# Bench sheet · EnGeL-BIOINF-OBITOOLS4

**Metabarcoding bioinformatics with OBITools4, VSEARCH, LULU and BOLDigger3 (COI amplicons; 18S nematode variant)** · version 1.2.0 (2026-11-20). Full protocol: EnGeL-BIOINF-OBITOOLS4. Tick each step when done; write deviations in Notes.

| | |
|---|---|
| Date | |
| Operator | |
| Plate ID / samples | |
| Software versions (OBITools4, VSEARCH, R, LULU, BOLDigger3) | |
| BOLD database download date | |

### 1. Working folders and raw data

| ☐ | No. | Step |
|---|---|---|
| ☐ | 1 | Set the working directory and create one folder per step: |
| ☐ | 2 | Move all raw .gz files (usually in one folder per sample) into `1.raw` and decompress them: |
| ☐ | 3 | Rename the files to `<sample>_R1.fastq` and `<sample>_R2.fastq` (the source script adds a missing "R" and shortens names; adapt the patterns to the provider's file names; Novogene delivers `<sample>.raw_1.fastq.gz` and `<sample>.raw_2.fastq.gz`) [TO CONFIRM: general renaming rule]. |
| ☐ | 4 | Start the run report, one row per metric, and define a short helper to add rows to it: |

### 2. Sample sheets

| ☐ | No. | Step |
|---|---|---|
| ☐ | 5 | Put the master sample sheet (Before you start) and the helper script in `files`, and split the sheet into one file per sample in `samplesheets`. Keep the master sheet out of `samplesheets`: every `.txt` file there is treated as one sample. |
| ☐ | 6 | Report the number of samples and the raw reads: |

### 3. Merge paired-end reads

| ☐ | No. | Step |
|---|---|---|
| ☐ | 7 | Align forward and reverse reads. With these settings, pairs that overlap by less than 10 bp or have less than 80% identity in the overlap are concatenated instead of aligned (`mode` = joined): |
| ☐ | 8 | Remove the unaligned (joined) records: |
| ☐ | 9 | Report the pairing settings and the reads kept and removed: |

### 4. Primer removal and sample assignment

| ☐ | No. | Step |
|---|---|---|
| ☐ | 10 | Remove primers and annotate each sequence with its sample information, allowing four mismatches in the primers (`-e 4`); unassigned sequences go to a separate file: |
| ☐ | 11 | Report the reads assigned and not assigned to a sample: |

### 5. Dereplication and singleton removal

| ☐ | No. | Step |
|---|---|---|
| ☐ | 12 | Collapse identical reads into exact sequence variants and keep only the `count` and `sample` annotations: |
| ☐ | 13 | Record statistics per sample in `10.stats` (merged reads at the start, unique sequences, singletons) with `obicount`, and report the totals: |
| ☐ | 14 | Remove singletons and check that no sequence with count 1 remains: |

### 6. Merge samples and inspect

| ☐ | No. | Step |
|---|---|---|
| ☐ | 15 | Concatenate all samples and report the total: `cat ./6.nosingletons/*.fasta > all.fasta`, then `report "Sanity" "Merge samples" "Sequences after merging all samples" $(nseqs all.fasta) sequences "All samples"`. |
| ☐ | 16 | In R, plot the number of occurrences per variant and the sequence length distribution, and add the summary to the report. The same function is used again after denoising (step 7) and after chimera removal: |

### 7. VSEARCH: denoising, chimeras and clustering

| ☐ | No. | Step |
|---|---|---|
| ☐ | 17 | Convert the headers to VSEARCH format (`count` → `size`): |
| ☐ | 18 | Dereplicate across the whole dataset and relabel: |
| ☐ | 19 | Denoise with UNOISE, keeping sequences of the expected length (COI: 400–450 bp; target 418 bp): |
| ☐ | 20 | Remove chimeras (UCHIME3 de novo; by default a parent must be at least 16 times more abundant than its chimera): |
| ☐ | 21 | Report the sequences after each VSEARCH step and inspect the denoised and non-chimeric sequences with the R function of step 6 (count pattern `'(?<=size=)\\d+'`, `binwidth = 1`): |

| ☐ | No. | Step |
|---|---|---|
| ☐ | 22 | Cluster into OTUs at 96% similarity. Adapt the threshold to the intraspecific genetic variability expected for the target group: |
| **!** | | **CRITICAL** 96% is the lab rule: 3–4% divergence is acceptable within an invertebrate (e.g. arthropod) species. Use 99% (`--id 0.99`) only for groups known to have very low intraspecific variability, and record the threshold used. (The December 2025 scripts used 99%.) |

### 8. LULU curation (erroneous OTUs and NUMTs)

| ☐ | No. | Step |
|---|---|---|
| ☐ | 23 | Build the OTU table (OTU frequency per sample) and the match list (pairwise similarity among OTUs): |
| ☐ | 24 | Run LULU in R and write the curated table and the curated OTU sequences: |
| ☐ | 25 | Copy the curated OTU table to `reports`: `cp ./9.lulu/otutable_curated.tsv ./reports/otutable_curated.txt`, and report the OTU table size (OTUs, samples, reads assigned) and the number of curated OTUs: |

### 9. Taxonomic assignment (BOLDigger3)

| ☐ | No. | Step |
|---|---|---|
| ☐ | 26 | Identify the curated OTUs against the local BOLD database: |

### 10. Variant: 18S nematode amplicons (NF1/18Sr2b)

| ☐ | No. | Step |
|---|---|---|
| ☐ | 27 | Step 3: only merged (aligned) read pairs are used. Pairs that do not merge well are discarded for now; how to use unmerged pairs is under development in the lab and not part of this protocol. |
| ☐ | 28 | Step 7, UNOISE: keep sequences of 280–430 bp (`--minseqlength 280 --maxseqlength 430`), and use the same limits in the R inspection. This range comes from the December 2025 primer test, which also included TAReukFWD1/TAReukREV3 (about 390 bp); it is provisional for NF1/18Sr2b alone [TO CONFIRM: final range]. Amplicon length varies a little between species. |
| ☐ | 29 | Step 7, clustering: 96%, as for COI, unless the group is known to have very low intraspecific variability. |
| ☐ | 30 | Instead of step 9, identify the curated OTUs against 18S-NemaBase with VSEARCH (global alignment, at least 84% identity over at least 90% of the query; all hits up to 10 are kept, sorted by identity): |
| ☐ | 31 | Split the taxonomy of the best hit into ranks in R. Each 18S-NemaBase header holds the accession followed by 16 ranks separated by semicolons (domain to species); missing ranks are written as `NA_<rank>`: |
| ☐ | 32 | Accept a name only at the rank its identity supports: species at ≥ 99%, genus at ≥ 97%, family at ≥ 95%. Below these, report the next higher rank that is supported (for example the family, or only “Nematoda”); never force a genus or species name from a weak match: |
| **!** | | **CRITICAL** The search keeps matches from 84% so that distant OTUs are still placed in Nematoda; 84% does not support a genus or species name. Record the thresholds used. 18S reference coverage for nematodes is incomplete, so many OTUs will stop at family or above. |

### 11. Read losses per step and per sample

| ☐ | No. | Step |
|---|---|---|
| ☐ | 33 | Collect the reads of every sample at each step into one table (long format: sample, step, reads): |
| ☐ | 34 | In R, compute for each sample and step the reads in, reads out, reads lost and percentage lost, flag any step that loses more than half of its input, and add dataset totals: |
| ☐ | 35 | Check every flagged step before going on. A large loss at merging usually means short overlaps (amplicon too long for the reads), at primer assignment a sample-sheet or primer problem, and at singleton removal a sample with very few reads. |

**Notes / deviations**

| |
|---|
| |
| |