**Work package 4.** Data Infrastructure and **bioinformatics pipeline**

# Description (Project proposal)

This work package addresses a number of gaps in existing workflows in order to meet the data management needs specified in WP2, and more generally in order to improve the availability of omics data, and more specifically metabarcoding data from biofouling communities, to global data repositories and hence to decision support systems**. By connecting and building on a number of existing initiatives**, we will create a workflow for genetic data from raw data generated in the lab or even in the field to information products that are ready for use by a policy decision support system (Figure 3).

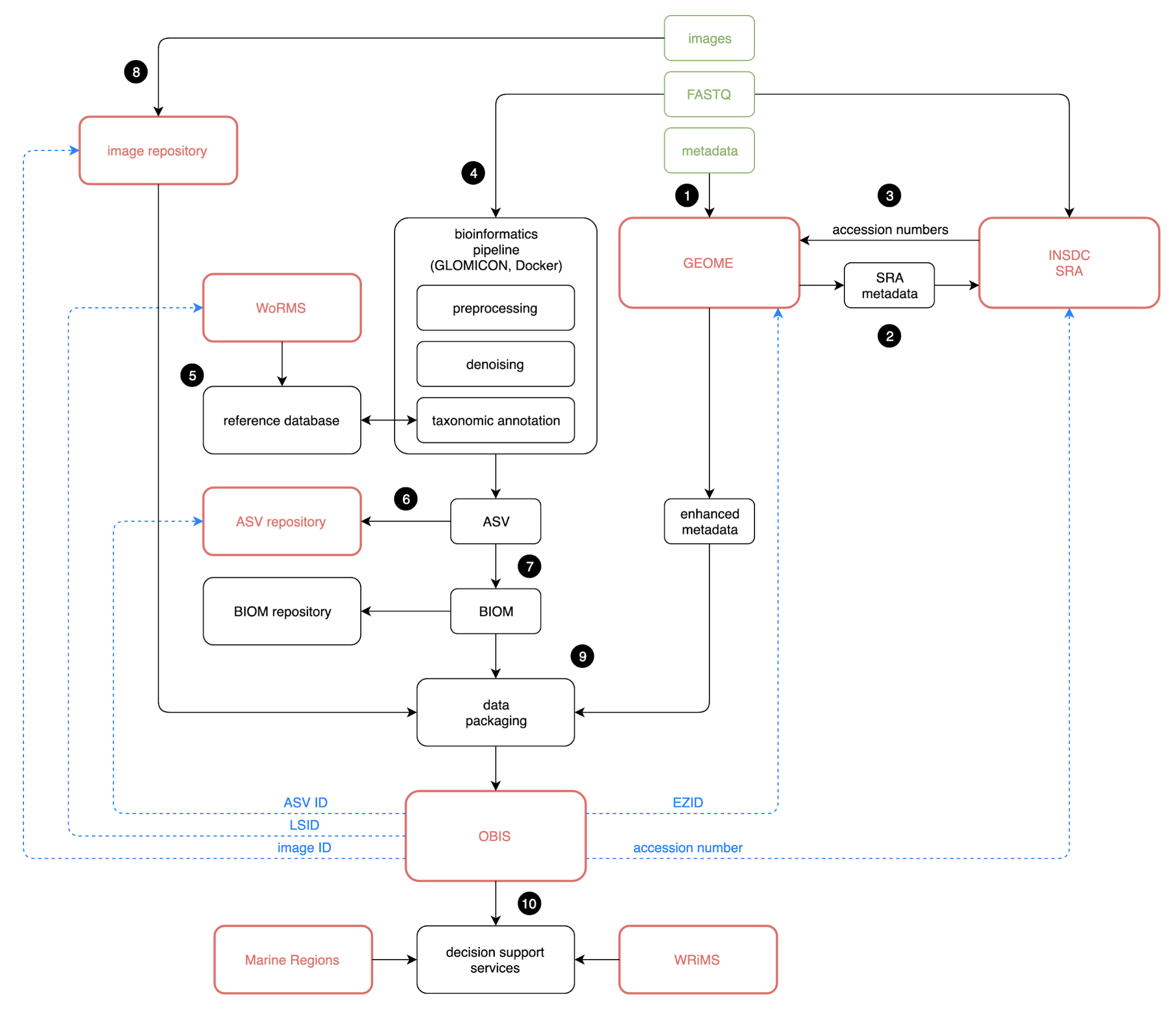


Figure 3. Data and metadata processing workflow from raw sequence reads to standardized OBIS occurrence data.

In this document we will go over existing bioinformatics pipelines and give recommendations for the pipeline that OBIS would use.

Many different bioinformatics pipelines exist for processing raw reads from next-generation sequencing (NGS) platforms. These pipelines are often composed of existing open-source software tools and commonly involve the following high-level steps: **preprocessing (read pairing, demultiplexing, dereplication), denoising and/or clustering, and taxonomic annotation.** traditionally, barcode sequences are clustered into Operational Taxonomic Units (OTUs) in order to get rid of sequence variation introduced by amplification and sequencing errors, which cannot easily be distinguished from actual intraspecific genetic variation. More recently, algorithms have been developed to infer exact sequences down to the level of single-nucleotide differences without imposing arbitrary similarity thresholds. These amplicon sequence variants (ASVs) have several advantages over the classic OTU approach, including higher taxonomic resolution and interoperability between studies13. Notable bioinformatics software suites and pipelines include PEMA (used by ARMS), QZIP (used by AWI), Tourmaline (used by NOAA), and Banzai (used by MBON). Recently, the Global Omics Observatory Network (GLOMICON; www.glomicon.org) has started an initiative to document and align existing bioinformatics workflows that generate amplicon sequence variants (ASVs). In particular, this initiative aims to establish an ASV registry where ASVs are assigned a persistent universally unique identifier, and to come up with a common data and metadata exchange format which can then be used to feed data into global aggregators such as OBIS. We will work with GLOMICON to come up with an optimal bioinformatics pipeline, implemented as shared open source software (Figure 3, 4).

# Review of existing pipelines

The approaches developed by the pipelines in table 1 are reviewed. These pipelines were decided on by a focus on ASV -detection, environmental DNA analyses, or their workflow completeness.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Institution/Author** | **Publication** | **Code** |
| **abyss** | Ifremer | <https://doi.org/10.1101/717355> | <https://gitlab.ifremer.fr/abyss-project> |
| **Anacapa** | UCLA | <https://doi.org/10.1111/2041-210X.13214> | <https://github.com/limey-bean/Anacapa/> |
| **Banzai** | MBON |  | <https://github.com/jimmyodonnell/banzai> |
| **CASCABEL** | NIOZ | <https://doi.org/10.3389/fgene.2020.489357> | <https://github.com/AlejandroAb/CASCABEL> |
| **MetabarTOAD** | Luke Holman |  | <https://github.com/leholman/metabarTOAD> |
| **MgNify** | ENA | <https://www.ebi.ac.uk/metagenomics/pipelines/5> |  |
| **PEMA** | ARMS | <https://doi.org/10.1093/gigascience/giaa150> | <https://github.com/hariszaf/pema> |
| **QZIP** | AWI |  | <https://github.com/PyoneerO/qzip> |
| **SLIM** | University of Geneva | <https://doi.org/10.1186/s12859-019-2663-2> | <https://github.com/yoann-dufresne/SLIM> |
| **Tourmaline** | NOAA |  | <https://github.com/aomlomics/tourmaline> |

**ABYSS**

The Abyss pipeline combines Amplicon Sequence Variants (ASV) assignment (dada2) with MOTU clustering (swarm) and additional post-clustering curation (LULU). The approach of clustering is used especially to enable more realistic biodiversity estimates as species differ greatly in their intraspecific variation. It could be tested if clustering and curating of the datasets could be helpful also for invasive species detections, however this may be more relevant for other types of analyses.

**Currently the steps are:**

1. A. Ligation ligation preprocessing : classify and clean R1F/R2R R2F/R1R reads with cutadapt then sort reads with bbmap repair
2. B. Double PCR data preprocessing :
   * Cutadapt : remove adapters/primers for R1/R2 reads
   * Cleaning : keep only valid couples for R1/R2 reads with bbmap repair
3. Dada2 : read correction and production of Amplicon Sequence Variants (ASVs), Taxonomy assignment with RDP Classifier
4. Blast : Blast taxonomy assignation for Amplicon Sequence Variants (ASVs), merged to RDP taxonomy table
5. (optional) FROGS : clustering with swarm v2 / additional chimera removal / affiliation of Operational Taxonomic Units (OTUs) with RDP and BLAST (Escudie et al. 2017)
6. Tag switching correction : sort the samples by abundance of each ASV/OTU and eliminate the reads of the samples corresponding to a cumulative frequency of less than 3% for each particular ASV/OTU (Wangensteen & Turon 2016, Metabarpark project)
7. Lulu : ASV/OTU table curation based on similarity and co-occurence rates (Froslev et al. 2017)

**ANACAPA**

The Anacapa pipeline is largely based on the dada2 pipeline, which is gaining popularity. It uses sequence quality information for denoising data and inferring ASV tables. Anacapa combines this with some initial trimming, taxonomic inference as well as building a custom reference database.

Anacapa consists of 4 modules:

1. Building reference libraries using [**CRUX**](https://github.com/limey-bean/CRUX_Creating-Reference-libraries-Using-eXisting-tools)
2. Running quality control (QC) and assigning Amplicon Sequence Variants (ASV) using Dada2 (Sequence QC and ASV Parsing),
3. Assigning taxonomy using Bowtie 2 and a Bowtie 2 specific Bayesian Least Common Ancestor (BLCA) (Assignment) and
4. Running exploratory data analysis and generating ecological diversity summary statistics for a set of samples using [ranacapa](https://f1000research.com/articles/7-1734/v1).

All scripts and several prebuilt reference libraries are available and can be utilized on a local computer, and HPC server or a Singularity container. Anacapa is developed for the analysis of datasets of the [CALeDNA](https://ucedna.com/) project, making it suitable for the needs of PacMAN. The use of automated reference library generation, could be a useful tool for OBIS, updating taxonomic assignments, and allowing several different markers to be analysed in an otherwise identical manner. Currently available crux datasets are built solely on Genbank sequences, but it should be possible to add other datasets as well, like BOLD or MIDORI, for the COI gene.

**BANZAI**

**banzai** is a shell script (bash) that links together the disparate programs needed to process the raw results from an Illumina sequencing run of PCR amplicons into a contingency table of the number of similar sequences found in each of a set of samples. Banzai is the MBON bioinformatics workflow,

1. **Read merging**: pear
2. **Quality filtering**: vsearch
3. **Demultiplexing**: awk
4. **Primer removal**: cutadapt
5. **Dereplication**: python
6. **Chimera detection**: vsearch
7. **OTU clustering**: swarm
8. **Taxonomic annotation:** blastn
9. **Taxon resolution**: R:taxize

**CASCABEL**

Cascabel is a pipeline built with snakemake and developed at NIOZ, though mainly for prokaryotic community analyses. Due to snakemake, it is flexible and easy to run with the modules of choice and with different configurations. Currently Cascabel supports both analyses based on clustering, or analyses based on ASV inference with dada2, and relies also largely on qiime algorithms. Cascabel produces a report of the analysis run, and can also warn users of any problems in the sequence data during the run in interactive mode. Currently steps for the ASV workflow are:

1. Quality control: FASTQC
2. Assemble fragments (merge forward with reverse reads) (Shouldn’t be done for dada2)
3. FastQC on merged/assembled fragments
4. Demultiplex input files
   1. Extract barcodes
   2. Barcode correction
   3. Split libraries
5. dada2 trim and filter reads
6. dada2 merge pairs
7. dada2 generates ASV
8. Assign taxonomy: dada2 uses RDPClassifier
9. Length filtering for OTU or ASV
10. Filter OTUs / ASV from table (remove singletons)
11. Biom convert
12. +Summary, alignment, phylogeny steps.

**METABARTOAD**

This is a simple pipeline for analysing metabarcoding data. It strings together functions provided in several packages and has been extensively tested on marine eDNA data generated by the Illumina MiSeq. It supports either analysis with UPARSE/UNOISE for OTU inference or analysis with dada2 for ASV inference. Doesn’t contain steps for taxonomic assignments. The dada2 pipeline is preceded by a custom dadaReadPrep() function, to link sequences coming from the sequencer to the denoising pipeline.

**MGNIFY**

MGnify is a freely available hub for the analysis and exploration of [metagenomic](https://emg-docs.readthedocs.io/en/latest/glossary.html#term-Metagenomic), [metatranscriptomic](https://emg-docs.readthedocs.io/en/latest/glossary.html#term-Metatranscriptomic), [amplicon](https://emg-docs.readthedocs.io/en/latest/glossary.html#term-Amplicon) and [assembly](https://emg-docs.readthedocs.io/en/latest/glossary.html#term-Assembly) data. The resource provides rich functional and taxonomic analyses of user-submitted sequences, as well as analysis of publicly available metagenomic datasets held within the European Nucleotide Archive ([ENA](https://www.ebi.ac.uk/ena)). It is a promising platform for OBIS, as it is well-documented and standardized, has access to and analyzes large amounts of public sequencing data. However, currently it doesn’t support COI-analysis, and has a very different approach (different tools) from the other pipelines analyzed here. It might not therefore be best suited for environmental DNA analyses, or keeping data comparable with other sources (however a big source in itself.) The current amplicon analysis pipeline (v. 5) consist of the following steps:

1. **Merge reads:** SeqPrep
2. **Count reads, and trim:** FastQC+Trimmomatic
3. **QC filtering/Quality control:** Biopython
4. **Search agains Rfam library (rRNA prediction):** cmsearch
5. **Extract coordinates and sequences**
6. **Classify:** MapSeq
7. **Visualise:** Krona

**PEMA**

Pema is the pipeline developed and currently in use for the ARMS program. This makes it a well-standardized workflow, and used by a biodiversity community already. PEMA can be used for analysing 16S, 18S, ITS or COI metabarcoding data. The main steps are:

1. **Quality control and preprocessing of fastq files**
   1. FastQC, cutadapt, trimmomatic, PandaSEQ, OBITools, Vsearch-uchime
2. **ASV inference/OTU clustering**
   1. Vsearch, Swarm, Crop
3. **Taxonomic assignment**
   1. Algorithms: EPA-NGA, Crest, RDPClassifier
   2. Reference libraries: Silva, UNITE, MIDORI
4. **Output and biodiversity analysis**
   1. Phyloseq

The possible downsides of this pipeline are, that ASV inference is done through the older quality control methods, which are being replaces by the use of dada2 at the moment. It could be possible to discuss with the group on the addition of dada2 to the pipeline for ASV inference. PEMA has a nice documentation page: <http://hariszaf.github.io/pema_documentation>

**QZIP**

Qzip is a bash script which wraps and connects common bioinformatics tools to produce sample by observation contingency tables from demultiplexed Illumina paired-end amplicon sequences. The usage of GNU parallel for most parallelizable processing steps and the usage of Swarm as fast clustering method makes it extremely speedy. On a 20 core linux system (physical cores) the processing of about 300 million sequences from a large scale project only takes ~ 6 h and the processing of 10 million sequences from a handy-sized project only takes a few minutes.

1. filter and trim of the reference sequences (cutadapt)  
   Only sequences which contain both primers are passing the filter. Flanking primers and segments beyond are removed to obtain a homolog region block.
2. 3'-quality trimming of paired end query reads (trimmomatic)
3. Overlap-merge of paired-ends (vsearch)
4. Adjustment of orientation 5'-> 3' and trimming of primer segements (cutadapt)
5. Feature filter (vsearch)  
   A Min and max length filter and a "max number of expected errors" filter are implemented.
6. Chimera filter (vsearch)  
   Prediction of putative chimeric sequences sample-wise and removal of them from the sample if they do not occur in any other sample.
7. Swarm clustering (swarm)
8. Annotation by naive bayesian classifier (mothur)

**SLIM**

SLIM is an open-source web application that simplifies the creation and execution of metabarcoding data processing pipelines through an intuitive Graphic User Interface (GUI). It has been developed for environmental DNA purposes and is therefore an interesting choice for PacMAN. Currently it supports the following algorithms:

1. **Demultiplexing:** DTD: Demultiplex libraries from illumina outputs
2. **Paired-end read joiner**: Pandaseq, vsearch mergepair, CASPER
3. **Chimera detection:** vsearch uchime
4. **ASVs inference / OTUs clustering:** DADA2, vsearch uclust, swarm
5. **Sequence assignment:** vsearch usearch, IDTAXA
6. **Post-clustering:** LULU

Each step is added as a ‘module’ allowing flexibility for the user and project to decide on the exact workflow. While it looks easy to upload data, still the user needs to make the decisions at each step of the process and will require extensive knowledge on sequence analysis. Lacking a bit in explaining the steps and suggesting a workflow, but an interesting approach, which could be helpful for a project like PacMAN to share a workflow with all data collectors/owners.

**TOURMALINE**

Tourmaline is an amplicon sequence processing workflow for Illumina sequence data that uses [QIIME 2](https://qiime2.org) and the software packages it wraps. Tourmaline manages commands, inputs, and outputs using the [Snakemake](https://snakemake.readthedocs.io/en/stable/) workflow management system. **Tourmaline is an alternative amplicon 'pipeline' to** [**Banzai**](https://github.com/jimmyodonnell/banzai)**, which was developed for** [**MBON**](https://github.com/marinebon/MBON) **(Marine Biodiversity Observation Network) and uses** [**Swarm**](https://github.com/torognes/swarm) **for OTU picking.** Tourmaline provides rapid and reproducible workflows for [Deblur](https://github.com/biocore/deblur) and [DADA2](https://github.com/benjjneb/dada2) and is extensible to other OTU picking algorithms. Currently supported Qiime2 commands:

1. FASTQ sequence import using a manifest file, or use your pre-imported FASTQ qza file
2. Denoising with DADA2 (paired-end and single-end) and Deblur (single-end)
3. Feature classification (taxonomic assignment) with options: **consensus BLAST, naive Bayes**
4. Feature filtering by taxonomy and by feature ID
5. Interactive taxonomy barplot (qzv)
6. De novo multiple sequence alignment with MAFFT (with masking) and tree building with Fasttree
7. Tree visualization using Empress (qzv)
8. Interactive alpha-rarefaction plot (qzv)
9. Alpha diversity and alpha-rarefaction with choice of metrics (qzv)
10. Beta diversity and beta group significance with choice of metrics (qzv)
11. Principal coordinates plots with Emperor (qzv)

Tourmaline requires already demultiplexed (and primers removed?) files as dada2 also.

# Summary of available pipelines

As is clear this overview of pipelines made specifically for eDNA analysis the pipelines ultimately use mostly the same set of bioinformatic tools, despite one exception (MgNIFY), which has more broad use. The pipelines can be divided roughly in three types: those that use vsearch, uclust or swarm for OTU analysis (Banzai, PEMA, QZIP), those that use dada2 for ASV inference (Anacapa, Tourmaline), or those that have both options (Cascabel, MetabarToad, SLIM). In addition the Abyss pipeline optionally clusters and curates OTUs, after ASV inference. It is clear that as there is no strong standards for this work, each group makes a slightly modified version of the workflow to meet their needs. A short summary of the different tools utilized in each pipeline is shown in table 2.

The choice for OBIS and PacMAN then is if we should do the same, develop our own workflow which would contain everything we need and could be possibly developed into an easy-to-use GUI, or if we should use on of the existing pipelines.

While the PEMA pipeline is developed and standardized for a large project (ARMS) it does not incorporate dada2 for ASV detection at the moment, and therefore may not be the best choice for moving forward.

The most promising pipeline currently for PacMAN would be ANACAPA. This pipeline is meant for a large eDNA project in California, and has an automatic reference library generation step. However it is not sure how flexible it is in terms of developing it further and choosing different steps. In addition it doesn’t generate a report of the followed pipeline, which would be highly recommended for provenance.

A second choice, the CASCABEL pipeline is based on the flexible snakemake system, which allows easy documentation, version control, and use of the pipeline. It is largely based on Qiime, and contains many choices in terms of the different analysis steps. The workflow is also flexible in the sense that it could be possible to update it when necessary with new algorithms and workflows inserted in the pipeline. Tourmaline is similarly based on snakemake and in this case Qiime2 algorithms, however as of yet it does not contain as many of the tools and choices that Cascabel has.



# Recommendations

The minimal requirements for a pipeline for the PacMAN project would include the following:

1. Quality control
   1. Initial quality control before proceeding with the analysis
   2. Raises flags, if any of the values are worrying for manual inspection
2. Demultiplex
   1. Demultiplex sequences/sample. Currently not many of the pipelines do this
   2. Requires barcode information
3. Remove primers
   1. Remove primer sequences using cutadapt (recommended by dada2)
4. Denoise
   1. Dada2
   2. This pipeline is quickly gaining popularity
   3. It comprises steps for denoising, dereplication, pairing and chimera detection
5. Assign taxonomy
   1. Could take a multi-stepped approach
      1. First search for curated sequences of known invasive species
      2. Search a biomarker databases (BOLD, MIDORI)
      3. Search unclassified sequences with NCBI-nt
6. Package data
   1. Biom format
   2. DwC-A core format (built-in?)

If biodiversity assessments from this data are required it might make sense to cluster and curate the ASV’s as the variation in the sequences differs largely between species, masking the number of species that are detected.