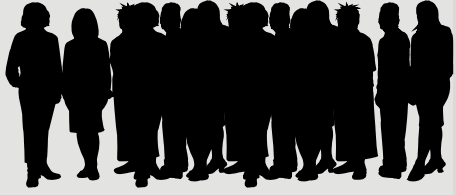
Read type	Mappable length (bp)				Error rate (Proportion of overall error) (%)			
	Mean	Median	Standard deviation	Maximum	Overall	Insertion	Deletion	Mismatch
PacBio CCS	1772	1464	1132	8006	1.72	0.087 (5.06)	0.34 (19.48)	1.30 (75.46)
PacBio subread	1570	1299	1076	16040	14.20	5.92 (41.71)	3.01 (21.17)	5.27 (37.12)
ONT 2D	1861	1754	882	9126	13.40	3.12 (23.30)	4.79 (35.70)	5.50 (40.99)
ONT 1D	1695	1602	824	9345	20.19	2.93 (14.51)	7.52 (37.24)	9.74 (48.25)

The fractions of each error types are in parenthesis. The fractions of the most predominant error types in each data are in bold.

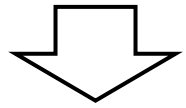
	NS	PacBio	Illumina	Ion Torrent
Read length	Variable (200 bp up to 2 Mbps)	Up to 20 kb	Up to 600 bp (2x300 PE)	Up to 400 bp (SE)
SNV error rate	1%–5%	0.1%*	<0.1%	<0.1%
Indel error rate	5%–10%	4%*	<0.1%	1%

PE, pair-end; SE, single-end; *Error-rate estimation of PacBio circular consensus sequencing (CCS) method.

Population cohort



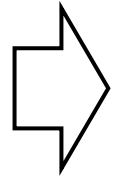
Metadata:
Anthropometrics
Diagnosis
Lifestyle
Medication



Sample collection

Stool

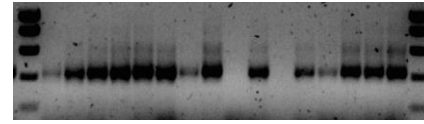
Bead-beating!



DNA



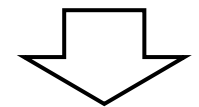
V4 region, 515f-806r



Amplicon visualization
Electrophoresis

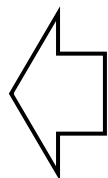


Amplicon quantification
Fragment Analyzer

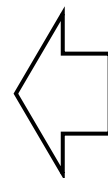


Pooling and library purification

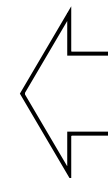
Analysis:
R libraries



Data crunching
Amplicon
NGS pipelines

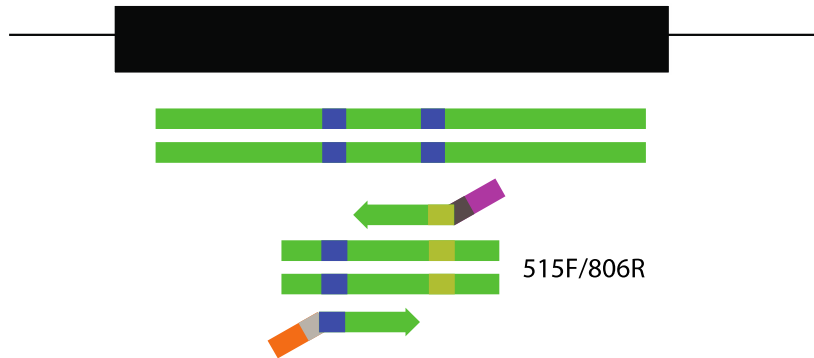


MiSeq / HiSeq 2x250



Dual Index and controls!

16S rRNA



TruSeq adaptor Index1 primer



CAAGCAGAAGACGGCATAACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT NNNNNNNNNNN GTGYCAGCMGCCGCGGTAA

515F

TruSeq adaptor Index2 primer

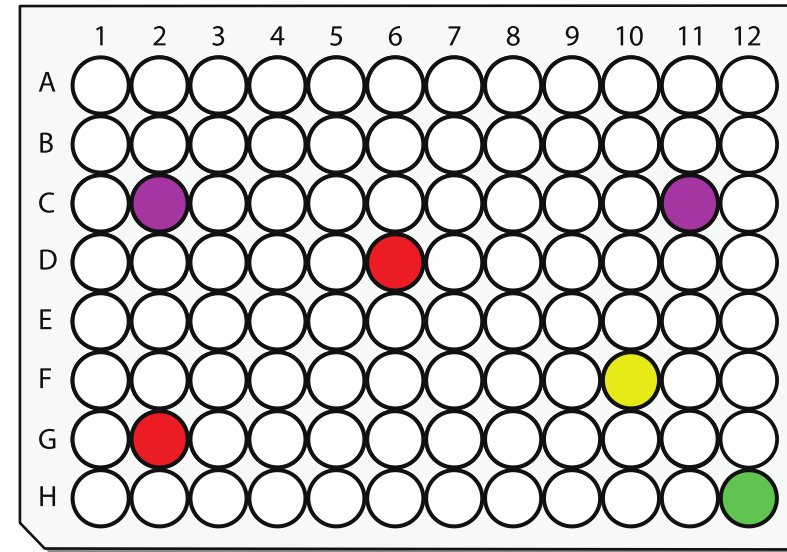


AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCT NNNNNNNNNNN GGACTACNVGGGTWCTAAT

806R

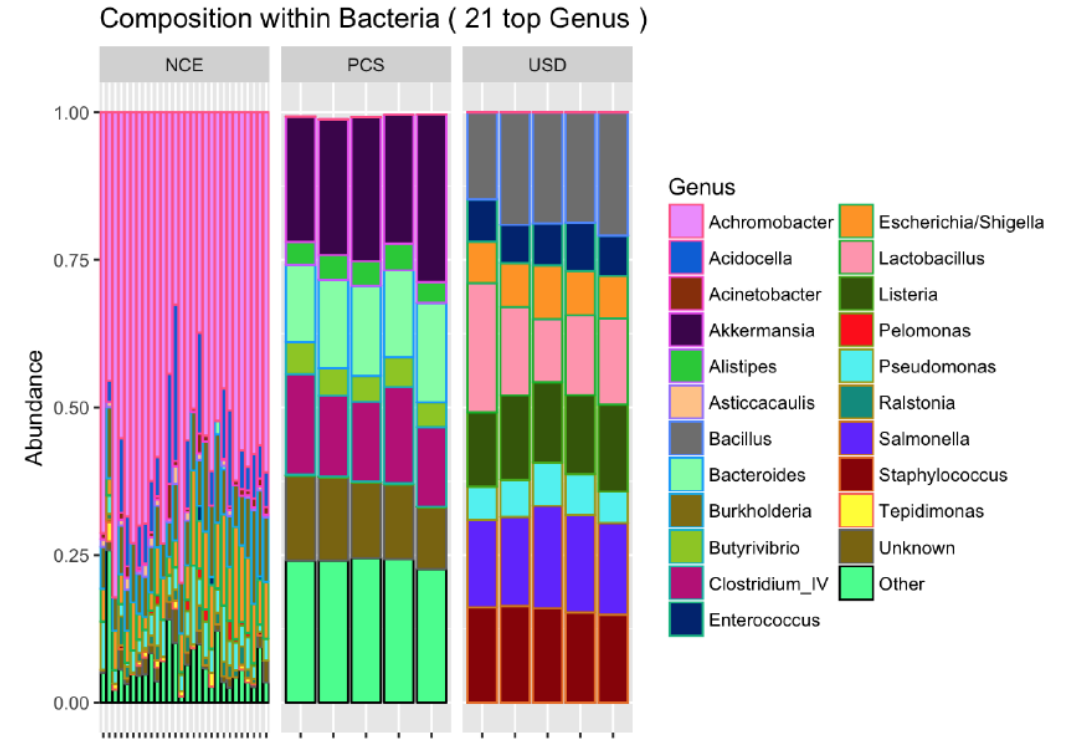
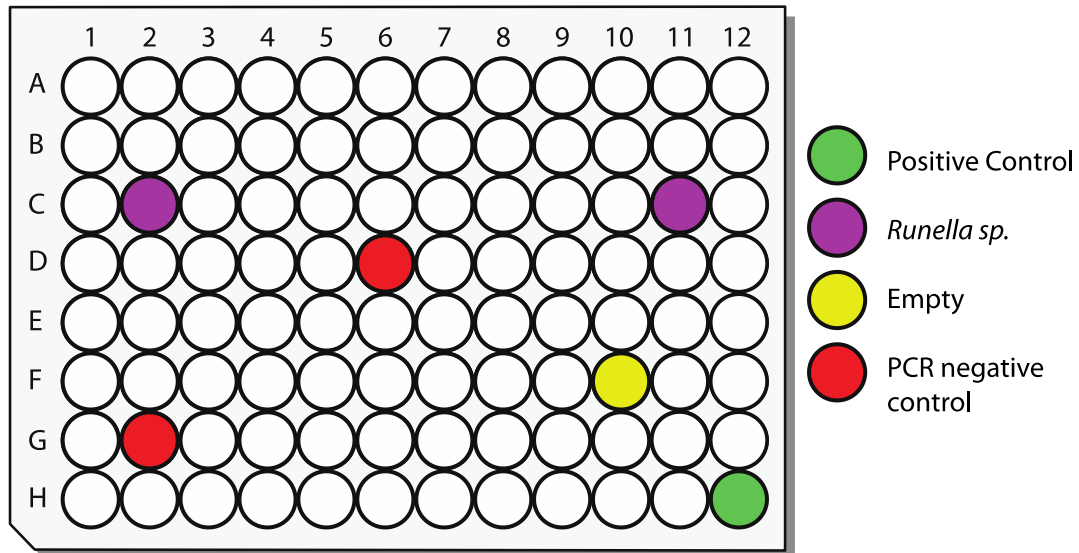
R2

R1

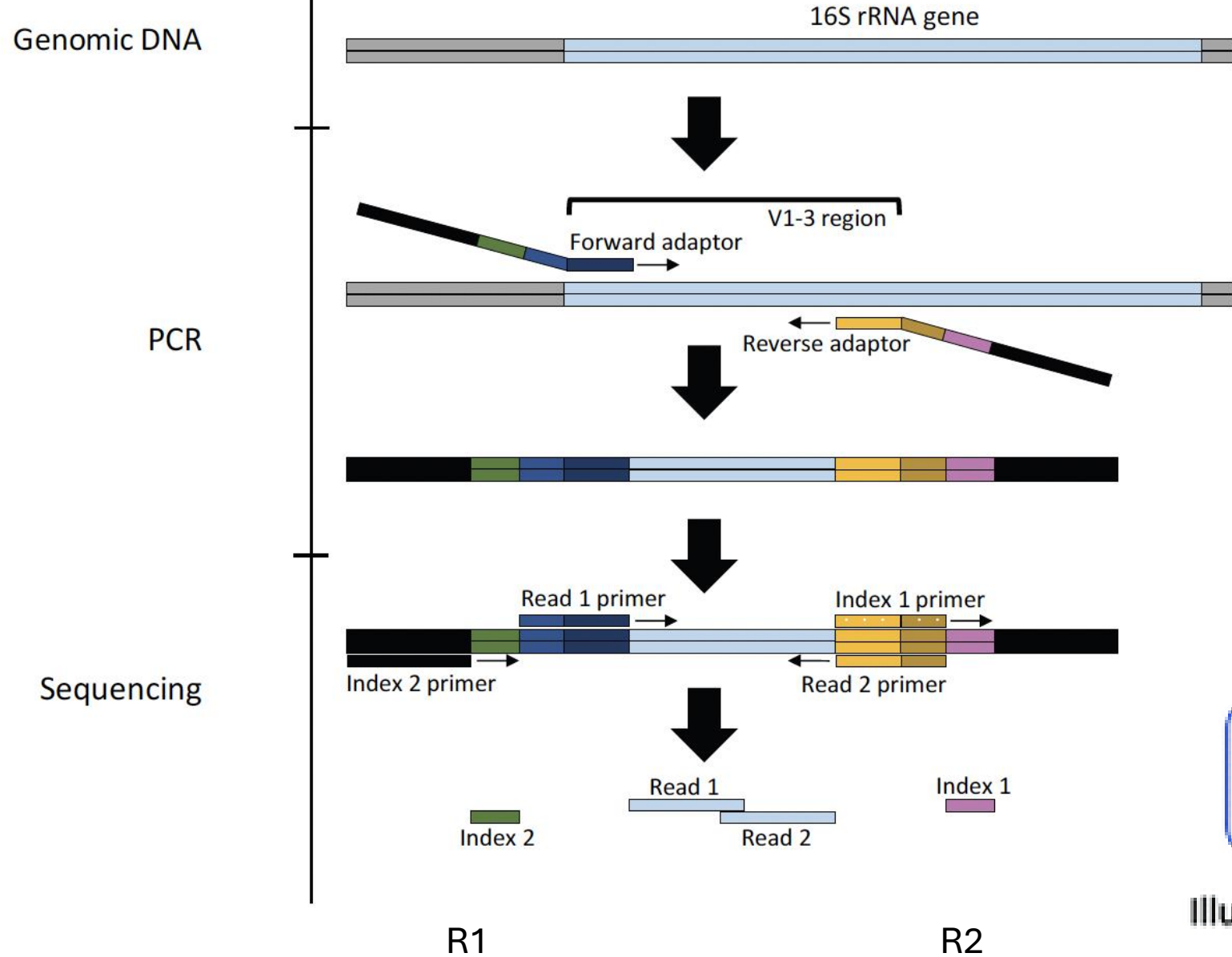


- Positive Control
- *Runella sp.*
- Empty
- PCR negative control

Dual Index and controls!

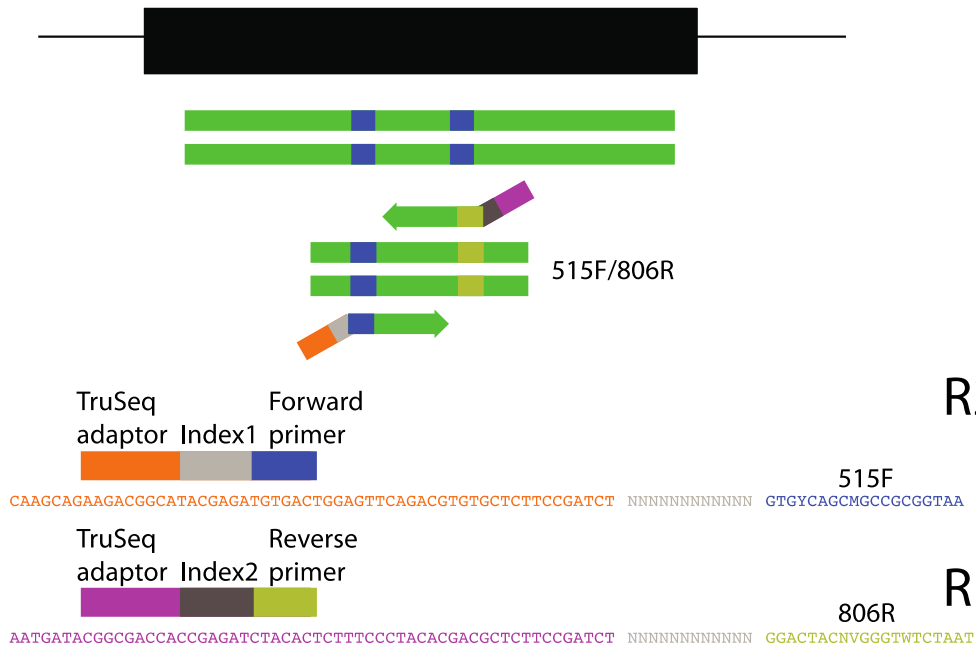


General pre-processing flow



Sequencing format (MiSeq, HiSeq, Novaseq)

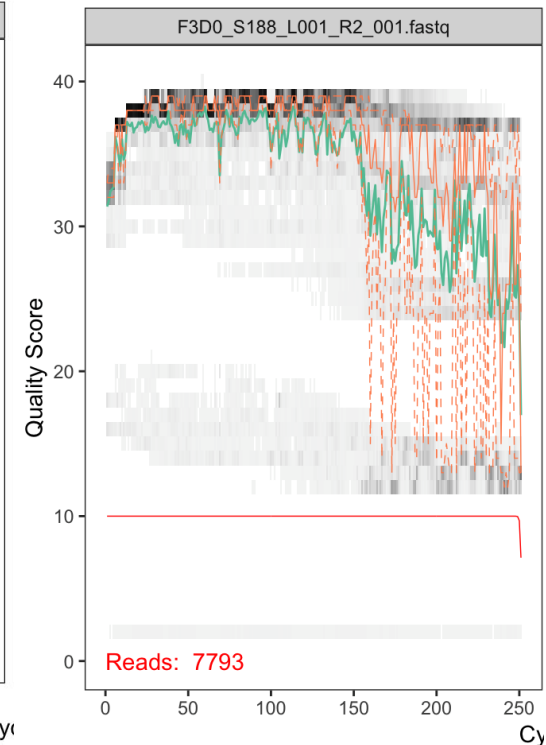
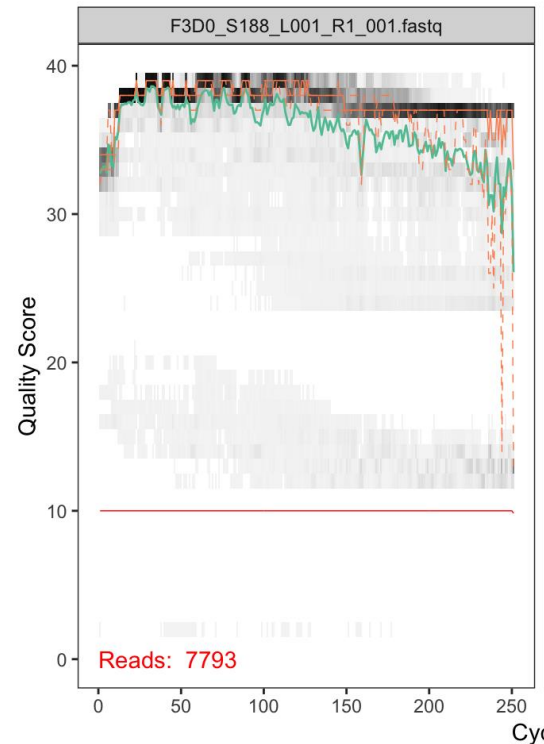
16S rRNA



MiSeq PE 2X250
 ~ 12-18 millions of reads PE
 R1 and R2
 Fastq files

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%



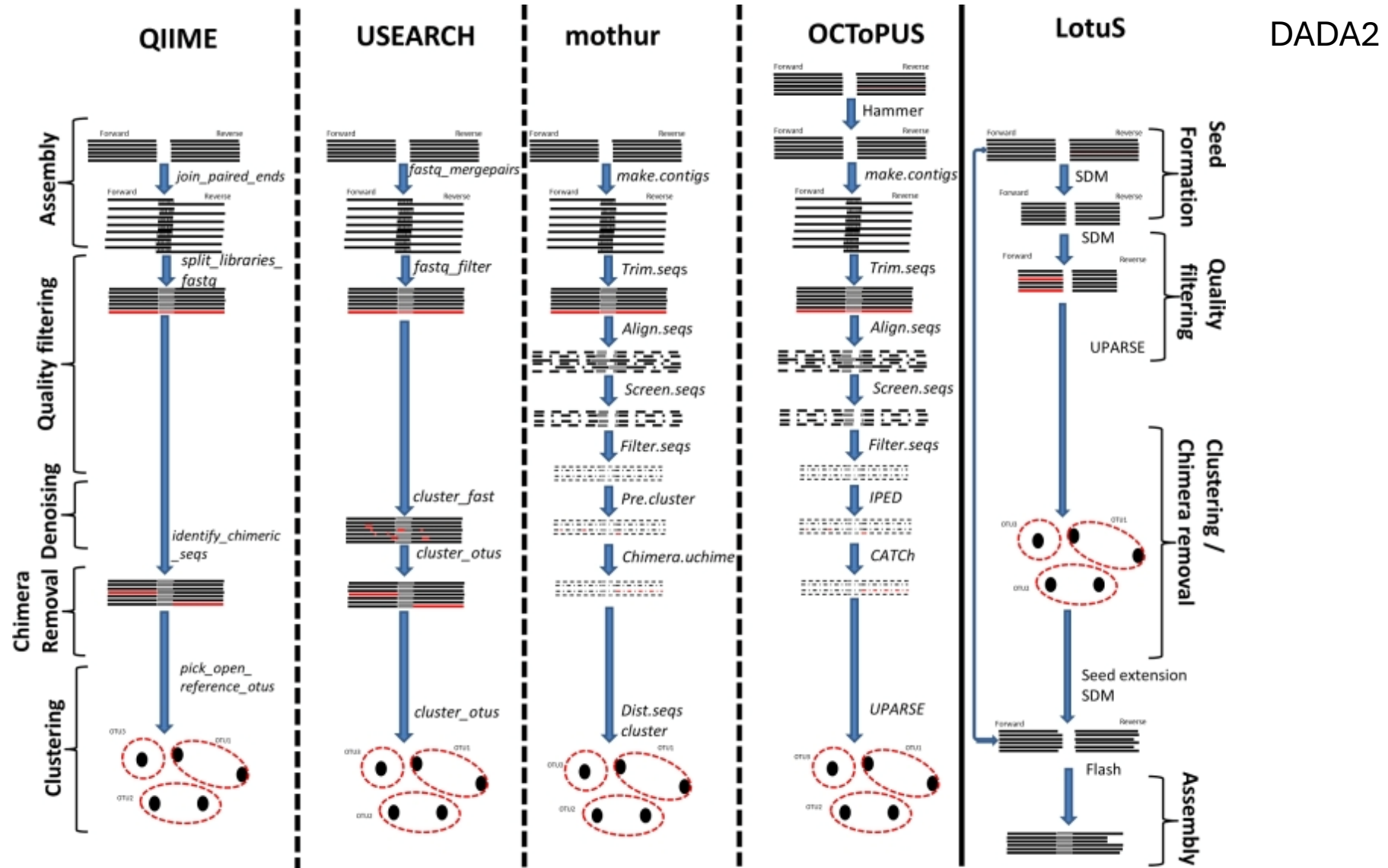
Data crunching: general steps

Basic pipeline for 16S data preprocessing

- 1. Demultiplexing**
- 2. Quality control and trimming**
- 3. Merging R1 and R2**
- 4. Chimera removal**
- 5. Closed-reference OTU picking**
- 6. De novo OTU picking**
- 7. Taxonomic annotation**
- 8. Tree building**

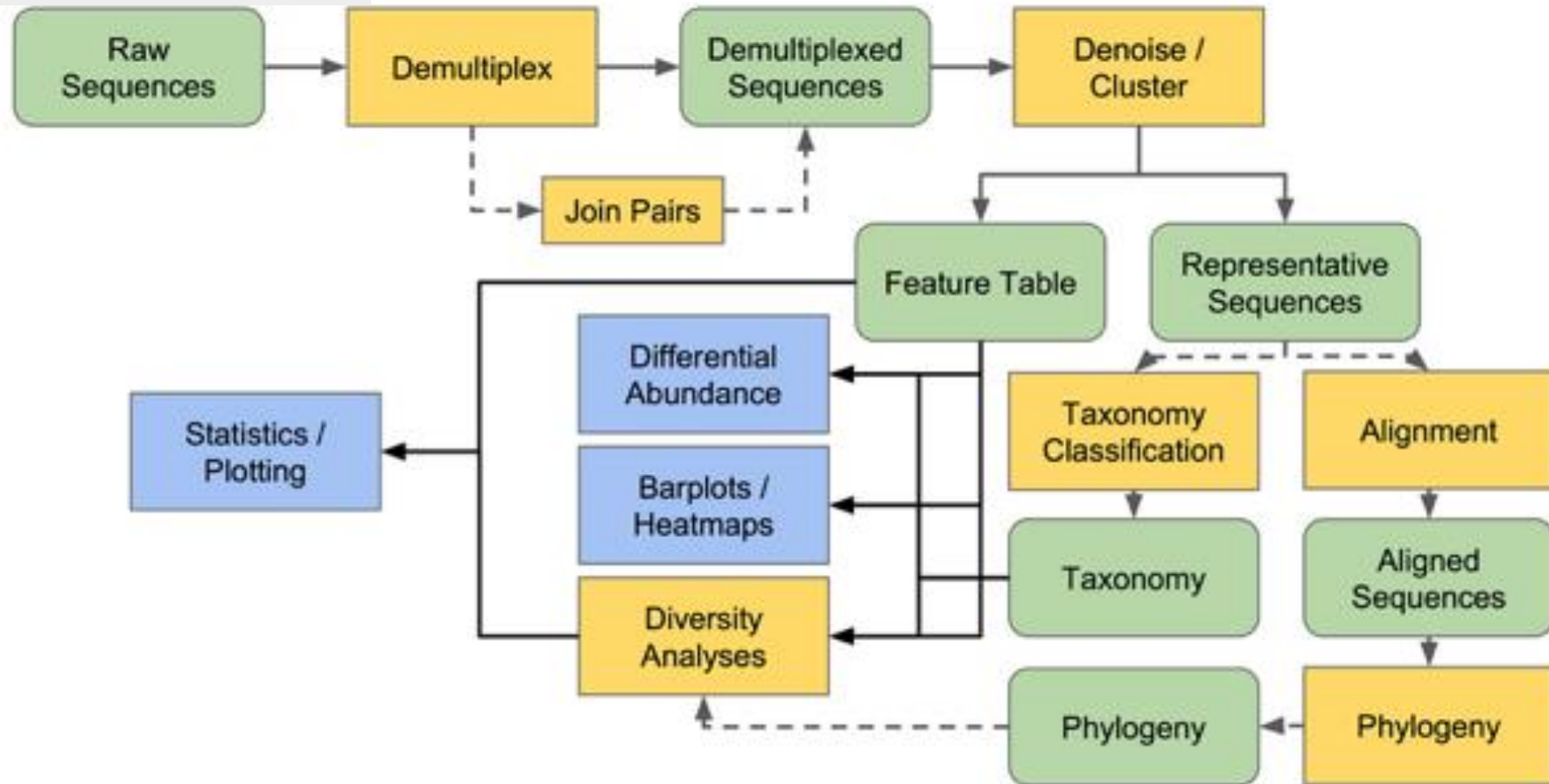
Pipelines to analyzed microbiota
amplicon sequencing data
(ribosomal markers genes)

Different flavors

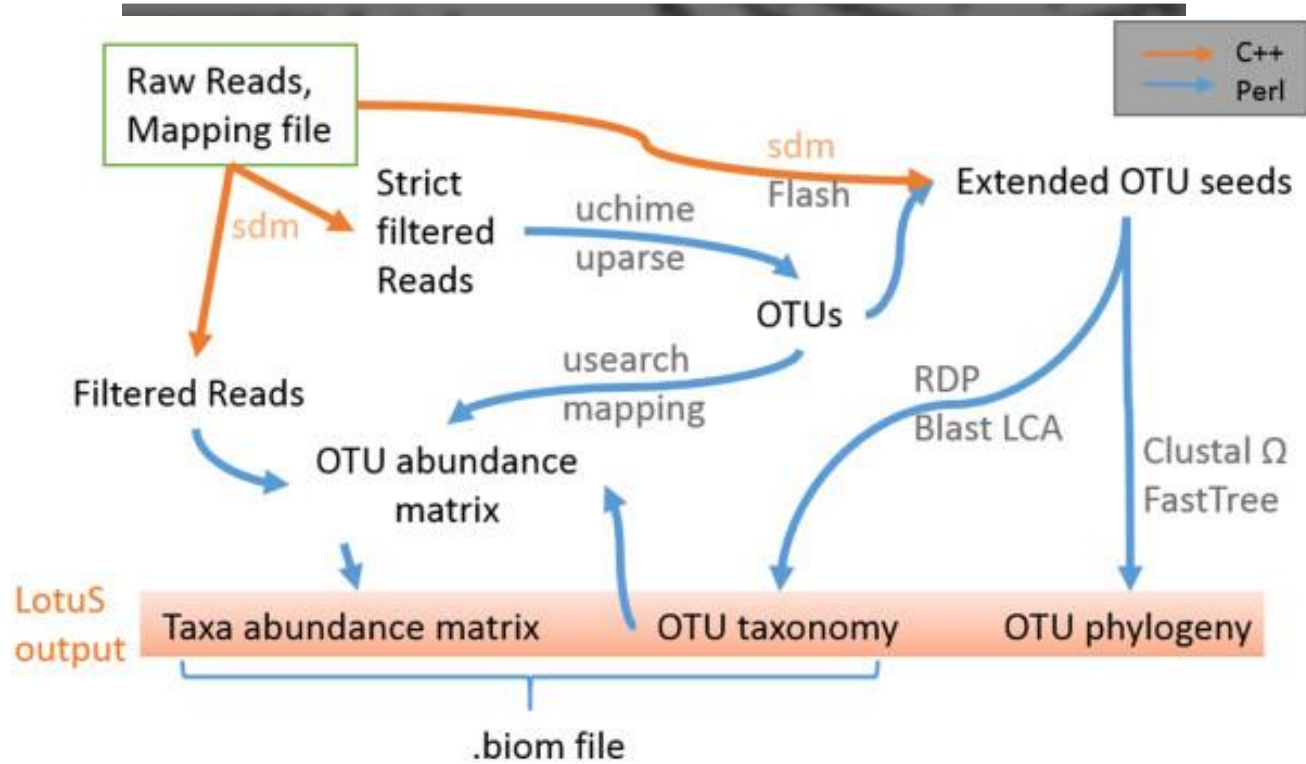




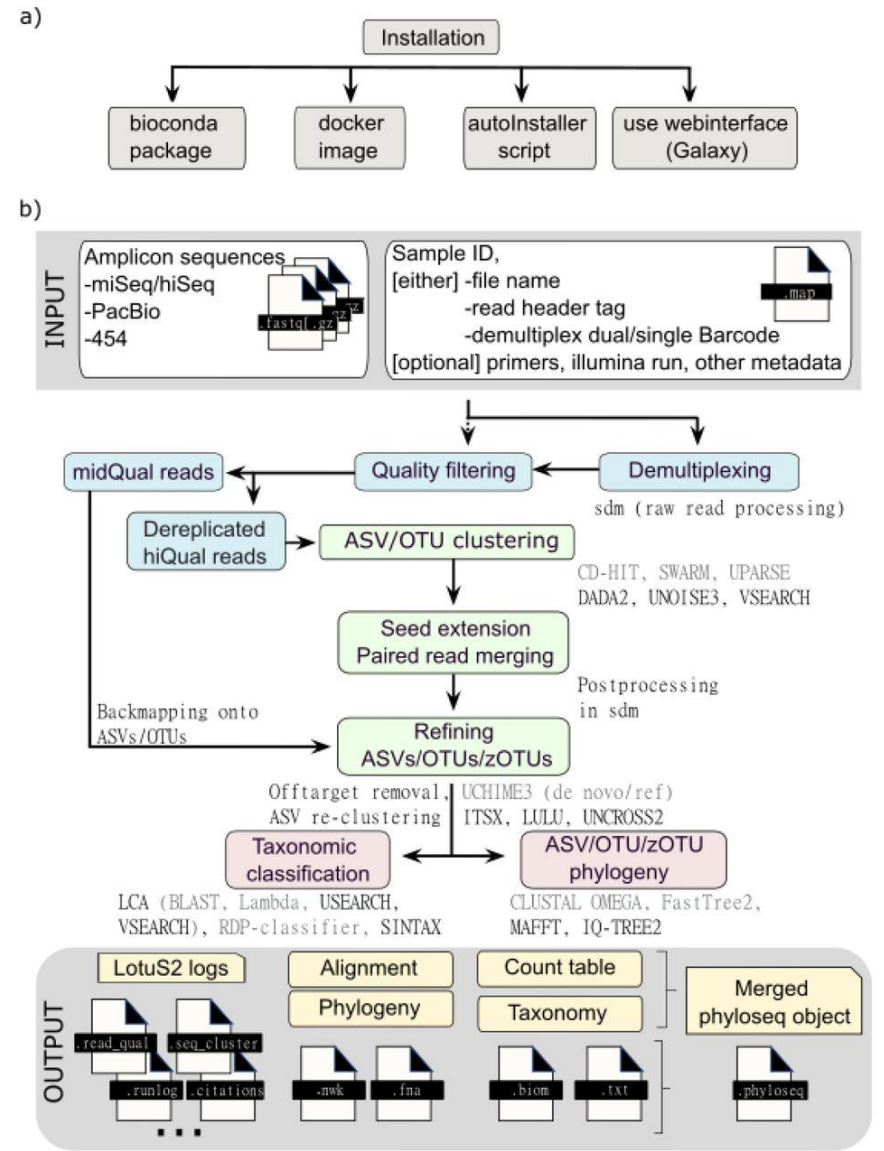
QIIME2 Pipeline

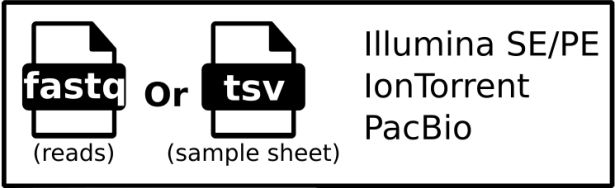


LotuS



<https://github.com/hildebra/lotus2?tab=readme-ov-file>





--- Or ---



Quality control

Primer trimming
cutadapt

Remove multiple or read-through primers
cutadapt

Evaluation
FastQC
plotQualityProfile

Quality filtering
filterAndTrim

Infer Amplicon Sequence Variants (ASVs)

DADA2
derepFastq
learnErrors
dada
removeBimeraDenovo
mergeSequenceTables

Extract ITS region
ITSx

Annotate rRNA
Barnap

Taxonomic classification

DADA2
cutadapt
assignTaxonomy
addSpecies

QIIME2
extract-reads
fit-classifier-naive-bayes
classify-sklearn

Reference taxonomy

SILVA UNITE DADA2 only

PR2 GTDB RDP

Taxonomic filtering
taxa filter-seqs
taxa filter-table

Abundance & prevalence filtering
feature-table filter-features

Abundance tables
taxa collapse
feature-table relative-frequency

Visualisation
barplot

Differential abundance
ANCOM

Alpha- & beta-diversity
qiime diversity
qiime diversity adonis

Quality control
alpha-rarefaction

Predict function
PICRUSTt2

Reporting
MultiQC

Visualisations & tables
html tsv fasta
biom nwk

Legend

default (green rounded rectangle)
on demand (white rounded rectangle)
mandatory (black rounded rectangle)
optional (dashed black rounded rectangle)



CC-BY 4.0. Design originally by Zandra Fagernäs

Pipelines

Input data: 16S Illumina raw data (pair end reads). Primer V3–V4 region

QIIME2
(Bolyen, 2019)

Bioconductor
(Callahan 2016b)
v 29 OCT 2018

USEARCH
(Edgar, 2010)

mothur
(Schloss, 2009)
v 17/10/2018

Main bioinformatic steps

Pair reads merging

Primer trimming
Poor quality read removal

Denoise
Dereplication
Error rate estimation
Substitution and indel errors removal
Chimera filtering

Sequence clean-up

Primer trimming
Poor quality read removal

Dereplication
Error rate estimation
Substitution and indel errors removal

Pair read merging

Primer trimming
Poor quality read removal
Dereplication
Discard singletons

Primer trimming
Poor quality read removal

Dereplication
Alignment
Clean alignment
Dereplication
Precluster
Chimera filtering
Non bacterial sequence removal

Pair read merging

Generating OTU/RSV table

Chimera filtering

Sequence clean-up:
Chimera filtering

Taxonomic assignments using SILVA 132 ribosomal RNA (rRNA) database as reference

Outcome

ASVs

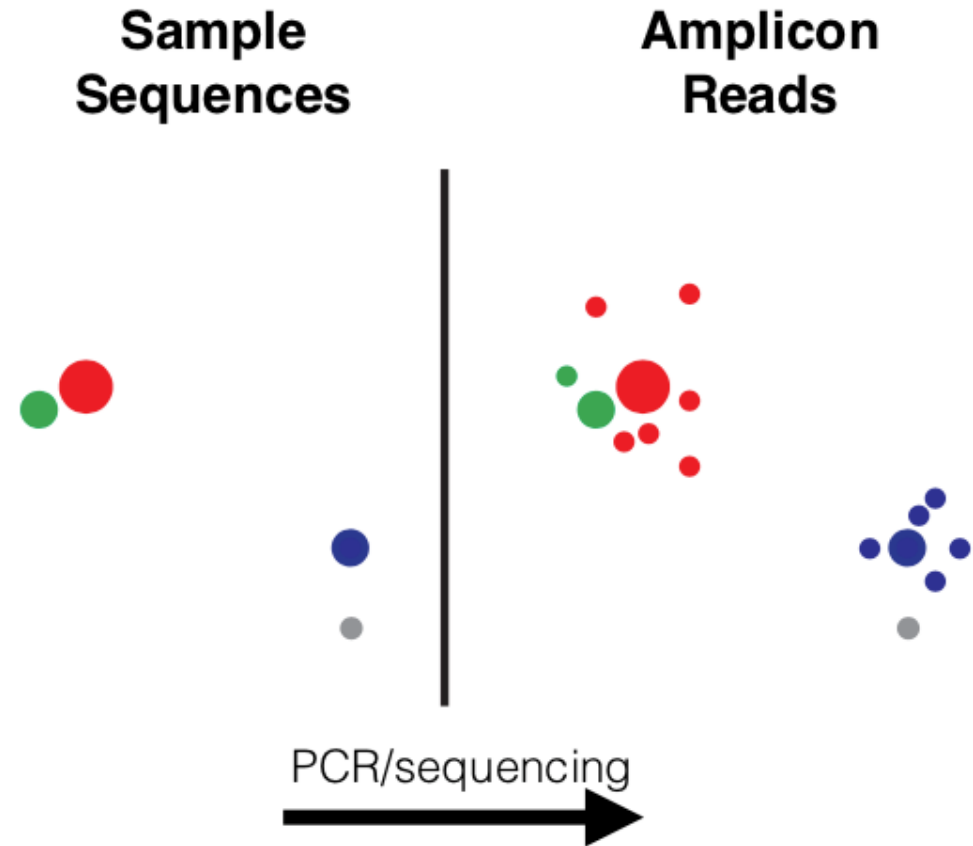
ASVs

OTUs

OTUs

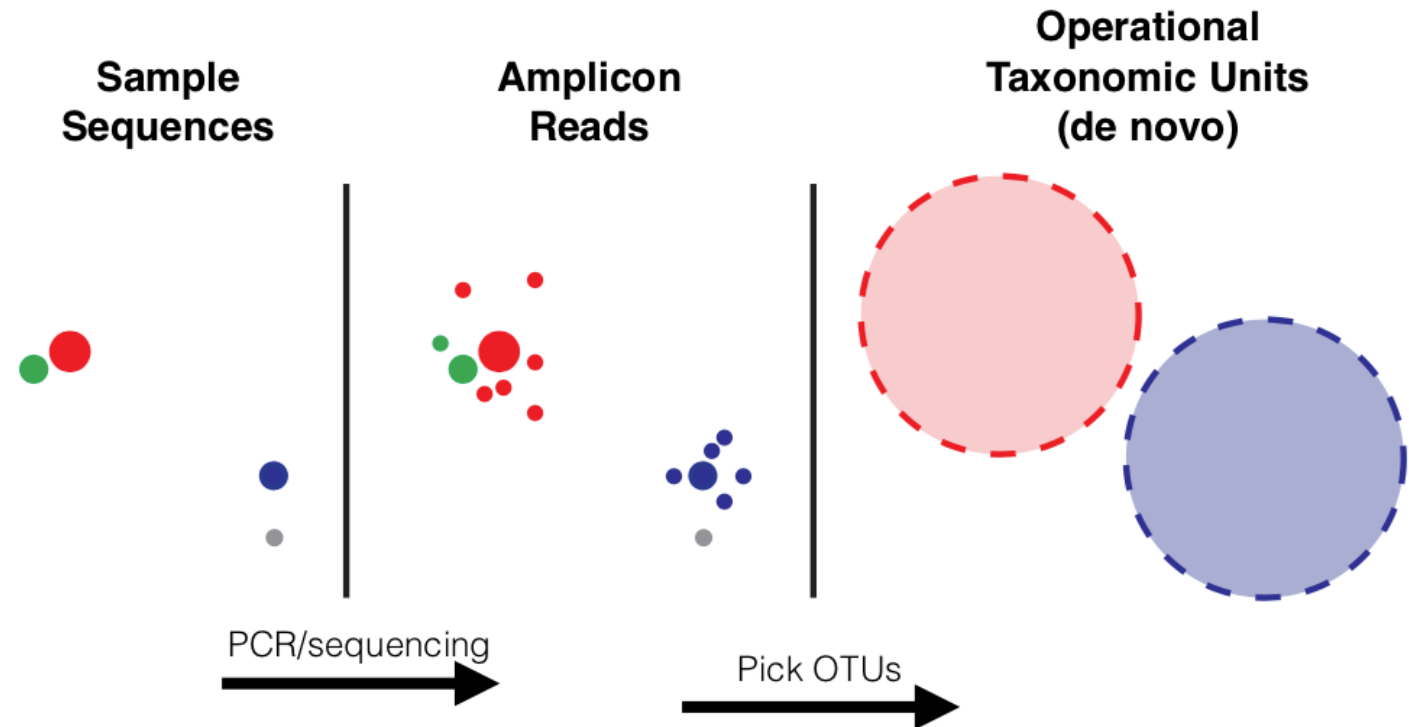
Original sequences and amplified sequences

GCGGC	TCAAC	CGTAA	AATTG	CAGTT
GCGGC	TCATC	CGTAA	AGTTG	CAGTT
.....	...Y.R...
GGTGC	TTAAC	CCTAA	AATCG	CAGTT
GGTGC	TTAAC	CGTAA	AATCT	CAGTT
.....S...K



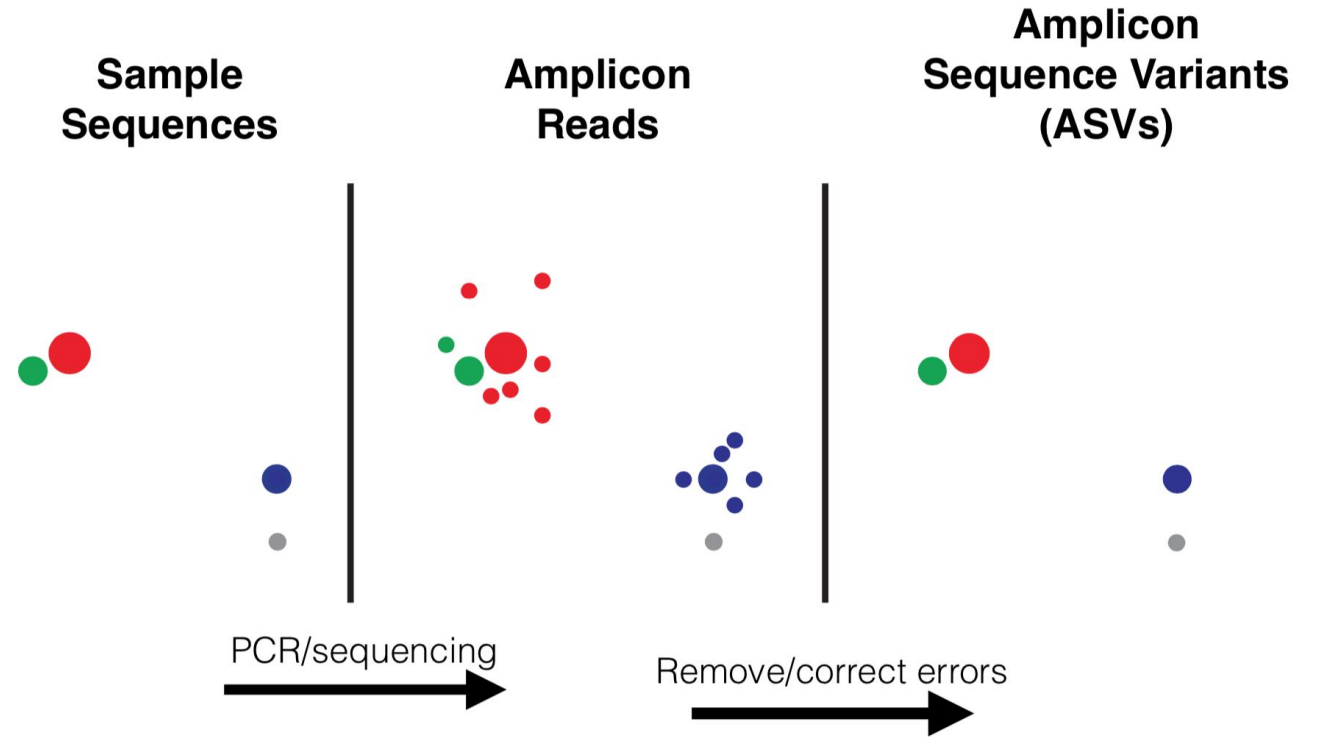
Operational Taxonomic Unit (OTU)

OTUs are defined as a cluster of sequences with an identity above a given threshold, often set to 97% .



Amplicon Sequence Variant (ASV)

'True' biological sequences after a denoising procedure to control for errors of the sequencing technology per batch/run.



OTUs vs ASVs

OUT @ 97%

ASV

Can be subject to reference bias

A reference is not used until after taxonomy assignment

OTU-tables cannot be combined between studies

ASV-tables can be combined across studies

Represented by a consensus sequence

Represented by an exact sequence

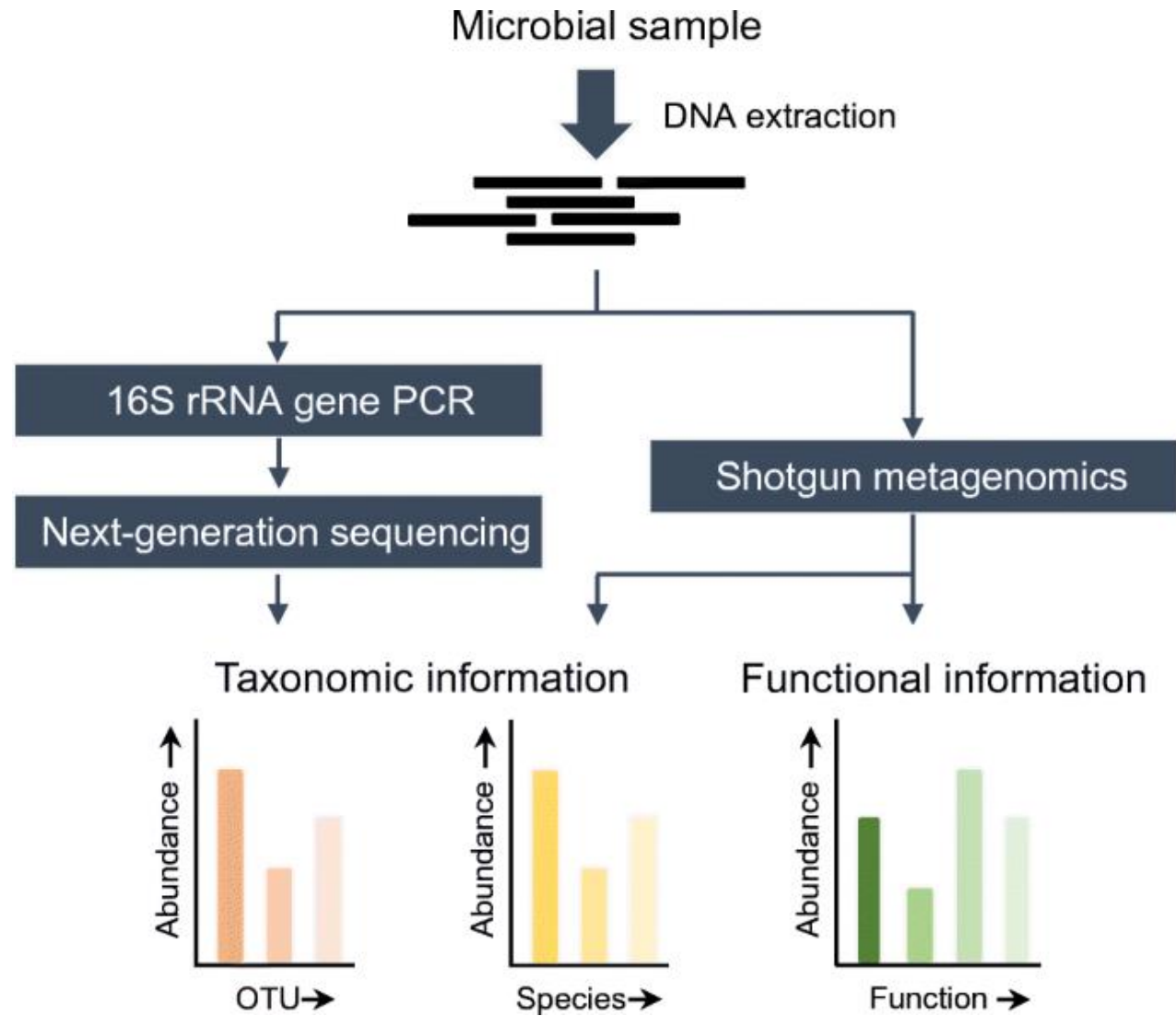
A consensus sequence can represent multiple species with different sequences

If it represents multiple species, it is because they share an identical sequence

Limitations and biases

Experimental	Mitigation
Step 1: sample collection	
Transport and storage conditions	Immediate freezing at – 20 °C or lower, followed by long-term storage at – 80 °C.
Step 2: DNA extraction	
Different methods	The same method should be used in a whole project
Step 3: PCR amplification	
No 16S rRNA gene PCR primer pair is truly ‘universal’ and different primer pairs may have different proportions of ‘conserved’ sequences.	The same method should be used in a whole project
All protocols are sensitive to contaminating DNA throughout the process	Negative (extraction) controls should be included
Step 4: Next-generation sequencing	
Short sequences (few hundred bases)	Keep method consistent within a project
Step 5: Bioinformatics analysis	
16S rRNA gene NGS results are generally presented as proportional abundances of OTUs/ASVs	The use of protocols that determine the absolute quantity of OTUs/ASVs improves the interpretation.

Limitations and biases



Take home message

- There are numerous options for analysis of amplicon sequencing, be consistent within a project.
- Whichever method is selected, be aware of the limitations and acknowledge them as a part of your discussion.
- There are multiple free web resources/pipelines for omics data processing, filtering and analysis, be aware of default settings and options as they may not be comparable from tool to tool.