STANDARD OPERATING PROTOCOL · EnGeL-BIOINF-COMMUNITY

From OTU table to community data: identified OTUs, incidence and shared OTUs between groups (R)

Protocol ID	EnGeL-BIOINF-COMMUNITY
Version	1.0.1
Date	2026-11-17
Status	draft
Area	Bioinformatics (BIOINF)
Authors	Luísa Fraga (lead; analysis script; ORCID 0000-0003-3153-4593); Luís Cunha (review and approval; ORCID 0000-0002-5870-2537)
Reviewed and approved by	Luís Cunha
Laboratory	Environmental Genomics Laboratory (EnGeL-UC), Environmental Ecology & Ecotoxicology Group (E3G), Centre for Functional Ecology, Department of Life Sciences, University of Coimbra · Laboratório Associado TERRA
Contact	luis.cunha@uc.pt · ewormuc@gmail.com
Licence	Text: CC BY 4.0 · Code: MIT
Main hazards	None: computational protocol

1 Purpose and scope

This protocol takes the curated OTU table and the taxonomic identifications produced by EnGeL-BIOINF-OBITOOLS4 and prepares them for community analysis: it keeps the OTUs that received an identification, converts read counts into presence/absence (incidence) data, and shows which OTUs are shared between groups of samples (for example treatments, sites or methods) with a Venn diagram. The incidence table is also the input for sampling-effort and completeness analyses (EnGeL-SAMPL-EFFORT).

2 Principle

The OTU table (OTUs × samples, read counts) is joined to the identification table by OTU name, so that only OTUs with an identification are kept. Because read numbers in metabarcoding do not translate directly into abundance, each count is turned into presence (1) or absence (0). An OTU is counted as present in a group when it occurs in at least one sample of that group; the overlaps between groups are drawn as a Venn diagram (Gao et al., 2021).

3 Before you start

Item	Details
Hands-on time	[TO CONFIRM]
Input	Outputs of EnGeL-BIOINF-OBITOOLS4 (curated OTU table and identification table); sample-group table
Prepare in advance	R project folder set up as in the lab's file and folder management guide (`data/raw`, `data/processed`, `scripts`, `results`); R packages installed
Key safety points	Not applicable

4 Materials

4.1 Computer

- Computer with R and RStudio (lab workstation or personal computer)

4.2 Consumables

- None

4.3 Software and input files

- R [TO CONFIRM: version] with the packages readxl, tibble, writexl, dplyr and ggVennDiagram [TO CONFIRM: package versions]
- Curated OTU table from EnGeL-BIOINF-OBITOOLS4 (9.lulu/otutable_curated.tsv, also copied to reports/otutable_curated.txt)
- BOLDigger3 identification table for the curated OTUs, matches only (.xlsx) [TO CONFIRM: exact file produced by BOLDigger3]
- Sample-group table (.csv): one row per sample, columns sample and group

5 Reagent preparation

Sample-group table. Create `data/raw/sample_groups.csv` with one row per sample column of the OTU table: column `sample` (exactly as in the OTU table) and column `group`. Negative controls (extraction blanks, PCR negatives) get their own group, for example `blank`.

```
sample,group
PA1,BF3_BR2_Novogene
PA2,BF3_BR2_Novogene
PCB,blank
```

6 Procedure

6.1 1. Set up

1. Open the project's R project (paths below are relative to the project folder; do not use `setwd()` with personal paths).
2. Load the packages:

```
library(readxl)
library(tibble)
library(writexl)
library(dplyr)
library(ggVennDiagram)
```

6.2 2. Read and join the tables

3. Read the identification table and the curated OTU table, and check the dimensions (OTUs × samples) against the EnGeL-BIOINF-OBITOOLS4 report:

```
tax <- as.data.frame(read_xlsx("data/raw/curated_OTUs_identification_result_MatchesOnly.xlsx"))
otu <- as.data.frame(read.delim("data/raw/otutable_curated.tsv"))
dim(otu)
```

4. Give both tables the same OTU column name and keep only OTUs that have an identification:

```
colnames(tax)[colnames(tax) == "id"] <- "otu"
otu <- rownames_to_column(otu, var = "otu")
otu_kept <- merge(otu, tax, by = "otu")
dim(otu_kept)
write_xlsx(otu_kept, "results/otutable_keptOTUs.xlsx")
```

5. Keep only the sample columns, with OTU names as row names:

```
groups  <- read.csv("data/raw/sample_groups.csv")
samples <- intersect(groups$sample, colnames(otu_kept))
counts  <- otu_kept[, samples]
rownames(counts) <- otu_kept$otu
```

CRITICAL Every sample column of the OTU table must appear in the sample-group table, spelled the same way. Check that `length(samples)` equals the number of samples you expect.

6.3 3. Incidence (presence/absence)

6. Convert read counts to presence/absence (1/0) and save the incidence table:

```
incidence <- counts %>% mutate(across(everything(), ~ ifelse(. > 0, 1, 0)))
write_xlsx(rownames_to_column(incidence, "otu"), "results/OTU_incidence.xlsx")
```

6.4 4. OTUs shared between groups

7. List, for each group, the OTUs present in at least one of its samples. Negative controls are never counted in a sample group:

```
use   <- setdiff(unique(groups$group), "blank")
venn  <- lapply(use, function(g) {
  cols <- groups$sample[groups$group == g]
  rownames(incidence)[rowSums(incidence[, cols, drop = FALSE]) > 0]
})
names(venn) <- use
```

8. Draw and save the Venn diagram (8 × 6 inches, 300 dpi):

```
venn_plot <- ggVennDiagram(venn) +
  scale_fill_gradient(low = "#EDD9A3", high = "#F8765C") +
  scale_x_continuous(expand = expansion(mult = .2))
ggsave("results/venn_plot.png", plot = venn_plot, width = 8, height = 6, dpi = 300)
```

The diagram works for two to four groups; with more groups, use a table of shared OTUs instead [TO CONFIRM: preferred alternative, e.g. an UpSet plot].

7 Calculations

Not applicable.

8 Quality control

Check	Where	Acceptance criterion
Number of samples read	`dim()` after reading; `length(samples)`	Equal to the number of samples sequenced and listed in the sample-group table
OTUs kept after the join	`dim(otu_kept)`	[TO CONFIRM: what proportion of unidentified OTUs is acceptable]
Negative controls	Incidence table, `blank` group	[TO CONFIRM: how OTUs found in blanks are treated]

9 Data and metadata

- **Outputs:** `results/otutable_keptOTUs.xlsx` (identified OTUs with counts and taxonomy), `results/OTU_incidence.xlsx` (presence/absence), `results/venn_plot.png`.
- **Code:** keep the R script in the project's `scripts` folder; the script that follows this protocol is to be added to this repository (`code/`) [TO CONFIRM].
- **Where:** project folder in the lab's shared folder.

> **Minimum to record**
> - Input files used (names and dates) and the EnGeL-BIOINF-OBITOOLS4 run they come from
> - R and package versions (`sessionInfo()`)
> - Sample-group table used
> - Number of OTUs before and after the join; number of samples

10 Safety and waste

Not applicable (computational protocol).

11 Troubleshooting

Problem	Likely cause	Solution
Fewer samples than expected after selecting columns	Sample names differ between the OTU table and the sample-group table	Compare `colnames(otu_kept)` with `groups$sample`; correct the table.
Almost no OTUs kept after the join	OTU names differ between the two tables (e.g. a prefix), or the identification column is not called `id`	Compare the first OTU names in both tables and rename as needed.
Venn diagram unreadable	Too many groups	Use at most four groups, or a table of shared OTUs.

12 Remarks

The source analysis compared three primer and library-preparation combinations for COI metabarcoding of mesofauna (BF3/BR2 and III_BF3/Fol-degen-rev, prepared by two providers), with one sample group per combination and a blank; the groups and file names above follow it as an example. In the source, the blank was counted in one of the sample groups when drawing the Venn diagram; this is corrected here: negative controls form their own group and are left out of the sample groups.

13 References

Gao, C.-H., Yu, G., Cai, P. (2021). ggVennDiagram: an intuitive, easy-to-use, and highly customizable R package to generate Venn diagram. Frontiers in Genetics, 12, 706907. https://doi.org/10.3389/fgene.2021.706907

14 Version history and citation

Version	Date	Change	By
1.0.0	2026-11-17	First EnGeL-UC version, from the Quarto analysis "Data Analysis – Primer Comparison" by Luísa F. Dornellas (2026); sample groups read from a table instead of written into the script.	Luís Cunha
1.0.1	2026-11-17	Negative controls kept out of the sample groups (corrects the source analysis, where the blank was counted in one group).	Luís Cunha

How to cite

Fraga L, Cunha L (2026). From OTU table to community data: identified OTUs, incidence and shared OTUs between groups (R). EnGeL-UC protocol EnGeL-BIOINF-COMMUNITY, version 1.0.1. Zenodo. DOI [TO CONFIRM].

Licence

Text: CC BY 4.0. Code: MIT. Logos are the property of their institutions and are not covered by the licence.

15 Open questions

1. R and package versions; hands-on time.
2. Exact name of the BOLDigger3 identification file ("MatchesOnly") and the identification threshold it implies.
3. How OTUs found in negative controls are treated; acceptable proportion of unidentified OTUs.
4. Alternative to the Venn diagram for more than four groups.
5. Further community analyses to add (e.g. diversity indices, ordination)?
6. Add the R script to the repository (`code/`, MIT)?
7. DOI.