Metabarcoding Pipelines Overview

Ali Hakimzadeh

Ali.hakimzadeh@ut.ee

UT Summer School

August 2022

Overview

- Universal Pipelines
  - Anacapa Toolkit
  - Barque
  - Cascabel
  - CoMA
  - DADA2
  - FROGS
  - MetaWorks
  - OBITOOLS
- Marker Specific Pipelines
  - Nf-core / Ampliseq
  - BIOCOM-PIPE
  - DAnIEL
  - EzMAP
  - gDAT
  - JAMP
  - LOTUS2
  - MICCA
  - PEMA
  - PIPITS
  - QIIME2
  - SEED 2
- Benchmarking studies

Anacapa Toolkit

Linux

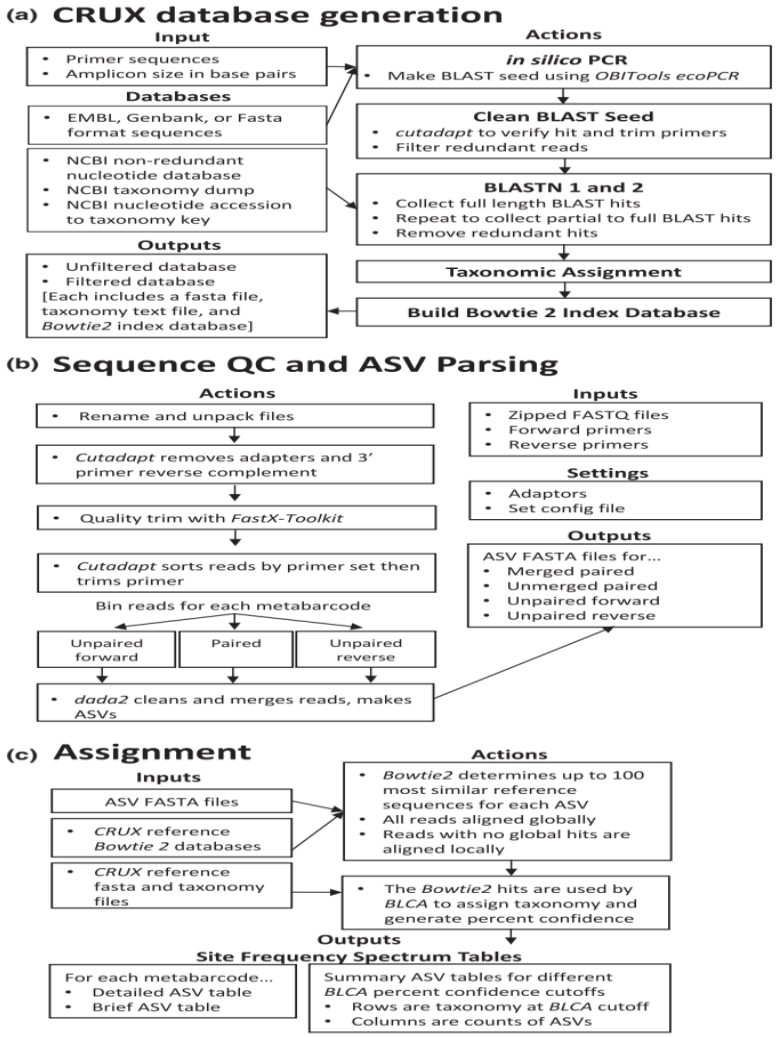
Paired-end short reads

Generates only ASV

Four modules
- CRUX
- DADA2
- Bowtie2
- ranacapa

NO demultiplexing and post clustering curation steps

Most used for Marine metabarcoding



Curd, et al.2019 *Methods in Ecology and Evolution*

Barque

Linux, macOS

Paired-end short reads

Annotates reads instead of generating OTUs

Can also produce OTUs (Annotating the reads with the OTUs that were previously found)

Only All steps in one go (Trimmomatic, Flash, and VSEARCH), no post-processing step

- More focused on species level ➡ Monitoring Invasive species and confirming the presence of specific species
- Mostly useful for COI and 12S amplicons

Mathon et al, 2021 Molecular ecology

Cascabel

Linux, macOS (Snakemake workflow manager)

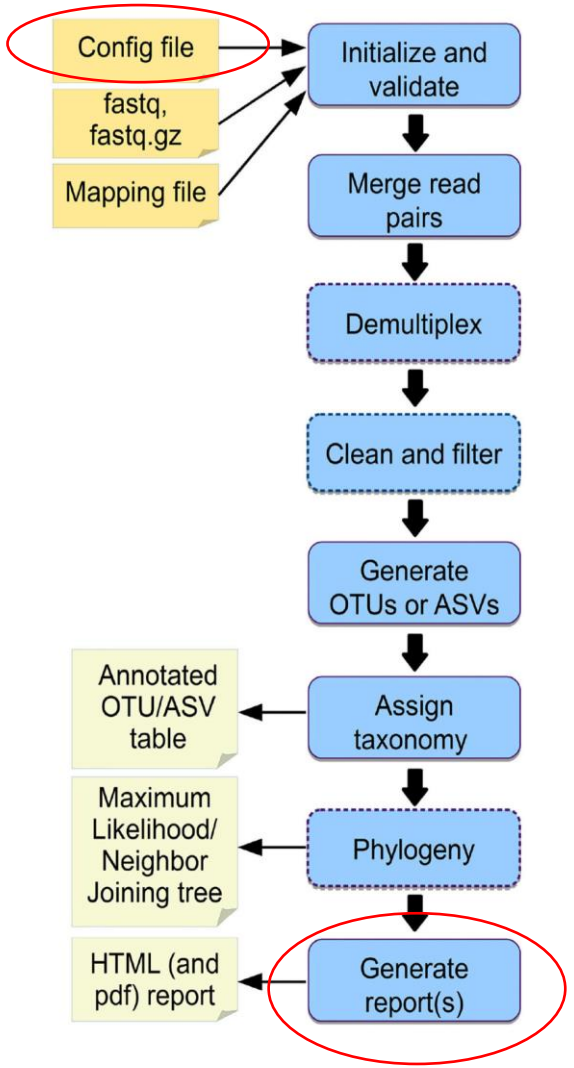
Paired-end short reads

Generates both ASV and OTU

Versatile software for customizing several steps

All analyses (runs) performed are completely documented and reproducible

Interactive and non-interactive mode



Step	Tools/Algorithms	Output
Initialize structure	Script	Project folder and file structure
Quality Control	FastQC (Andrews, 2010)	FastQC report
Merge reads	PEAR (Zhang et al., 2014)	Merged (assembled) sequences
Demultiplex	QIIME (Caporaso et al., 2010b), scripts	Sequences assigned to samples in one file and per sample
Align vs. reference	Mothur (Schloss et al., 2009)	Aligned sequences
Remove chimeras	usearch61 (Edgar, 2010), Uchime_denovo and uchime_ref (VSEARCH) (Rognes et al., 2016)	Chimera-free sequences
Remove adapters	Cutadapt (Martin, 2011)	Adapter-free sequences
Size filter	Script	Filtered sequences
Dereplicate	VSEARCH	Dereplicated sequences
Generate OTUs	Mothur (Schloss et al., 2009), prefix/suffix (Caporaso et al., 2010b), CD-HIT (Li and Godzik, 2006), SUMACLUST (Kopylova et al., 2016), Swarm (Mahé et al., 2015), UCLUST (Edgar, 2010), trie (Caporaso et al., 2010b) sortmerna (Kopylova et al., 2012)	OTU table
Pick representatives (OTUs)	Random, longest, most_abundant, first	Fasta file with representative sequences
Generate ASVs	DADA2 (Callahan et al., 2016)	ASV table
Assign taxonomy OTUs	QIIME [BLAST (Altschul et al., 1990), UCLUST, RDP (Wang et al., 2007)], blastn (BLAST+) (Camacho et al., 2009), VSEARCH	Taxonomic assignments for each OTU
Assign taxonomy ASVs	RDP	Taxonomic assignments for each ASV
Generate OTU table	QIIME, scripts	Annotated OTU table
Generate ASV table	DADA2	Annotated ASV table
Alignment	Pynast (Caporaso et al., 2010a), mafft (Katoh and Standley, 2013), infernal (Nawrocki and Eddy, 2013), clustalw (Larkin et al., 2007), muscle (Edgar, 2004)	Multiple sequence alignment
Make tree	Muscle, clustalw, raxml (Stamatakis, 2006), fasttree (Price et al., 2009)	Phylogenetic tree
Report	Scripts, Krona (Ondov et al., 2011)	HTML, pdf report, Krona charts

Abdala Asbun et al. 2020 frontiers in genetics

CoMA

Linux, macOS, Windows (GUI)

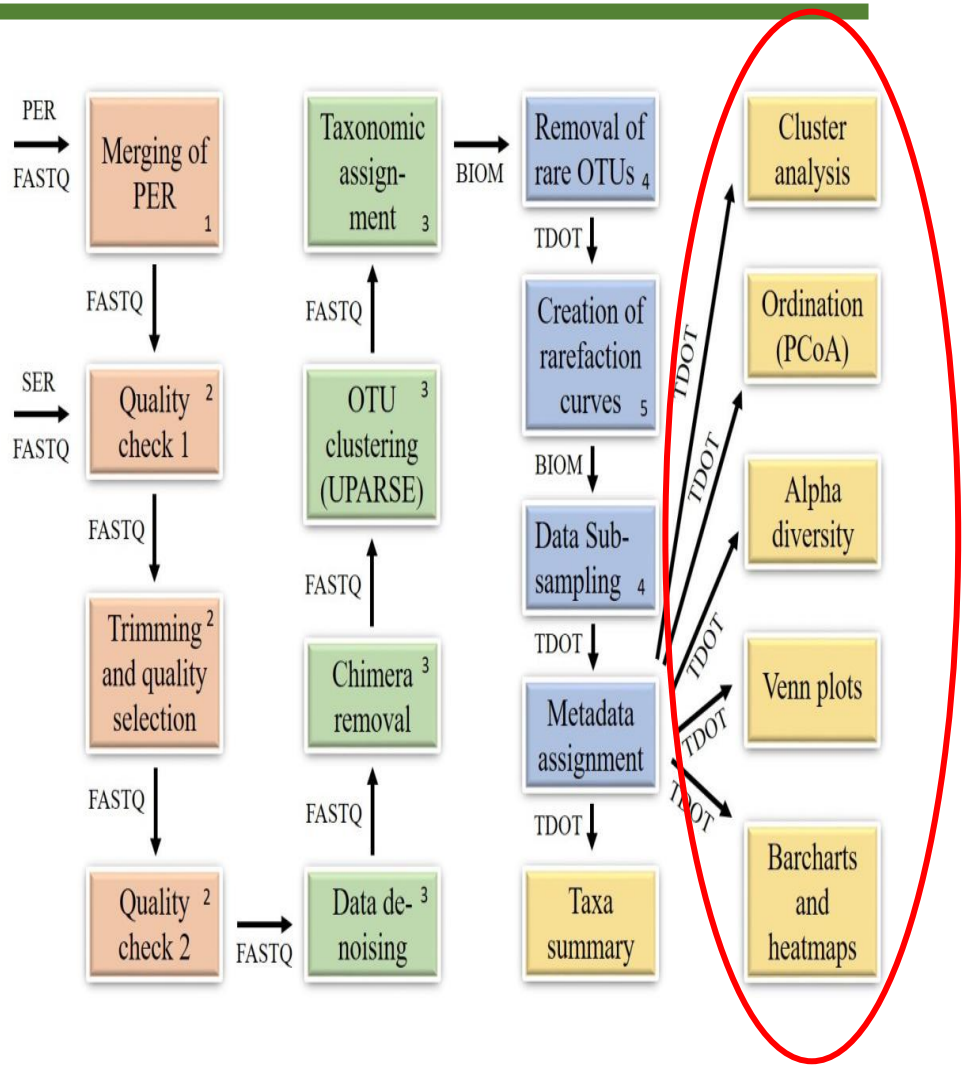
Paired-end short reads

Generates only OTU tables

Data post-processing

Data visualization

statistical appraisal



Hupfauf et al. 2020 PLOS ONE

DADA 2

Linux, macOS, Windows (R-based)

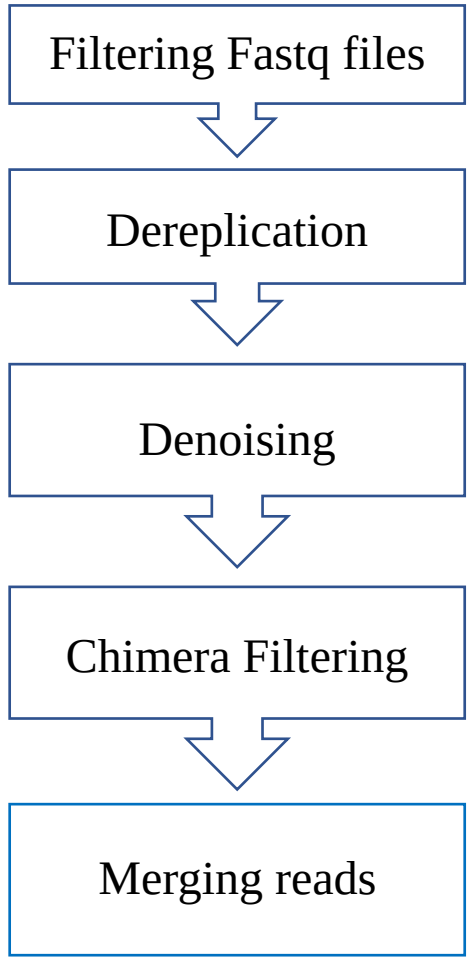
Single-end long reads, Paired-end short reads

Generates only ASV Fasta files

High Accuracy but slow

Species-level analysis

Performs merging of paired-end reads



Callahan et al, 2016 Nature Methods

FROGS

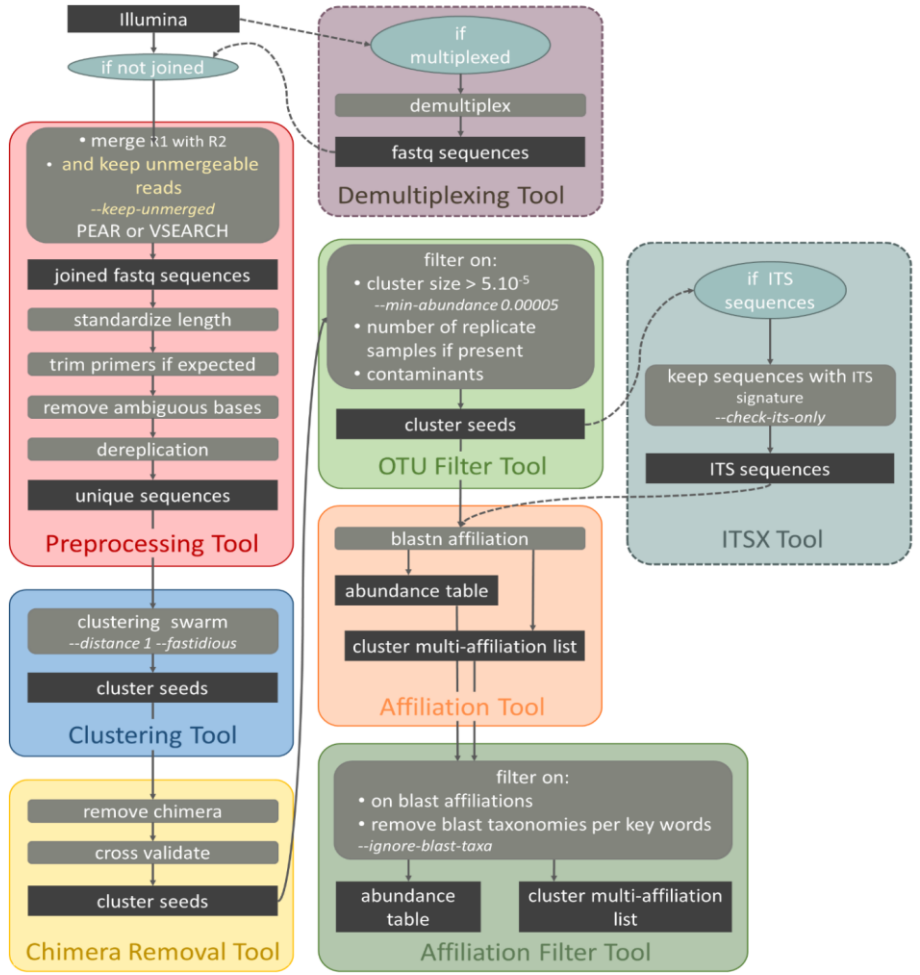
Linux, macOS (Conda)

Single-end long reads, Paired-end short reads

Generates only OTU tables

ITSx

Two OTU handling mode
- Mask (NA)
- Delete



Bernard et al. 2021 Briefing in bioinformatics

MetaWorks

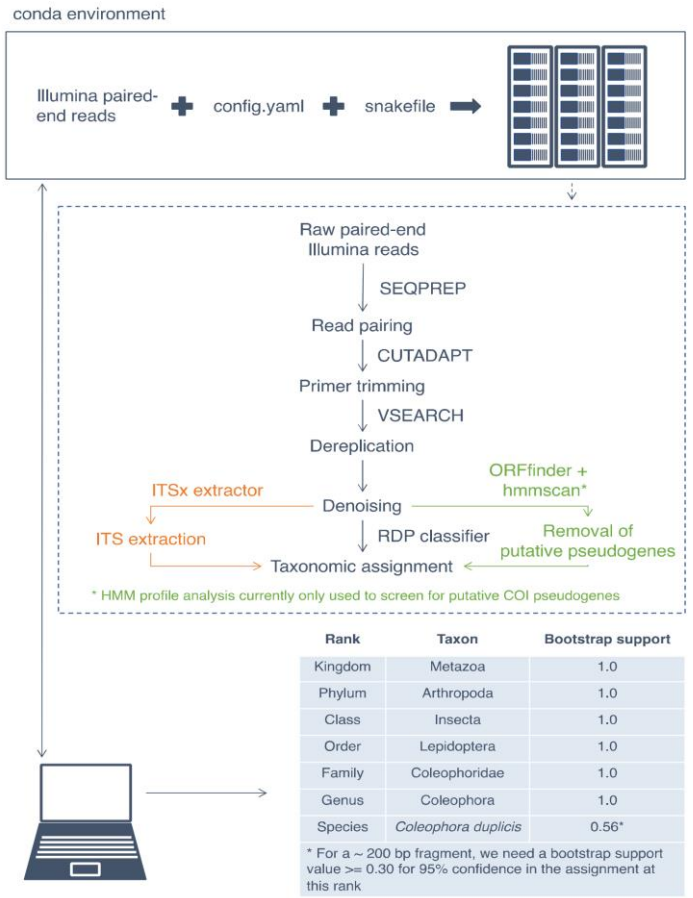
Conda(Snakemake)

Paired-end short reads

Generates only ESV (ASV)

ITSx

Remove pseudogenes (Hidden Markov model)



Rank	Taxon	Bootstrap support
Kingdom	Metazoa	1.0
Phylum	Arthropoda	1.0
Class	Insecta	1.0
Order	Lepidoptera	1.0
Family	Coleophoridae	1.0
Genus	Coleophora	1.0
Species	*Coleophora duplicis*	0.56*

* For a ~ 200 bp fragment, we need a bootstrap support value >= 0.30 for 95% confidence in the assignment at this rank

Porter et al. 2020

OBITOOLS

Linux, macOS

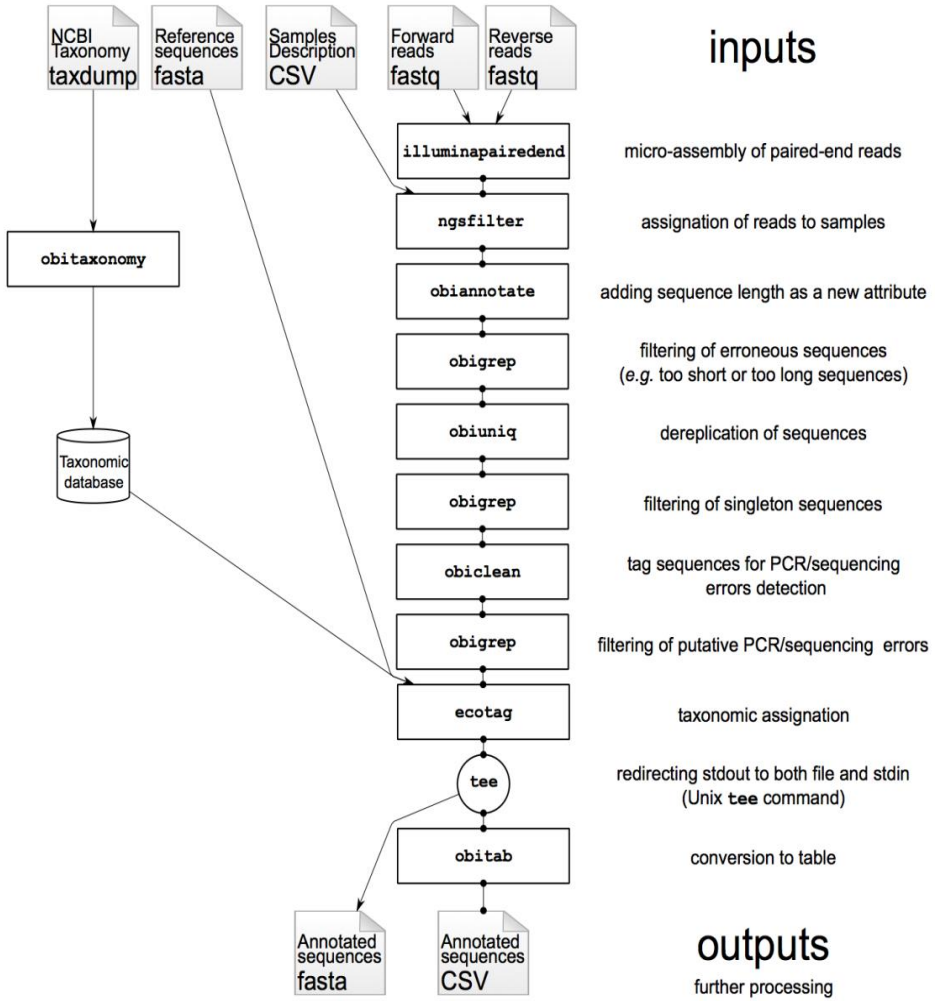
Paired-end short reads

Generates only OTU tables

Versatile software package

More used in fish and animal diet studies

Some of the modules are slow



Boyer et al. 2016 Molecular ecology resources

OBITOOLS

Linux, macOS

Paired-end short reads

Generates only OTU tables

Versatile software package

More used in fish and animal diet studies

Some of the modules are slow

Metabarcode design and quality assessment

ECOTAXSPECIFICITY	Evaluates barcode resolution

File format conversions

OBICONVERT	Converts sequence files to different output formats
OBIPR2	Converts PR2 database into an ECOPCR database
OBISILVA	Converts SILVA database into an ECOPCR database
OBITAB	Converts a sequence file to a tabular file

Sequence annotations

ECOTAG	Assigns sequences to taxa
OBIANNOTATE	Adds/edits sequence record annotations
OBIADDTAXIDS	Adds taxids to sequence records using an ECOPCR database

Computations on sequences

ILLUMINAPAIREDEND	Aligns paired-end Illumina reads
NGSFILTER	Assigns PCR product sequence records to their experiments/samples based on DNA tags and primers
OBICOMPLEMENT	Produces reverse complement sequences
OBICLEAN	Tags a set of sequences for PCR/ sequencing errors identification
OBICUT	Trims sequences
OBIJOINPAIREDEND	Joins paired-end reads
OBIUNIQ	Groups and dereplicates sequences

Sequence sampling and filtering

OBIEXTRACT	Extract samples from a data set
OBIGREP	Filters sequence file
OBIHEAD	Extracts the first sequence records
OBISAMPLE	Randomly resamples sequence records
OBISELECT	Selects representative sequence records
OBISPLIT	Splits a sequence file in a set of subfiles
OBISELECT	Selects representative sequence records
OBITAIL	Extracts the last sequence records

Statistics over sequence file

ECODBTAXSTAT	Gives taxonomic rank frequency of a given ECOPCR database
OBICOUNT	Counts the number of sequence records
OBISTAT	Computes basic statistics for attribute values

Utilities

OLIGOTAG	Designs a set of oligonucleotides with specified properties
OBISORT	Sorts sequence records according to the value of a given attribute
OBITAXONOMY	Manages taxonomic databases

Boyer et al. 2016 Molecular ecology resources

Marker Specific Pipelines

Nf-core / Ampliseq

Linux, macOS, Windows (Nextflow)

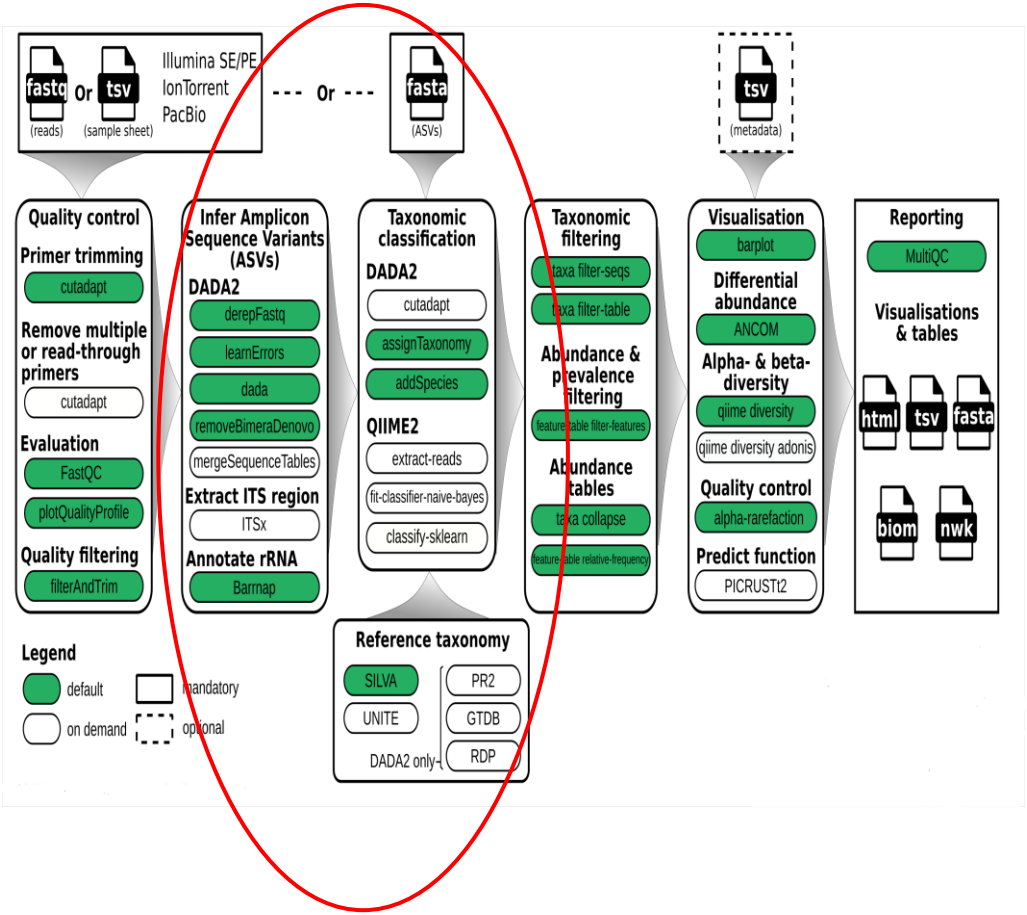
Single-end and Paired-end short reads,
Single end Long reads

Generates only ASV

16S, 18S, ITS (With ITSx)

QIIME2 + DADA2

Appropriate for 16S studies



Straub, et al. 2020 *Frontiers in Microbiology*

BIOCOM-PIPE

Linux, macOS

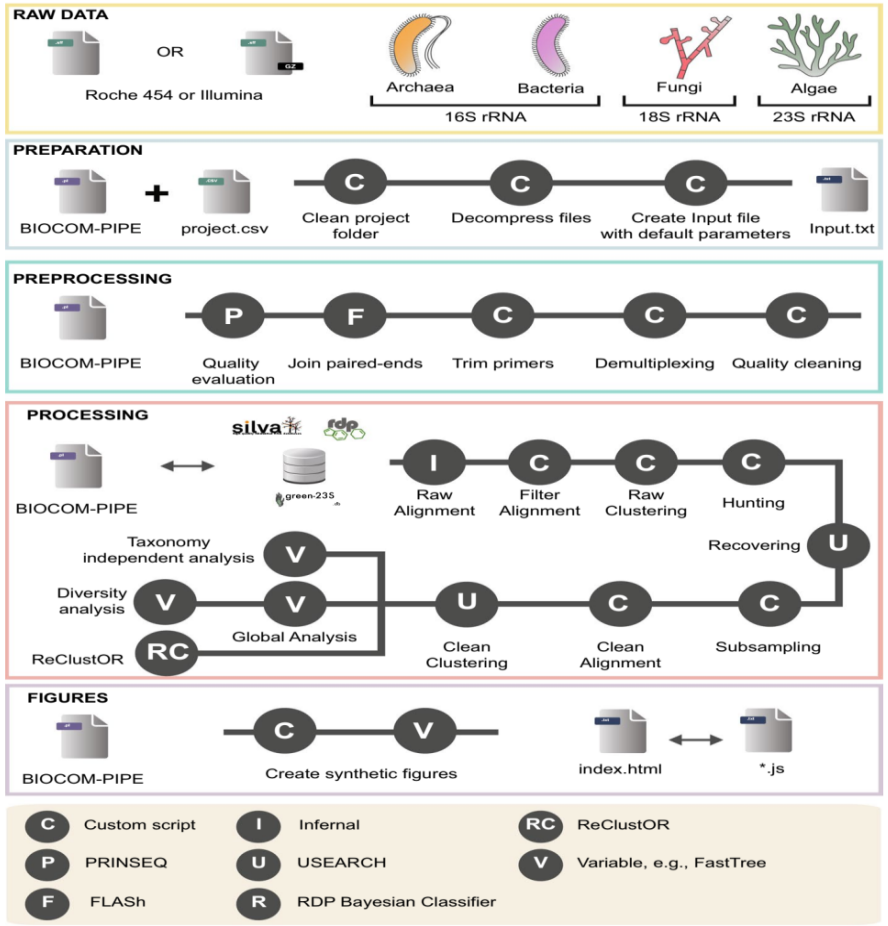
Paired-end short reads

Generates only OTU

16S, 18S, 23S

The “hunting–recovering” process

A complete classification of all high-quality reads, not just one representative read for each OTU



Djemiel et al. 2020 BMC bioinformatics

DAnIEL

Web-based, Linux

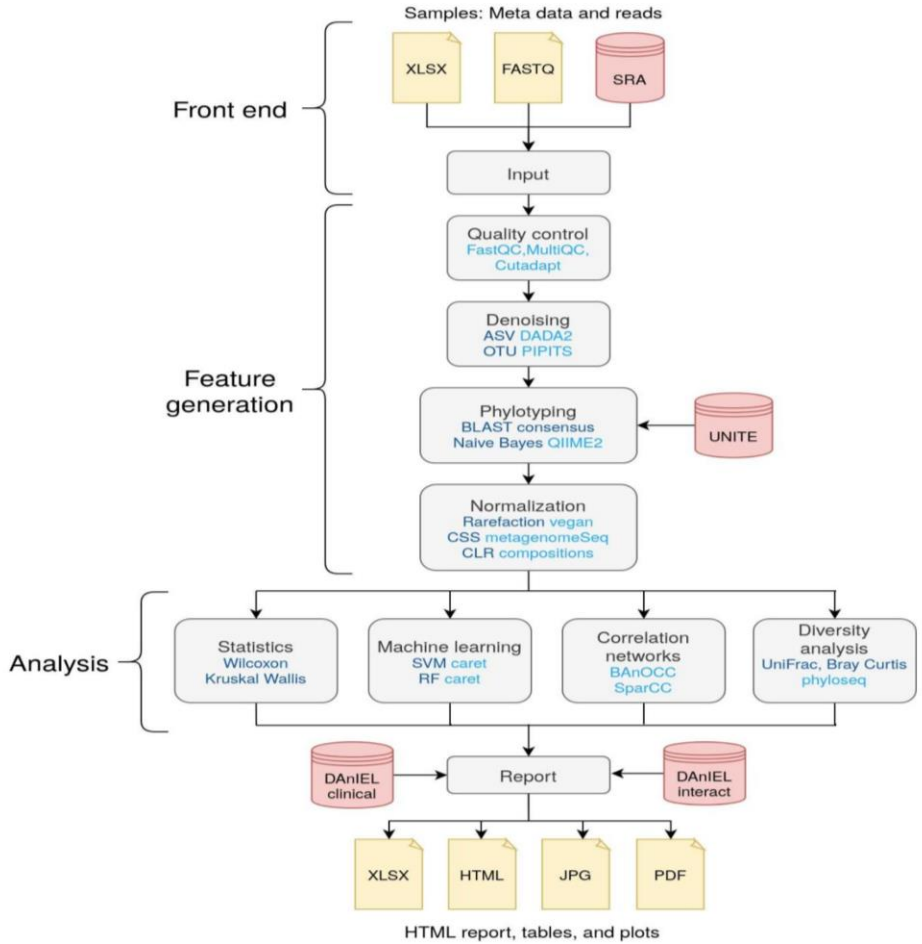
Paired-end short reads

Generates Both OTU and ASV files

Only ITS

Supports SRA

Vegan and Phyloseq implemented



Loos et al. 2021 Frontiers in Microbiology

EzMAP

Linux, macOS, Windows (Docker images), GUI

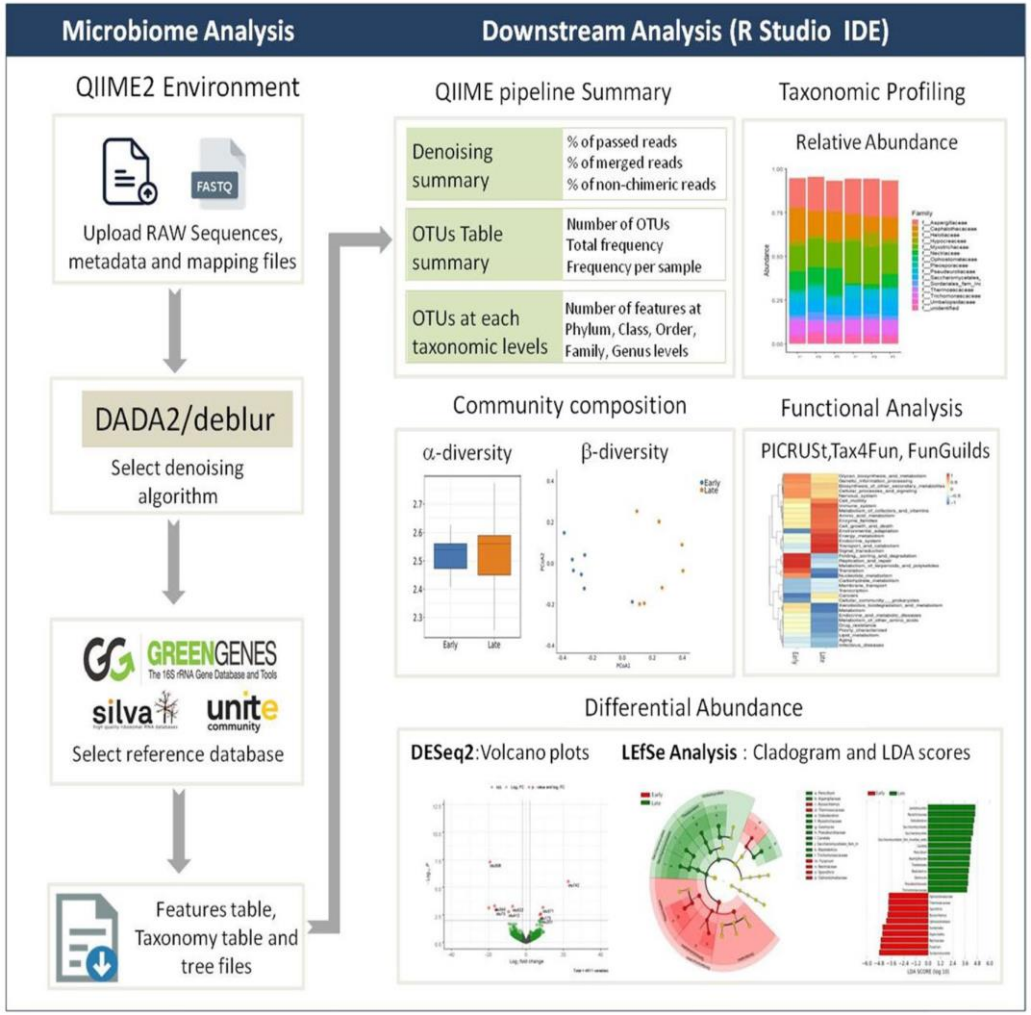
Paired-end or single-end short reads

Generates only ASV files

16S,ITS (without ITSx)

Wraps QIIME2 and DADA2 functions

Functional and Differential Abundance analysis



Shanmugam, et al. 2021 *BMC bioinformatics*

gDAT

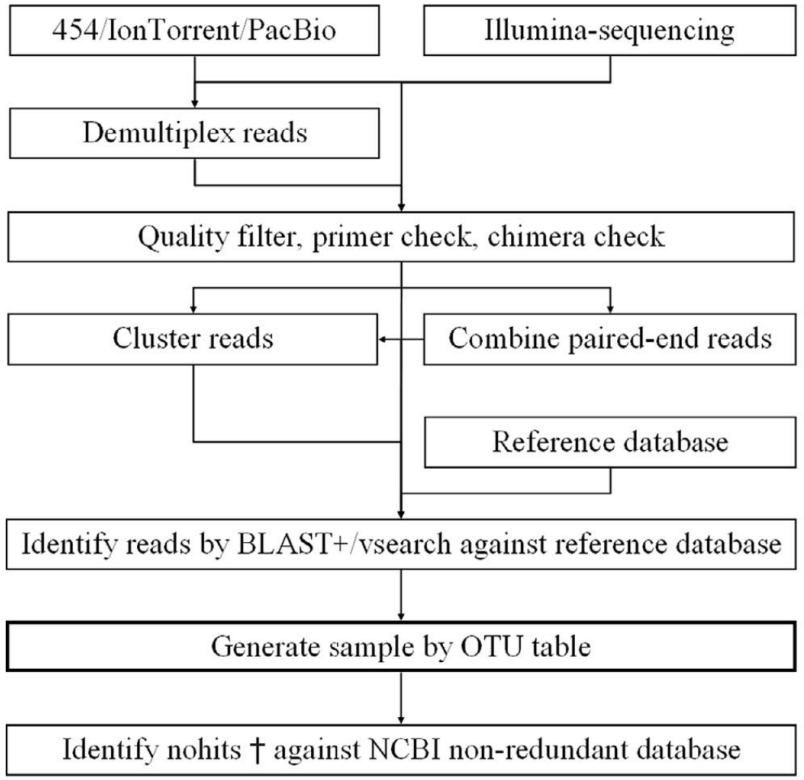
Linux, macOS, Windows, GUI

Single-end long reads, Paired-end short reads

Generates only OTU

16S, ITS (without ITSx)

Designed to work on a commodity computer with limited processing and memory capabilities



Vasar, et al. 2021 Molecular ecology Resources

JAMP

Linux, macOS (R-based)

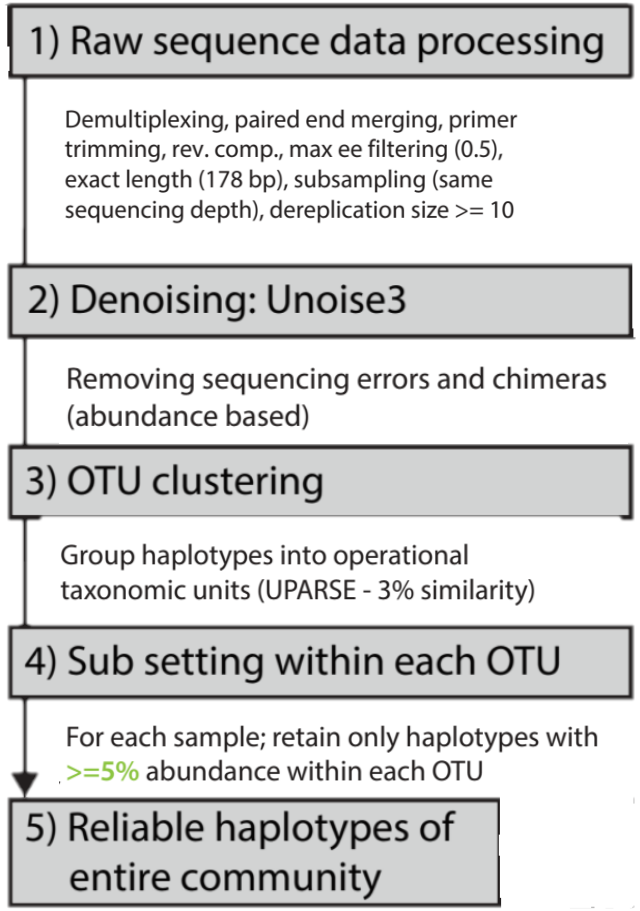
Paired-end short reads

Generates only OTU

COI

Cutadapt, Usearch and Vsearch

Haplotype Detection (single species)



Elbrecht et al. 2018 PeerJ

LOTUS2

Linux, macOS

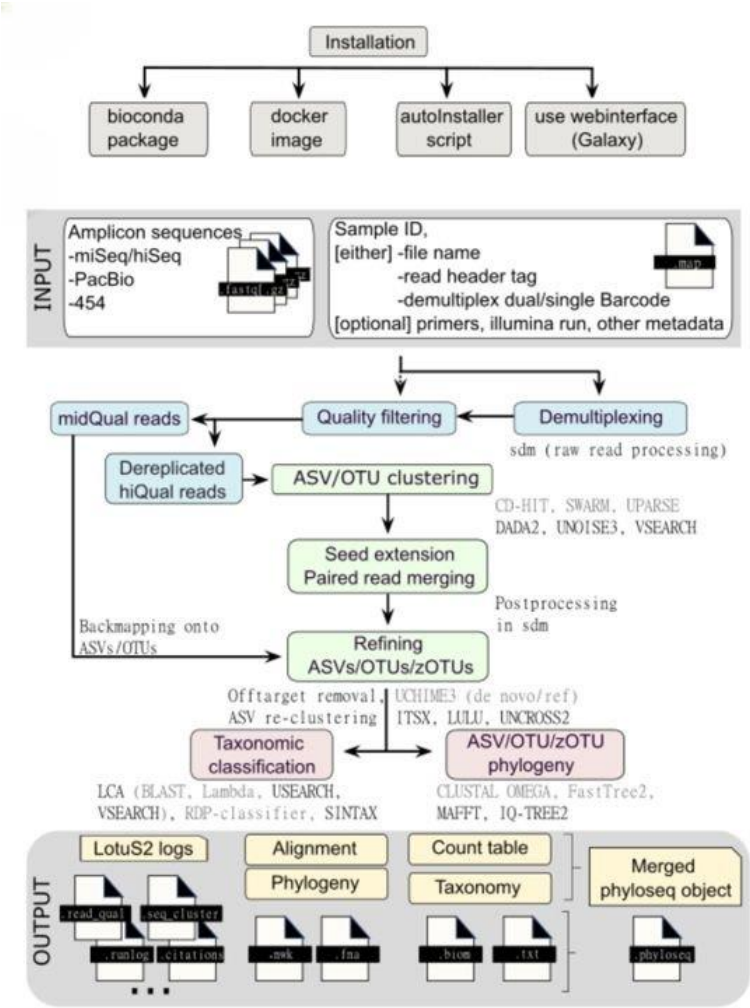
Single-end long reads, Paired-end short reads

Generates Both OTU and ASV files

16S, 18S, 28S, ITS (ITSx)

UNCROSS2 for reducing cross-talk error

Fast and low ram usage



Özkurt et al. 2021

MICCA

Linux, macOS , Windows (Docker images)

Single-end long reads, Paired-end short reads

Generates only OTU

16S, 18S, 28S, ITS

De novo built-in clustering method (OTUCLUST)

Command	Description	Tools	Notes
micca-preproc	• primer trimming both in the 5′ and 3′ ends of reads using semi-global alignments • quality trimming using sliding windows • minimum length filtering	• Cutadapt • SICKLE	supports gapped alignment and IUPAC codes for primer trimming
micca-otu-denovo	• de novo sequence clustering • de novo chimera filtering • taxonomic assignment with RDP classifier or BLAST+	• OTUCLUST • UCHIME • RDP Classifier • BLAST+	BLAST+: Greengenes, Silva and UNITE QIIME-formatted databases are supported. RDP: versions 2.6+ are supported.
micca-otu-ref	• reference-based clustering	• DNACLUST	Greengenes, Silva and UNITE QIIME-formatted databases are supported
micca-phylogeny	• de novo and template-based multiple sequence alignment (MSA) • phylogenetic tree reconstruction	• MUSCLE • T-Coffee • PyNAST • FastTree	

Albanese et al. 2015 Scientific Reports

PEMA

Linux, macOS, Windows (Docker images)

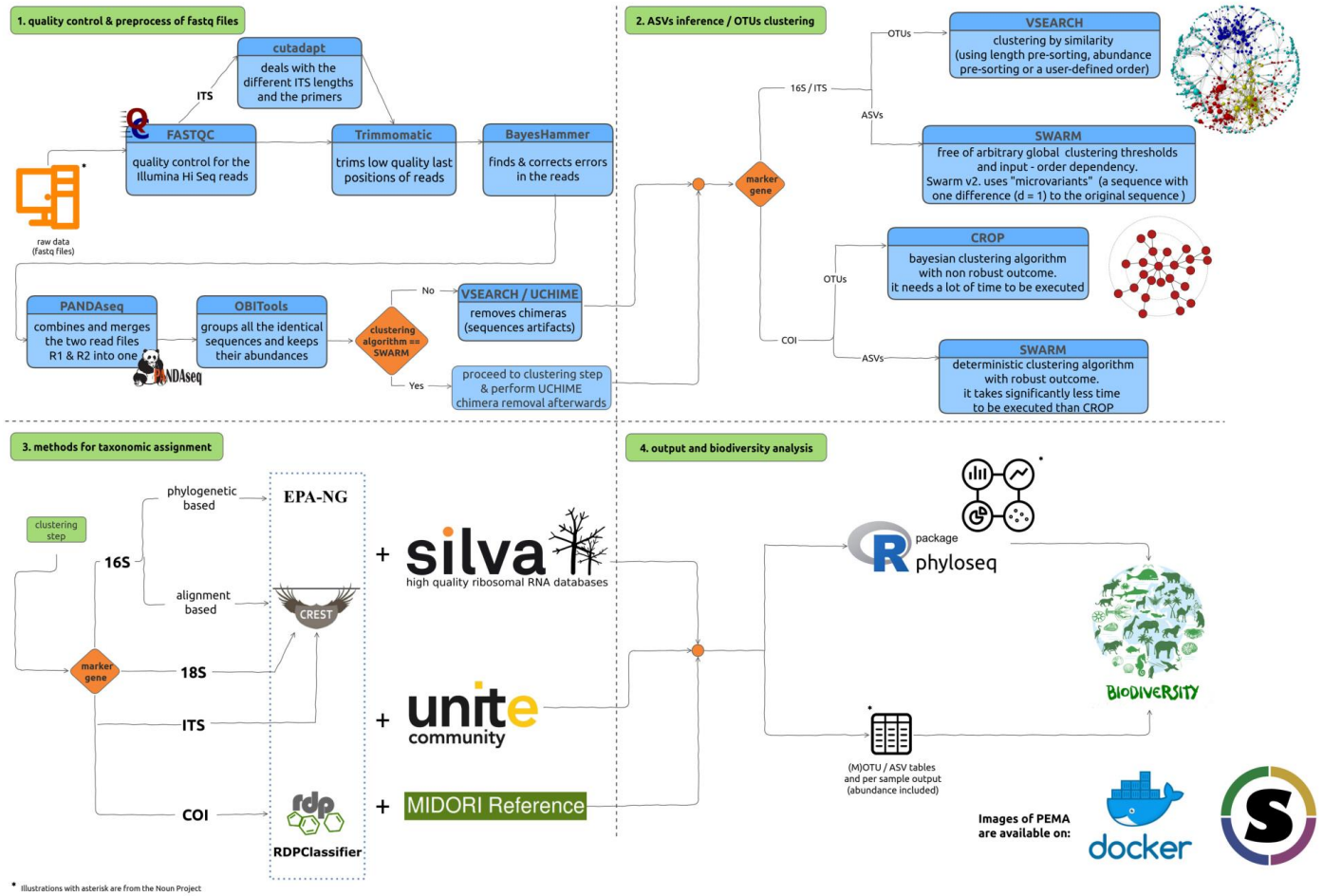
Paired-end short reads

Generates Both OTU and ASV files

Fixed reference database (16S,18S,ITS,COI)

Alternate programs for each Step

Phylogenetic based taxonomy assignment



Zafeiropoulos et al. 2020 GigaScience

PIPITS

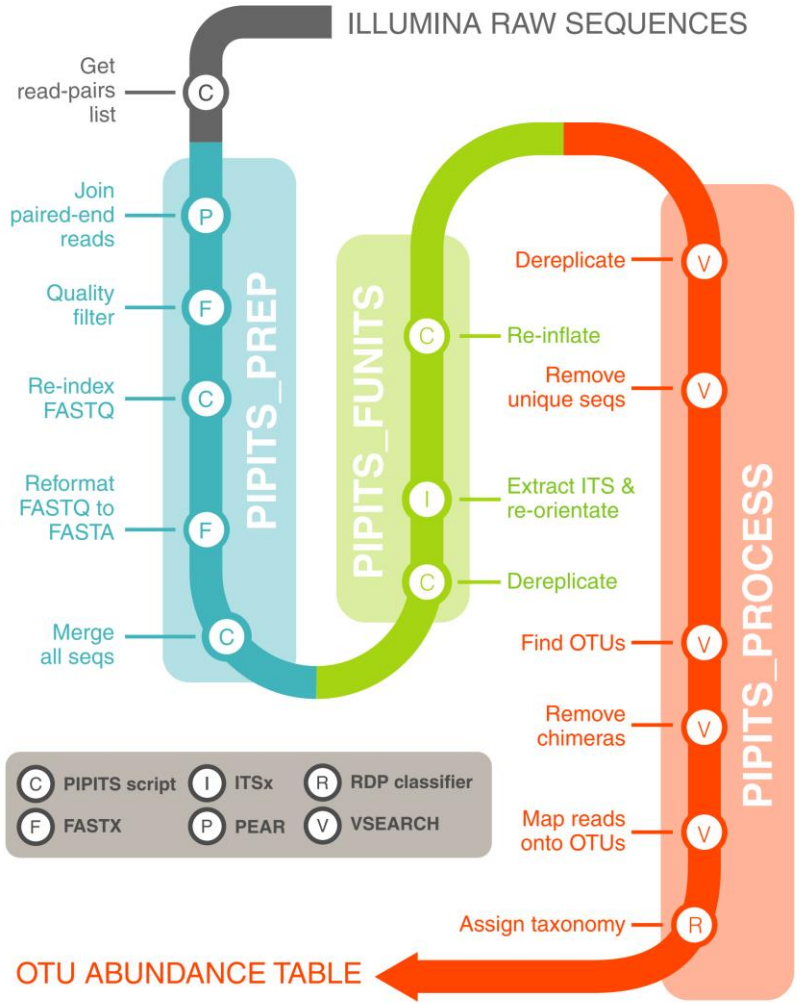
Linux

Paired-end short reads

Generates only OTU

ITS (ITSx)

RDP classifier + UNITE fungal database



Gweon et al. 2015 Methods in ecology and evolution

QIIME2

Linux, macOS, Windows (Docker images)

Paired-end short reads

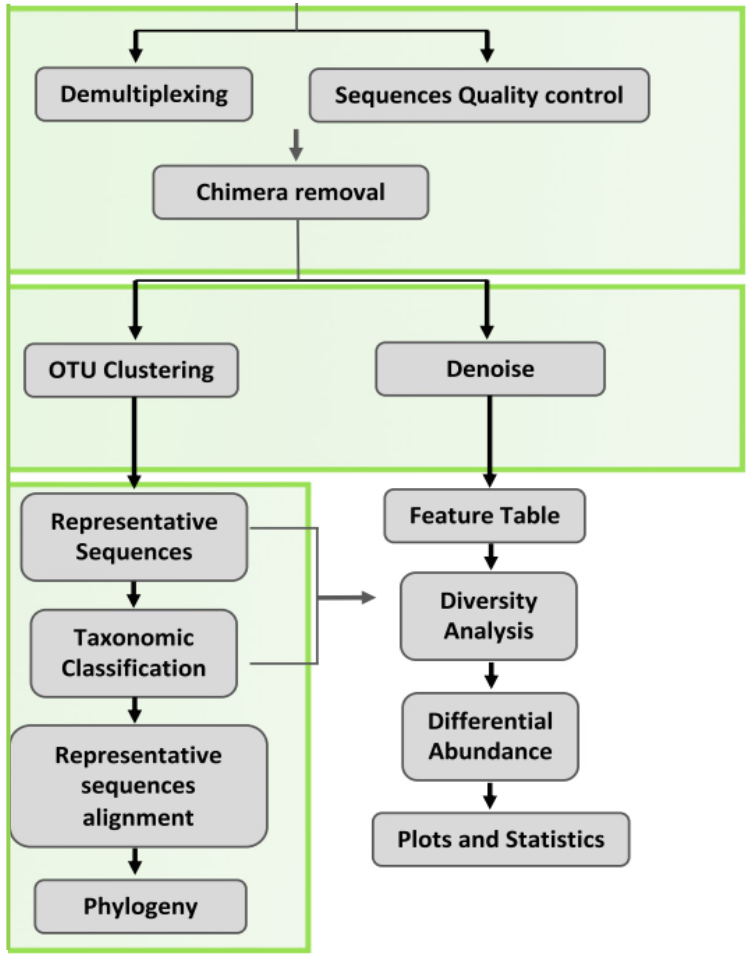
Generates Both OTU and ASV

16S,18S,ITS (without ITSx)

Anybody can create and distribute a plugin (https://docs.qiime2.org/2022.2/plugins/available/)

2 denoising software built-in: DADA2, Deblur

Mostly used For Microbiome Studies(16S)



Boylen et al. 2019 Nature Biotechnology

SEED2

Linux, macOS, Windows (GUI)

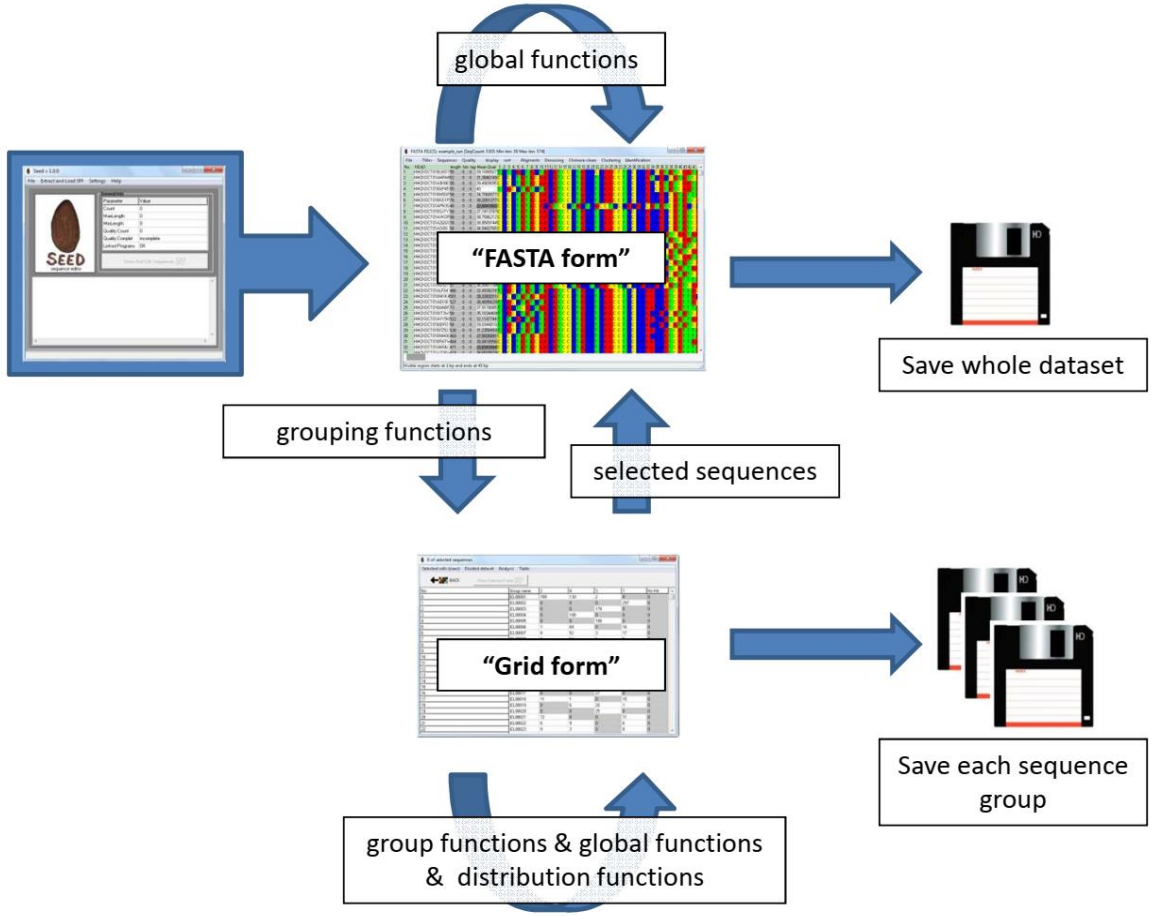
Paired-end short reads

Generates Only OTU

16S, ITS (with ITSx)

Sequence visualization capabilities

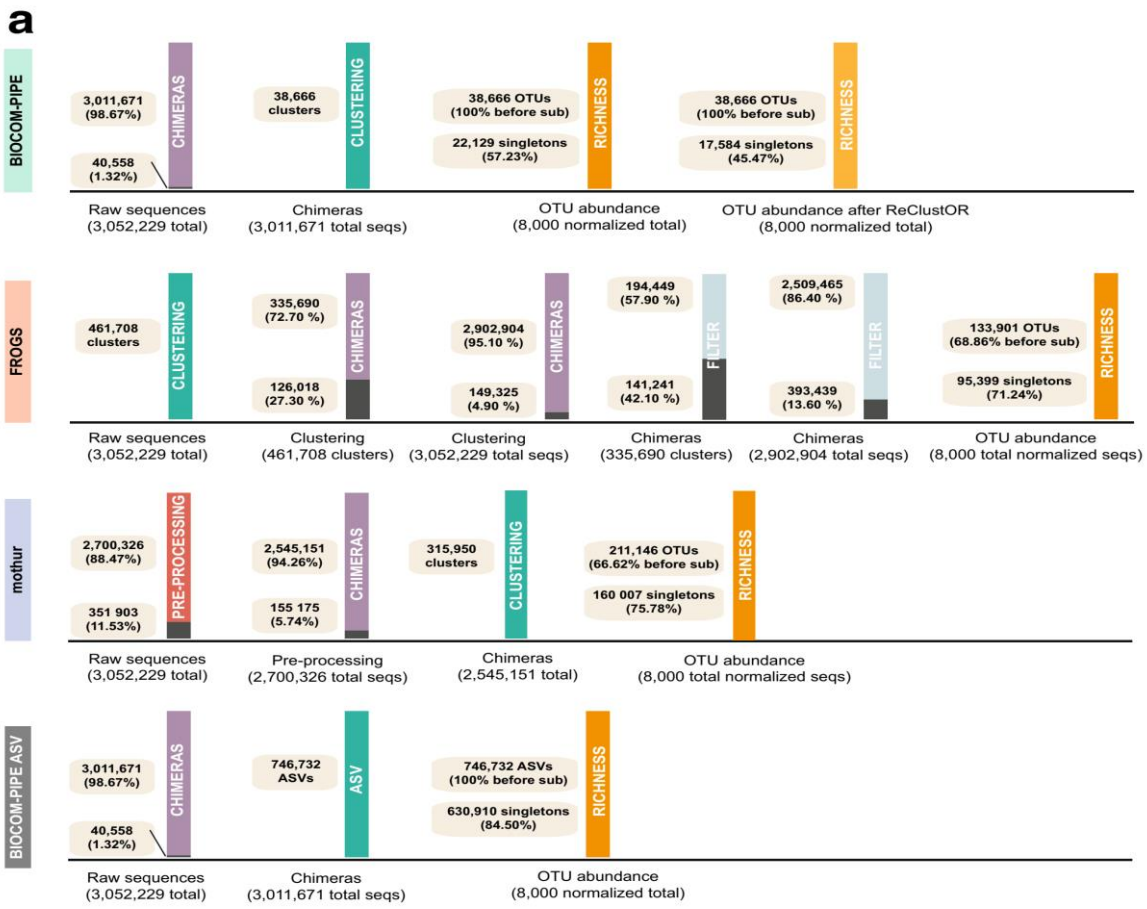
Generate phylogenetic trees

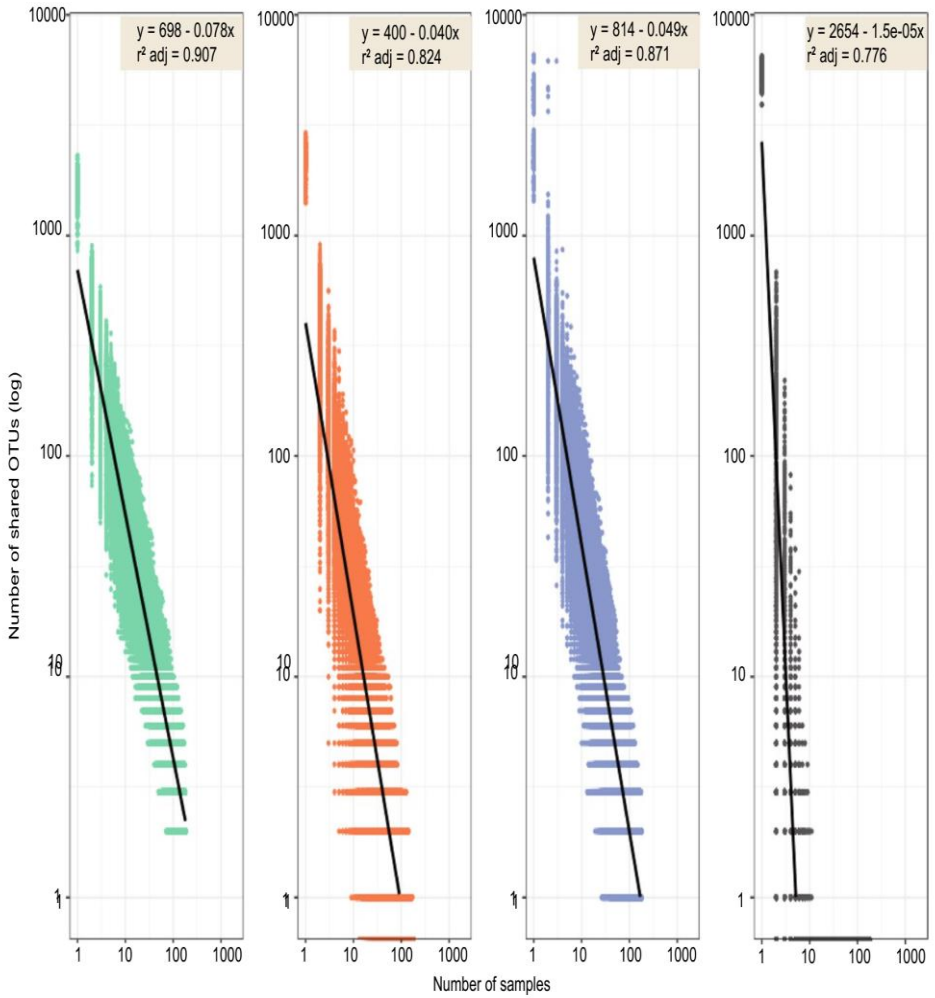


Vetrovsky et al. 2018 Bioinformatics

Benchmarking Studies

16S

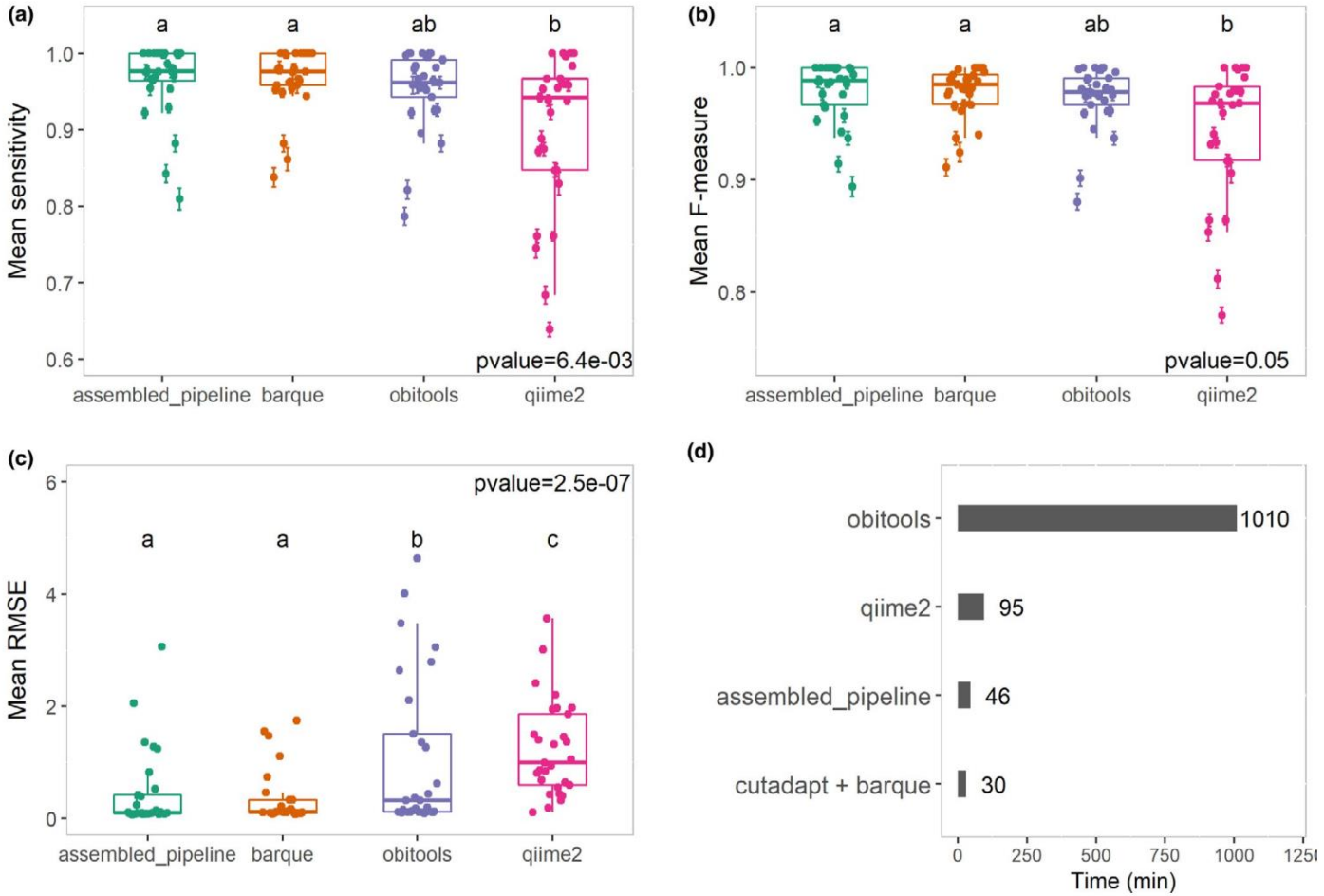




Djemiel *et al. 2020 BMC Bioinformatics*

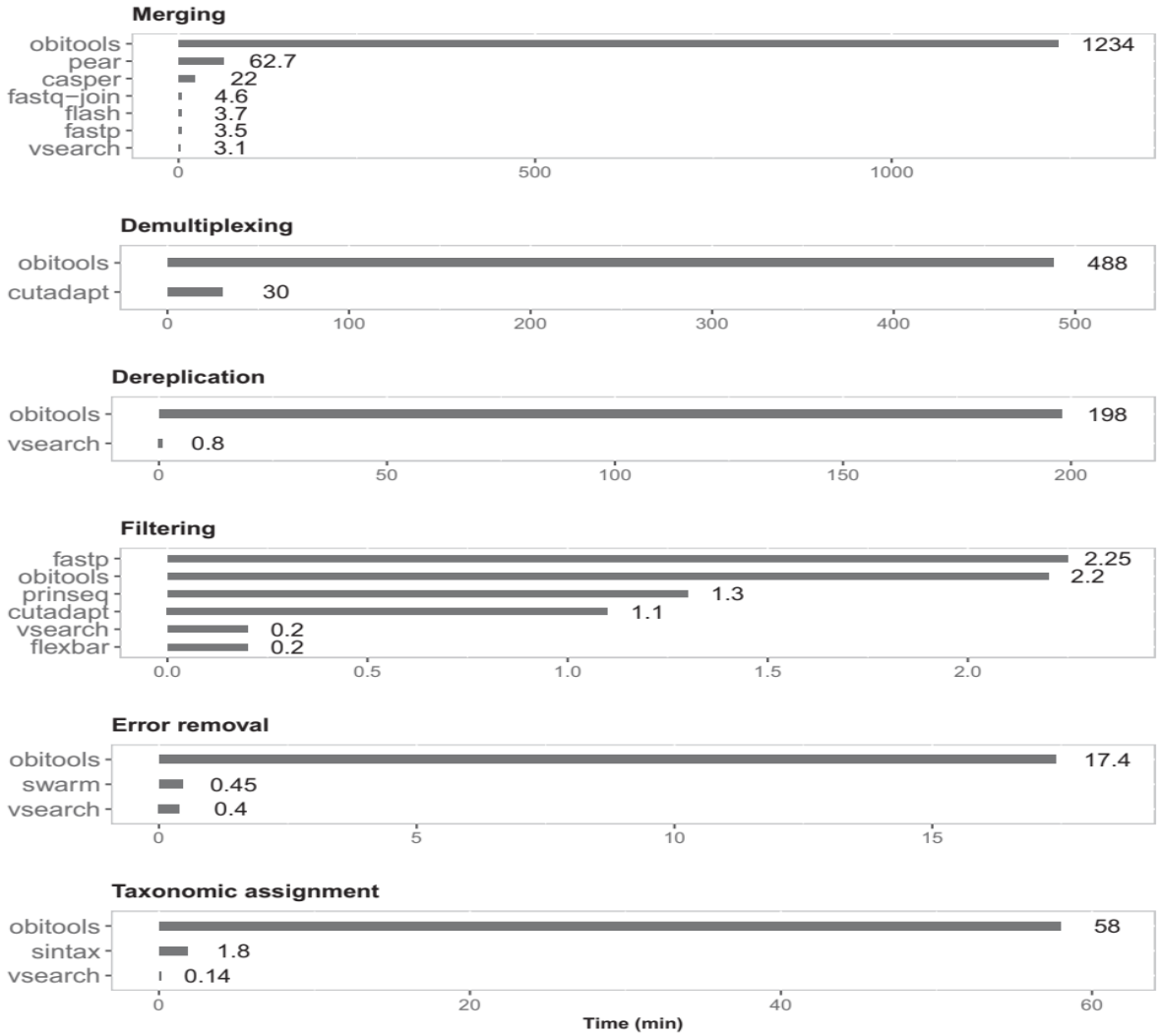
Barque-OBITOOLS-QIIME2

Fish samples (12S)

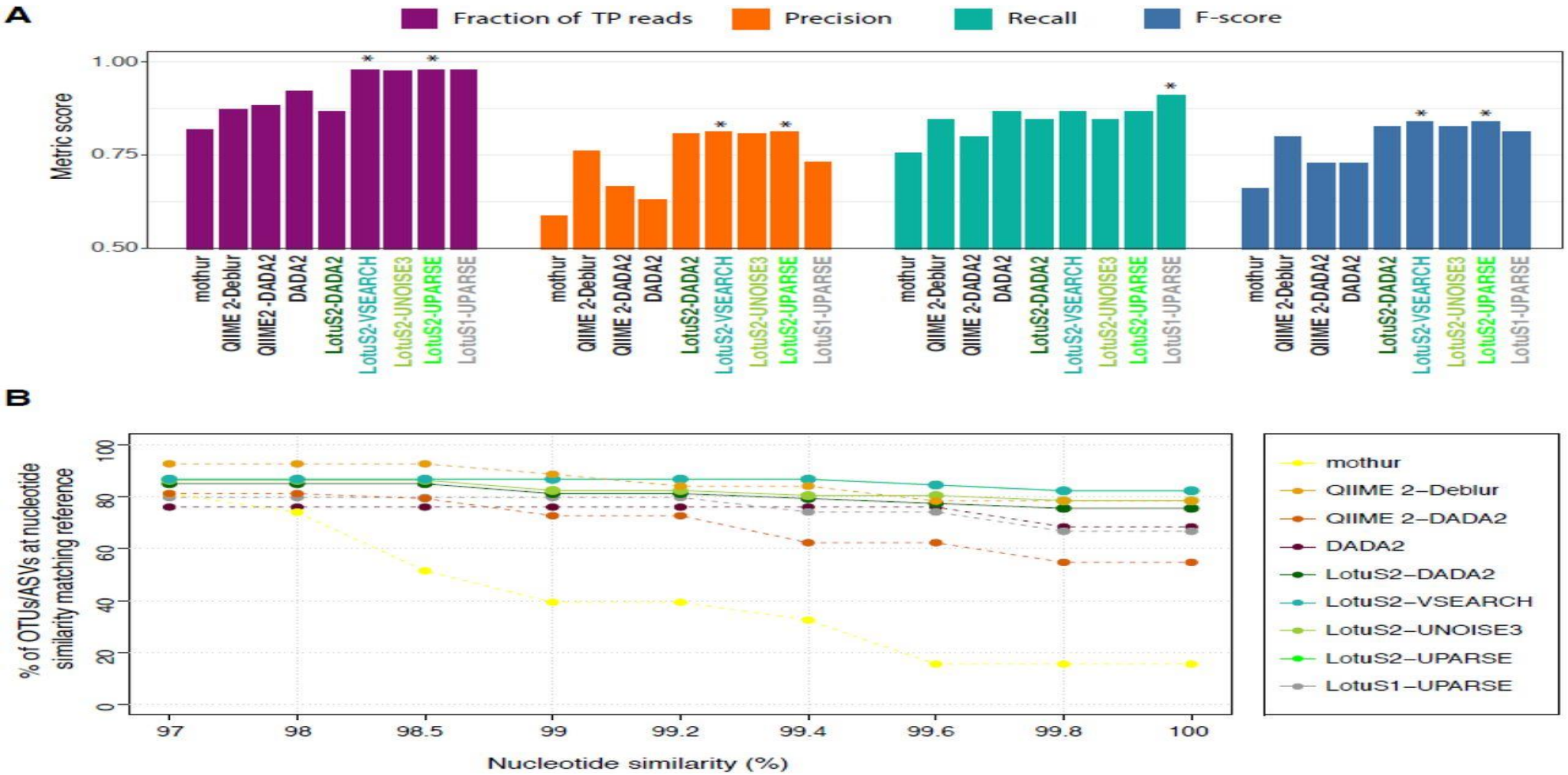


Mathon et al. 2021 MOLECULAR ECOLOGY RESOURCES

Barque-OBITOOLS-QIIME2



16S-ITS



Özkurt et al. 2021

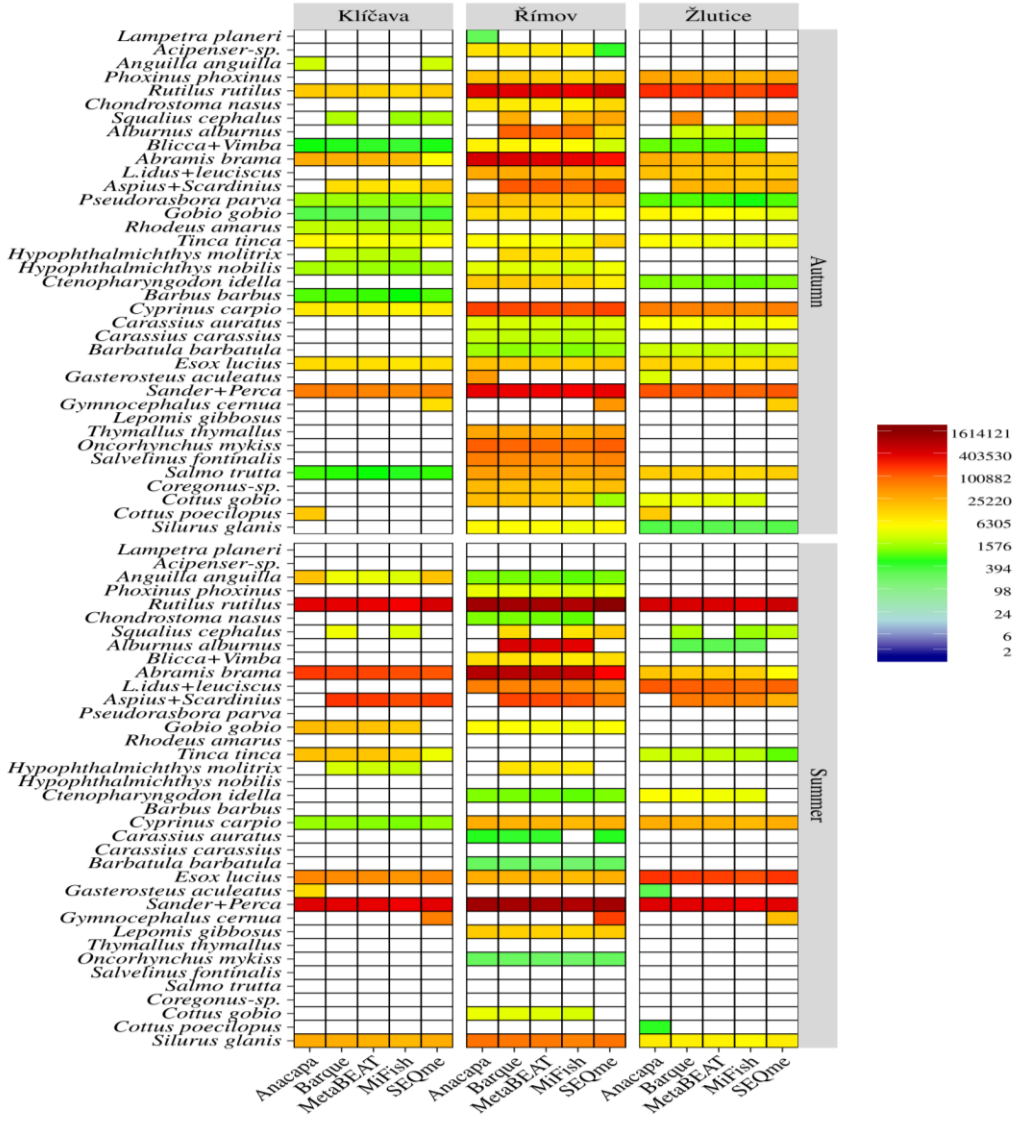
ANACAPA- BARQUE

The three reservoirs (Klícava, ˇ Rímov, and Žlutice)

Data Processing Steps	Anacapa	Barque
Total from original data	22,464,147	22,464,147
Total after demultiplexing	20,910,517	20,910,517
Trimmed and filtered	19,095,153	18,632,248
Merged	18,944,446	18,619,523
Filtered and chimera removed	18,271,442	17,889,794
Assigned	10,676,765	11,112,721
Unassigned (original data)	11,787,382	11,351,426
Unassigned (demultiplexed data)	10,233,752	9,797,796

Number of sequence reads assigned to pipelines, reservoirs, and seasons

Pipeline	Number of reads
Anacapa	7,816,625
Barque	8,410,037
MetaBEAT	7,859,744
MiFish	6,820,393
SeqME	8,200,342



Thank you