AtlantECO RoCSI-CPR eDNA Metabarcoding — Combined 18S & COI Report

AUTHOR
Nathan Hubot

🧬 Overview of Sequencing & Experimental Design

This report summarises the **two metabarcoding workflows** applied to the RoCSI-CPR eDNA samples:

- **18S rRNA V9 region** (targeting phytoplankton)
- **COI mitochondrial region** (targeting zooplankton)

Both datasets were sequenced at the
Centre for Genomic Research (CGR), University of Liverpool.

📊 Sequencing Summary

Feature	18S rRNA V9	COI (metazoans)
SSP ID	SSP202877	SSP200993
Purchase order	203631529	P10817-5
Sequencing platform	Illumina MiSeq v2 (2×150 bp)	Illumina MiSeq v2 (2×250 bp)
Target region	18S V9 (1389F–1510R)	COI Leray fragment (m1COIintF–jgHCO2198)
Expected amplicon size	~121 bp	~313 bp
Raw read depth (mean)	~150k	~200k
Pre-trimming QC	Cutadapt v4.5	Cutadapt v1.2.1 + Sickle v1.200
Downstream processing	Cutadapt → DADA2 → MZG 18S	Cutadapt → DADA2 → MZG COI

🧬 Primer Sets (with CGR Overhangs)

18S V9 Primers

- **Forward:**
  5′ **ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNN**
  TTGTACACACCGCCC 3′
  (CGR overhang + spacer in blue; 18S primer in red)

- **Reverse:**
  5′ **GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**
  CCTTCYGCAGGTTCACCTAC 3′

COI Primers

- **Forward:**
  5′ **ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNN**
  GGWACWGGWTGAACWGTWTAYCCYCC 3′
  (m1COIintF)

- **Reverse:**
  5′ **GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**
  TAIACYTCIGGRTGICCRAARAAYCA 3′
  (jgHCO2198)

🔧 Bioinformatic Workflow Overview

Both markers follow the same general pipeline:

1. **Primer removal** (Cutadapt)
2. **Quality filtering** (DADA2)
3. **Error learning and denoising** (DADA2)
4. **Read merging** (DADA2)
5. **Chimera removal** (DADA2)
6. **Taxonomic assignment:** (DADA2)
   - 18S → *MZG 18S "All Microbes + Protists", Mode-A* reference database
   - COI → *MZG COI "All Invertebrates", Mode-A* reference database

For each step, differences between the two pipelines are shown.

1️⃣ Primer Removal with Cutadapt

🔹 Summary of Cutadapt parameters used

Parameter	18S	COI
Forward primer (`-g`)	TTGTACACACCGCCC	GGWACWGGWTGAACWGTWTAYCCYCC
Reverse primer (`-G`)	CCTTCYGCAGGTTCACCTAC	TANACYTCNGGRTGNCCRAARAAYCA
`--match-read-wildcards`	✓	✓
Minimum overlap	**10**	**20**
Max error rate (`-e`)	**0.15**	**0.20**

Parameter	18S	COI
`--discard-untrimmed`	✓	✓
`--minimum-length`	80 bp	200 bp
Pre-processing by CGR	Adapters trimmed via Cutadapt 4.5	Adapters trimmed (Cutadapt v1.2.1) + Sickle quality filtering

🔹 Unified Cutadapt description

Both datasets used a looping bash command of the form:

```
cutadapt\
  -g <forward_primer> \
  -G <reverse_primer> \
  --match-read-wildcards \
  --overlap <OV> \
  -e <ERROR> \
  --pair-filter=both \
  --discard-untrimmed \
  --cores=0 \
  -o $out1 -p $out2 \
  $f $r
```

Where and differ between markers (see table above).

2️⃣ DADA2 Processing

🔹 Summary of Cutadapt parameters used

Parameters	18S	COI
`truncLen`	c(130,120)	c(210,210)
`maxEE`	c(2,3)	c(2,3)
`minLen`	80	200
`minOverlap` (mergePairs)	30	90
Ref database	MZG 18S	MZG COI

MZG Reference databases

Marker	Database	Notes
18S	`MZGdada2-18s__T2000000__o00__A.fastq`	phytoplankton –> "All Microbes + Protists", mode-A

Marker	Database	Notes
COI	`MZGdada2-coi__T4000000__o00__A.fastq`	zooplankton –> “All invertebrates”, mode-A

source: https://metazoogene.org/mzgdb/atlas/html-src/data__T4000000__o00.html

Unified DADA2 description

Both datasets were processed with the standard DADA2 workflow:

```
# Filtering and trimming (R1/R2 after cutadapt)
filtered_out <- filterAndTrim(
  fwd   = forward_reads,
  filt = filtered_forward_reads,
  rev   = reverse_reads,
  filt.rev = filtered_reverse_reads,
  truncLen   = <MARKER_SPECIFIC>,    # see table above
  maxEE      = c(2, 3),              # expected errors (stricter for R1)
  maxN       = 0,                    # discard reads with Ns
  rm.phix    = TRUE,                 # remove PhiX reads
  minLen     = <MINLEN>,             # marker-specific minimum length
  multithread = TRUE
)

# Error learning
errF <- learnErrors(filtered_forward_reads, multithread=TRUE)
errR <- learnErrors(filtered_reverse_reads, multithread=TRUE)

# Dereplication
derepF <- derepFastq(filtered_forward_reads)
derepR <- derepFastq(filtered_reverse_reads)

# ASV inference
dadaF <- dada(derepF, err=errF, pool="pseudo")
dadaR <- dada(derepR, err=errR, pool="pseudo")

# Merging
merged <- mergePairs(dadaF, derepF, dadaR, derepR,
                     minOverlap = <MARKER_SPECIFIC>,
                     trimOverhang = TRUE)

# ASV table
seqtab <- makeSequenceTable(merged)

# Chimera removal
seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus")

# Taxonomic Assignment
taxa <- assignTaxonomy(
  seqtab.nochim,
  refFasta = <REF_FASTA>,
  multithread = TRUE
)
```

Note

For 18S, shorter reads (150 bp) and a very short amplicon (~121 bp) justify relatively short truncLen (130, 120) and minLen = 80. For COI, the longer amplicon fragment (~313 bp) and 2×250 bp reads allow more aggressive truncation (210, 210) with large overlap (90) and a higher minLen = 200 to remove spurious short fragments.

🔍 Read Tracking Summary

The following table summarizes read counts at each step of the DADA2 pipeline:

DADA2 Counts for 18S and COI

	18S				COI			
sample	**input**	**filtered**	**nonchim**	**reads_retained (%)**	**input**	**filtered**	**nonchim**	**reads_retained (%)**
01-CPR_1_ID_1_	148213	128862	126330	85.2	144059	139400	104643	72.6
02-CPR_1_ID_2_	121841	105245	103807	85.2	91034	86400	82657	90.8
03-CPR_1_ID_3_	117632	96408	94945	80.7	77206	75069	65275	84.5
04-CPR_1_ID_4_	170195	135708	133962	78.7	99005	95927	88038	88.9
05-CPR_1_ID_5_	139036	124966	123177	88.6	143700	136891	130456	90.8
06-CPR_1_ID_7_	189844	172629	170302	89.7	100258	95319	91819	91.6
07-CPR_1_ID_8_	156940	134665	132550	84.5	163332	156331	150014	91.8
08-CPR_1_ID_9_	169359	155134	153346	90.5	56907	47951	46489	81.7
09-CPR_1_ID_10_	149202	134894	132964	89.1	164215	149580	141147	86.0
10-CPR_1_ID_12_	133358	109353	108318	81.2	82507	79597	75675	91.7
11-CPR_1_ID_13_	159171	119466	118219	74.3	133207	128803	122620	92.1
12-CPR_1_ID_14_	154920	82372	81798	52.8	127207	123701	117516	92.4
13-CPR_1_ID_15_	165533	100962	98756	59.7	181637	174765	167385	92.2
14-CPR_1_ID_16_	174055	154970	153105	88.0	137952	131771	126534	91.7

	18S				COI			
sample	input	filtered	nonchim	reads_retained (%)	input	filtered	nonchim	reads_retained (%)
15-CPR_1_ID_18_	187464	141613	136443	72.8	119723	113684	109776	91.7
16-CPR_1_ID_19_	199073	165283	163680	82.2	57654	55146	51706	89.7
17-CPR_1_ID_20_	121858	107172	105402	86.5	70524	67331	64356	91.3
18-CPR_1_ID_21_	150714	142004	140604	93.3	207441	199025	193426	93.2
19-CPR_1_ID_22_	121155	98484	96981	80.0	174619	168282	162395	93.0
20-CPR_1_ID_23_	80435	66396	65169	81.0	119162	113276	109138	91.6
21-CPR_1_ID_24_	104984	92580	90811	86.5	185903	179594	173934	93.6
22-CPR_1_ID_26_	236346	191192	188534	79.8	202369	196976	190007	93.9
23-CPR_1_ID_27_	156485	137223	134112	85.7	101725	97126	92636	91.1
24-CPR_1_ID_28_	127353	111643	109821	86.2	170221	162524	155675	91.5
25-CPR_1_ID_29_	132416	116805	115151	87.0	139384	134544	127957	91.8
26-CPR_1_ID_30_	172100	162217	160855	93.5	112727	107640	104009	92.3
27-CPR_1_ID_31_	178342	167885	165400	92.7	119777	114101	108237	90.4
28-CPR_1_ID_32_	183002	172658	170208	93.0	163244	156129	151718	92.9
29-CPR_1_ID_34_	158952	143494	141092	88.8	136586	128685	124018	90.8
30-CPR_2_ID_1_	194720	169302	166858	85.7	139205	135316	129682	93.2
31-CPR_2_ID_3_	194060	171082	168250	86.7	212589	206838	198641	93.4
32-CPR_2_ID_4_	179081	148520	145983	81.5	53088	51392	48520	91.4
33-CPR_2_ID_5_	166777	142691	140932	84.5	237970	219557	206262	86.7
34-CPR_2_ID_6_	149638	132537	130439	87.2	77732	75401	71710	92.3

	18S				COI			
sample	input	filtered	nonchim	reads_retained (%)	input	filtered	nonchim	reads_retained (%)
35-CPR_2_ID_7_	174663	145714	143138	82.0	81727	79459	75761	92.7
36-CPR_2_ID_9_	171703	141792	139921	81.5	207408	202506	195469	94.2
37-CPR_2_ID_10_	153543	137488	134897	87.9	144427	141116	135198	93.6
38-CPR_2_ID_11_	180363	152422	150795	83.6	221320	215203	207024	93.5
39-CPR_2_ID_12_	150636	122151	120213	79.8	365807	355825	342978	93.8
40-CPR_2_ID_13_	194784	180872	177058	90.9	135022	131227	124090	91.9
41-CPR_2_ID_15_	197928	174280	172050	86.9	193553	188006	180216	93.1
42-CPR_2_ID_16_	188135	175790	174228	92.6	147285	143187	137576	93.4
43-CPR_2_ID_17_	174942	145002	142740	81.6	174113	170004	163659	94.0
44-CPR_2_ID_18_	153984	135955	134263	87.2	131276	127604	122426	93.3
45-CPR_2_ID_19_	194789	161953	158427	81.3	128156	124223	117511	91.7
46-CPR_2_ID_21_	194608	174043	171831	88.3	160623	156266	149656	93.2
47-CPR_2_ID_22_	160941	131068	128824	80.0	183353	178208	171636	93.6
48-CPR_2_ID_23_	165931	148535	146897	88.5	89249	86797	83044	93.0
49-CPR_2_ID_24_	54433	50624	50116	92.1	116076	110431	104890	90.4
50-CPR_2_ID_25_	117781	101928	100424	85.3	125438	119556	113333	90.3
51-CPR_2_ID_27_	142756	107233	105385	73.8	30590	29864	27751	90.7
52-CPR_3_ID_1_	111133	85681	83994	75.6	188262	182359	174176	92.5
53-CPR_3_ID_3_	131736	110547	109180	82.9	149259	144930	138719	92.9
54-CPR_3_ID_4_	144501	127699	125997	87.2	348047	337956	325604	93.6

	18S				COI			
sample	input	filtered	nonchim	reads_retained (%)	input	filtered	nonchim	reads_retained (%)
55-CPR_3_ID_5_	154373	138771	137538	89.1	162972	157261	151439	92.9
56-CPR_3_ID_6_	150617	135967	134698	89.4	235204	228457	220339	93.7
57-CPR_3_ID_7_	199644	184698	182638	91.5	165181	159075	150736	91.3
58-CPR_3_ID_9_	176030	161246	159690	90.7	101926	98194	93307	91.5
59-CPR_3_ID_10_	138047	128819	127705	92.5	78978	75907	72554	91.9
60-CPR_3_ID_11_	127549	111467	109994	86.2	764251	743533	713247	93.3
61-CPR_3_ID_12_	99488	89824	88534	89.0	152306	147417	141727	93.1
62-CPR_3_ID_13_	144619	129452	128198	88.6	127896	124206	119095	93.1
63-CPR_3_ID_15_	173235	144180	142702	82.4	292673	282276	271096	92.6
64-CPR_3_ID_16_	178289	143108	141138	79.2	34818	33694	31628	90.8
65-CPR_3_ID_17_	162707	126581	124884	76.8	250949	244820	234509	93.4
66-CPR_3_ID_18_	131876	106565	104364	79.1	110678	108129	103754	93.7
67-CPR_3_ID_19_	128988	102637	100914	78.2	58035	56351	53356	91.9
68-CPR_3_ID_21_	107663	92109	90482	84.0	172563	165373	158612	91.9
69-CPR_3_ID_22_	112212	96684	94953	84.6	173570	168441	158656	91.4
70-CPR_3_ID_23_	132352	113912	112257	84.8	148617	143765	137513	92.5
71-CPR_3_ID_24_	196535	174115	171928	87.5	405832	395478	382290	94.2
72-CPR_4_ID_1_	196923	166713	164208	83.4	35382	33909	31733	89.7
73-CPR_4_ID_3_	213738	175220	171992	80.5	156704	143311	131255	83.8
74-CPR_4_ID_4_	205473	178940	176603	85.9	112970	109524	103878	92.0

	18S				COI			
sample	input	filtered	nonchim	reads_retained (%)	input	filtered	nonchim	reads_retained (%)
75-CPR_4_ID_5_	189932	165289	163135	85.9	35141	33968	31796	90.5
76-CPR_4_ID_7_	197530	166801	164410	83.2	194218	185074	176431	90.8
77-CPR_4_ID_8_	235558	194621	191317	81.2	113842	110778	103904	91.3
78-CPR_4_ID_9_	230985	193352	189591	82.1	176304	168346	160614	91.1
79-CPR_4_ID_11_	218815	181431	177308	81.0	98280	90284	84739	86.2
80-CPR_4_ID_12_	214875	172492	168539	78.4	43775	42215	40632	92.8
81-CPR_4_ID_13_	206448	170783	162683	78.8	343639	332962	312529	90.9
82-CPR_4_ID_15_	193011	156690	150368	77.9	262846	253570	243480	92.6
83-CPR_4_ID_16_	162913	125167	122596	75.3	11798	11246	10326	87.5
84-CPR_4_ID_17_	142434	107027	105677	74.2	102915	98756	94552	91.9
85-CPR_4_ID_19_	228236	196262	193991	85.0	185461	180209	167045	90.1

Negative Controls

The **No-Template Control (NTC)** and **Extraction Blank** samples in the 18S dataset did not pass the initial **DADA2 filtering step**, resulting in zero retained reads after quality filtering.
Consequently, these controls were **excluded from downstream analysis** (denoising, merging, and taxonomy assignment).

Their exclusion is consistent with expectations for negative controls, indicating the absence of detectable contamination above sequencing background levels.

💬 **Discussion — reads retained**

The proportion of reads that were kept following the DADA2 pipeline is good: **84%** and **91%** for *18S and **COI**, respectively. For 18S, samples 12 and 13 lost more then the rest: **53%** and **60%**, respectively.

🧮 Next Steps

At this stage, for each of the datasets/molecular markers, we have:

- An **OTU abundance table** (`seqtab.nochim`)
- A **taxonomy table** (`taxa`)

These files (provided) can be imported into **R** using the phyloseq package downstream analysis and visualization.

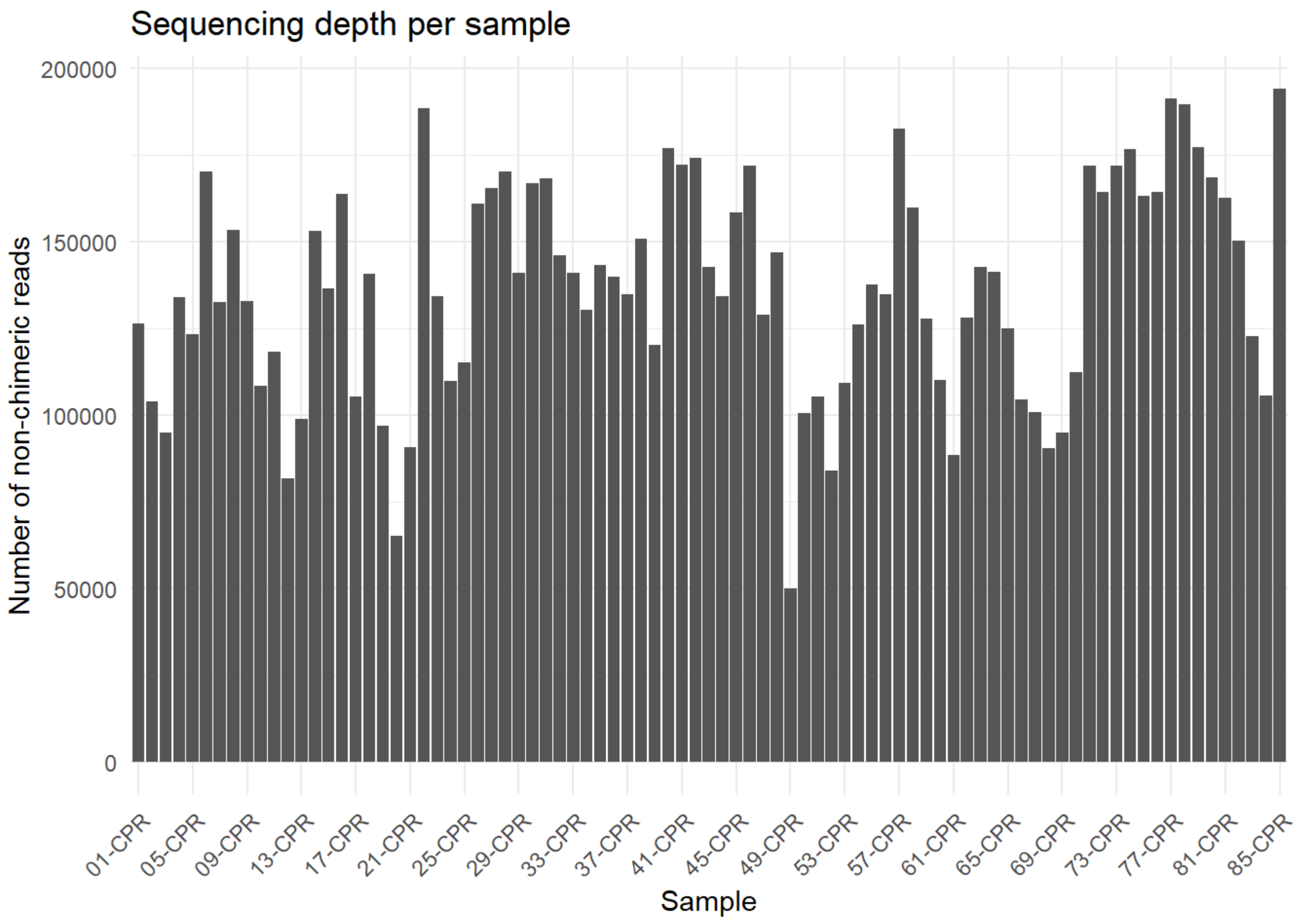
📦 Build the phyloseq object (18S)

```
🔹 ** A summary of the phyloseq object**

phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 5154 taxa and 85 samples ]
sample_data() Sample Data:       [ 85 samples by 8 sample variables ]
tax_table()   Taxonomy Table:    [ 5154 taxa by 20 taxonomic ranks ]

🔹 **These are the metadata variables:**

[1] "cpr"      "ID"       "Sample"   "Number"   "Inst"     "lat"      "lon"
[8] "Position"
```

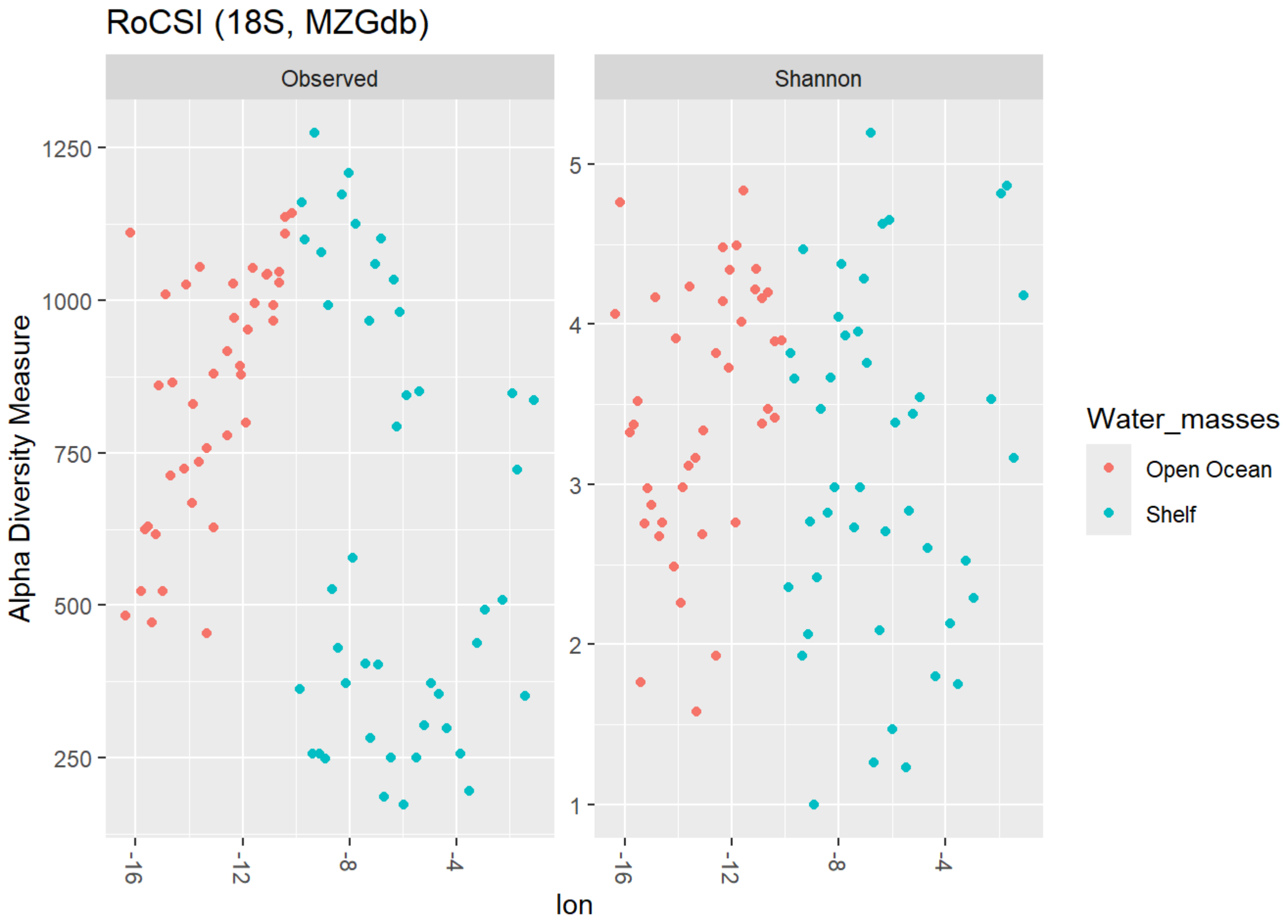
📊 Data visualisation

In this section, we explore the RoCSI-CPR 18S community structure using the `phyloseq` object (`ps`) generated above. We start by visualising read depth per sample, followed by basic taxonomic composition summaries.

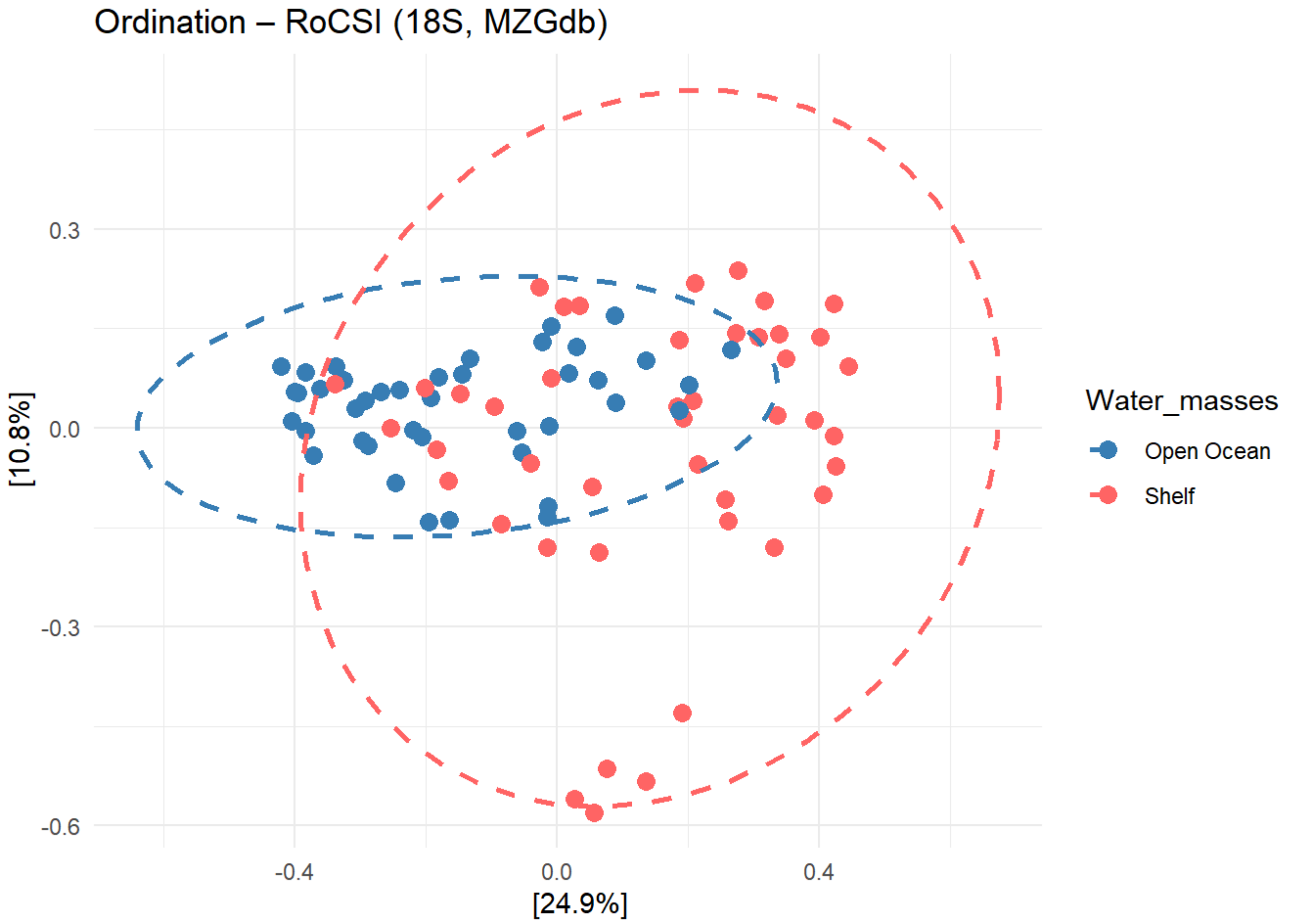
Read depth per sample



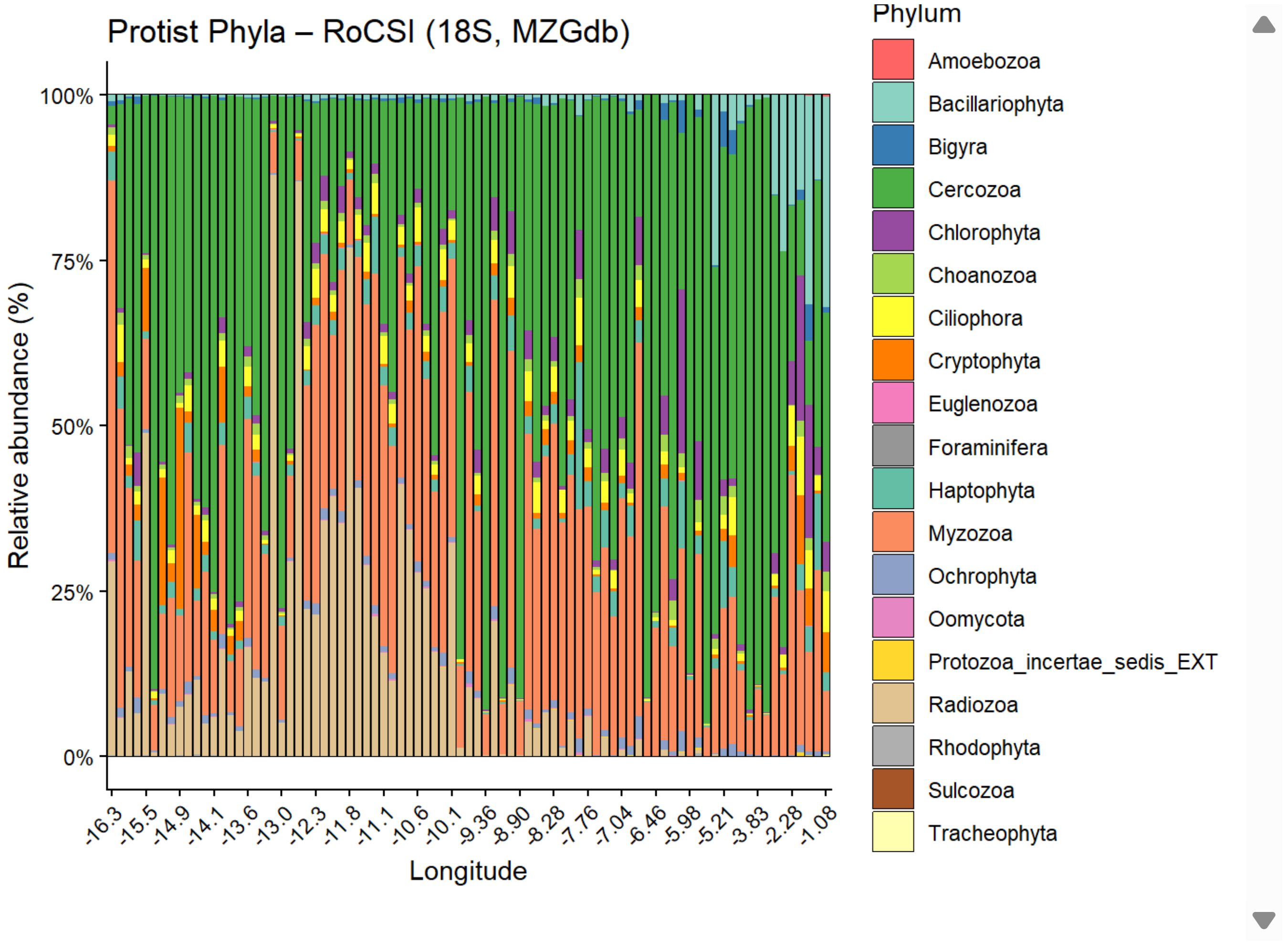
Alpha diversity



Ordination



Phylum composition

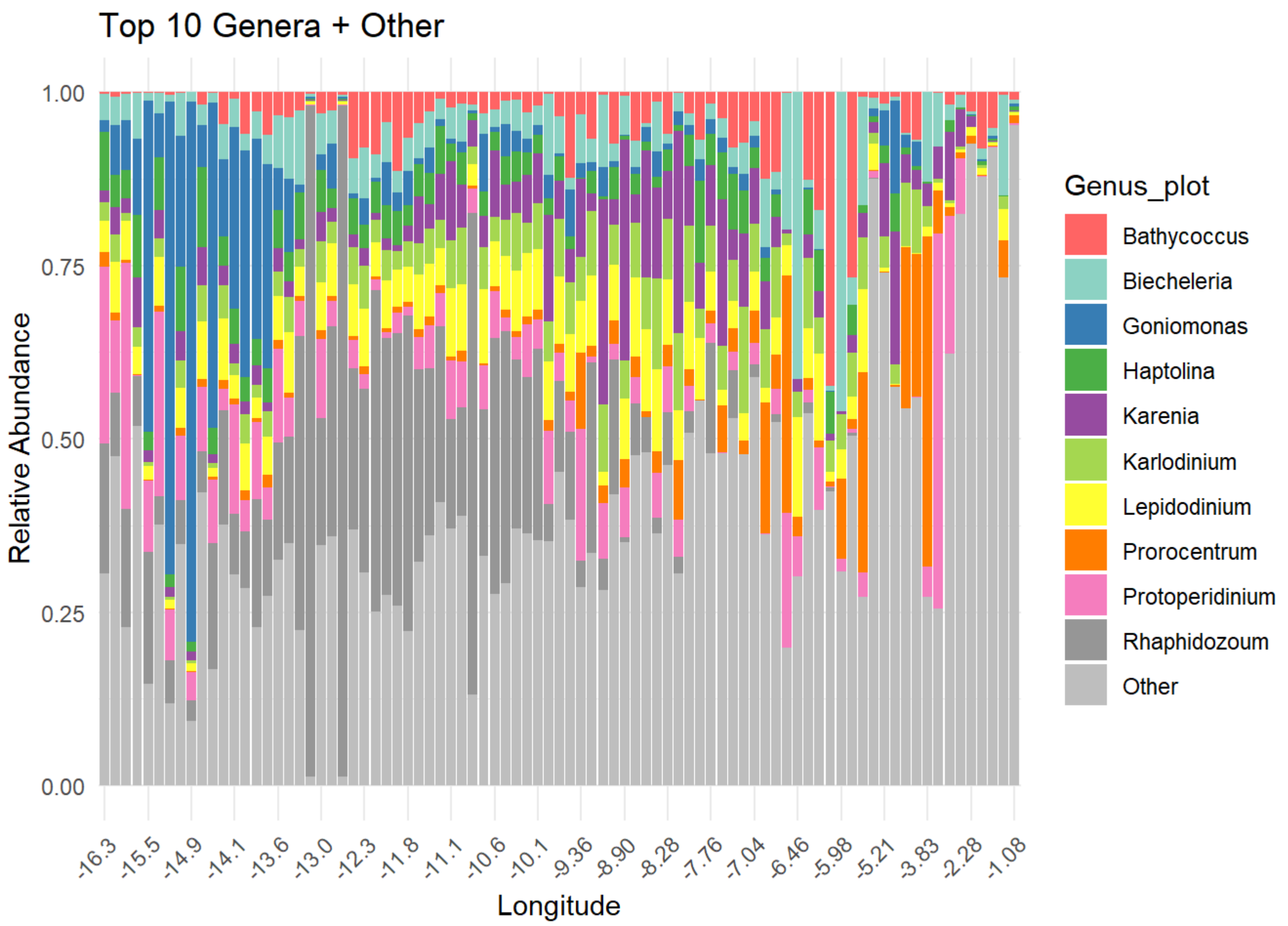


Discussion — Protist Community Structure

The 18S protist communities along the RoCSI transect are dominated by **Cercozoans** (Heterotrophic protists/grazers) and **Myzozoans** (subphylum Dinoflagellata). **Radiolarians** (unicellular eukaryotes) are abundant in the open-water section of the transect, while **Bacillariophyta** (Diatoms) are abundant at the beginning of the cruise.

Is this phyla composition ecologically coherent with expected shelf/open-ocean plankton dynamics during early spring in the English channel/Celtic Sea?

Top 10 genera



Discussion — Dominant protist genera

Breaking down the 18S community at the genus level reveals a mixture of **picophytoplankton (Bathycoccus), mixotrophic and autotrophic dinoflagellates (Karlodinium, Karenia, Prorocentrum and Lepidodinium), and heterotrophic grazers (Goniomonas, Protoperidinium)**.

The high abundance of dinoflagellates **Karlodinium, Karenia and Lepidodinium** at the transition between shelp to open-ocean might be linked to the observed phytoplankton blooms seen from the satelite data and chlorophll index of the CPR (see below).

Let’s note the high abundance of **Goniomonas** (Phagotrophic micrograzer) in the open-ocean!

Probably: shelf –> well mixed, Open-ocean –> more stratified…

Note: The “other” genius (including all the less abundant genera) is large, especially at low longitudes…

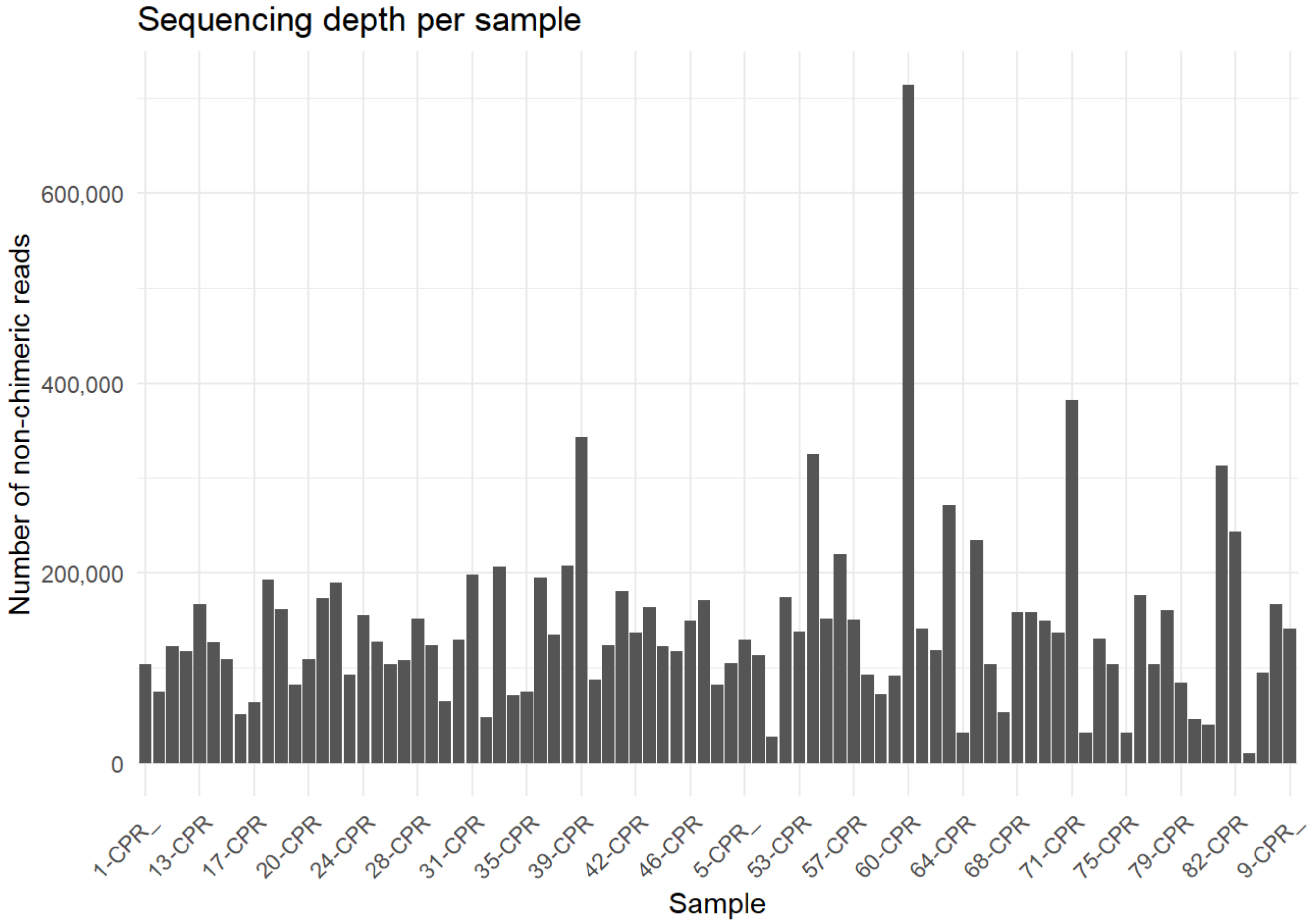
📦 Build the phyloseq object (COI)

```
🔹 ** A summary of the phyloseq object**

phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 18037 taxa and 85 samples ]
sample_data() Sample Data:       [ 85 samples by 8 sample variables ]
tax_table()   Taxonomy Table:    [ 18037 taxa by 20 taxonomic ranks ]

🔹 **These are the metadata variables:**

[1] "cpr"      "ID"       "Sample"   "Number"   "Inst"     "lat"      "lon"
[8] "Position"
```

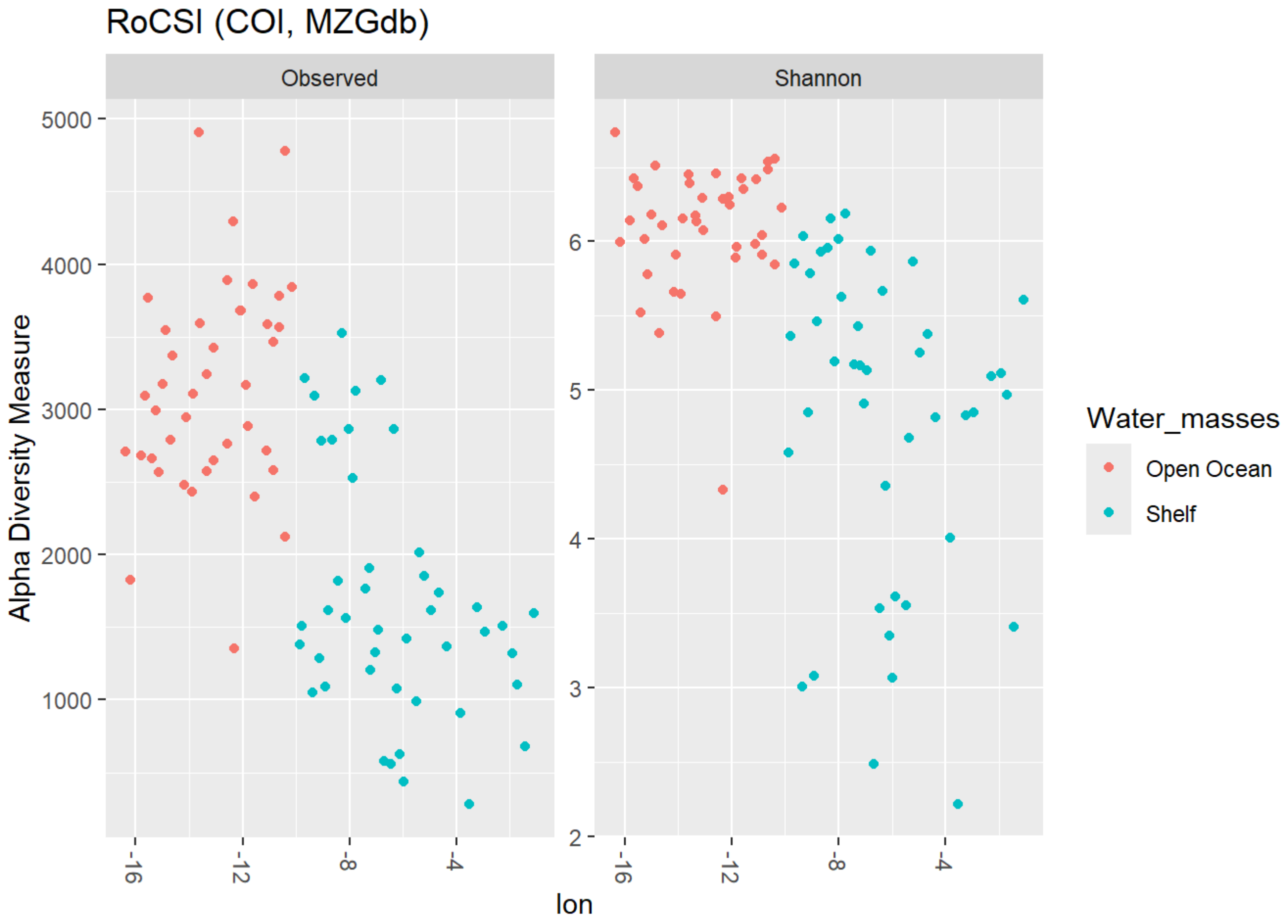
📊 Data visualisation

In this section, we explore the RoCSI-CPR 18S community structure using the `phyloseq` object (`ps`) generated above. We start by visualising read depth per sample, followed by basic taxonomic composition summaries.

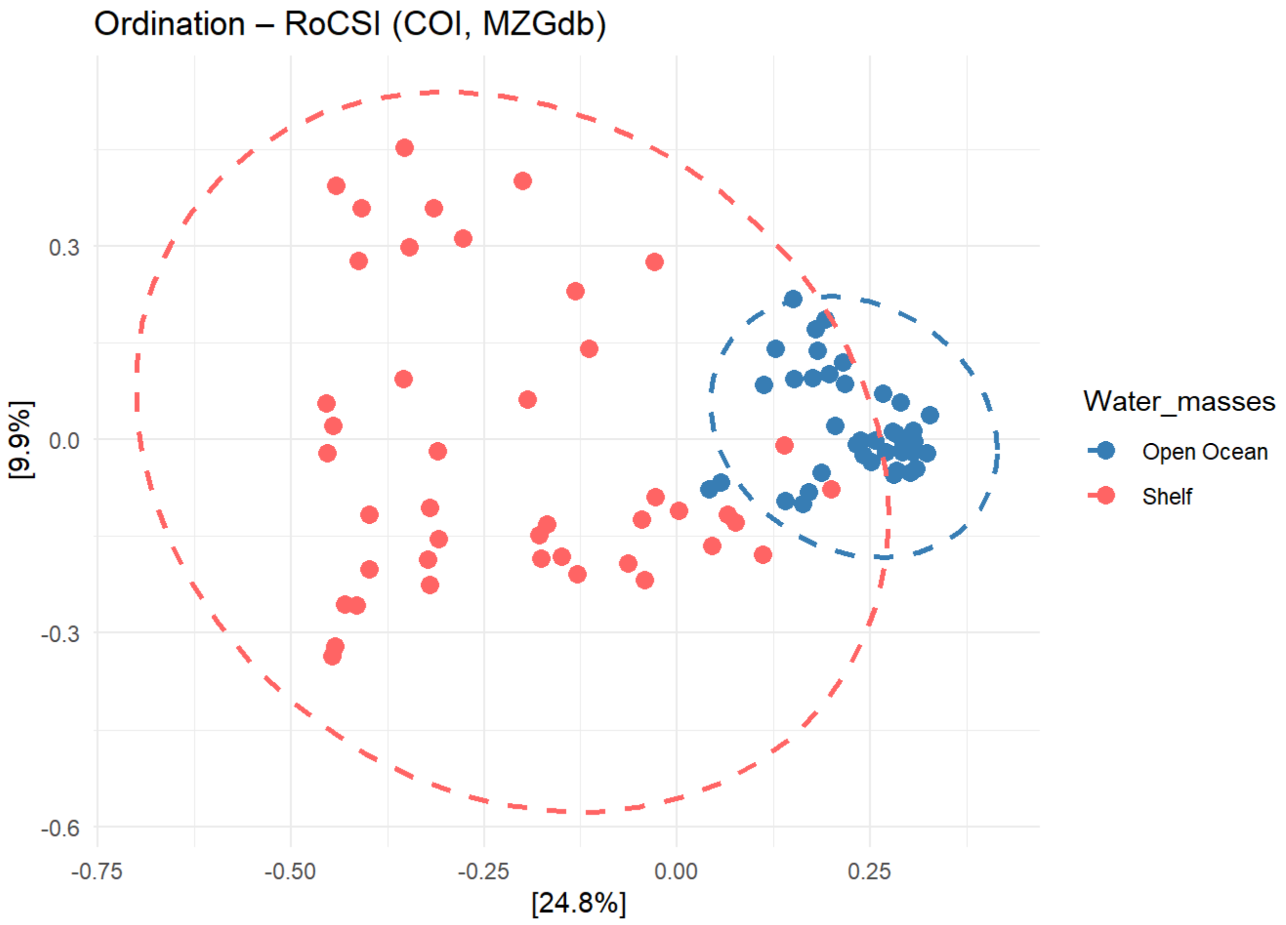
Read depth per sample



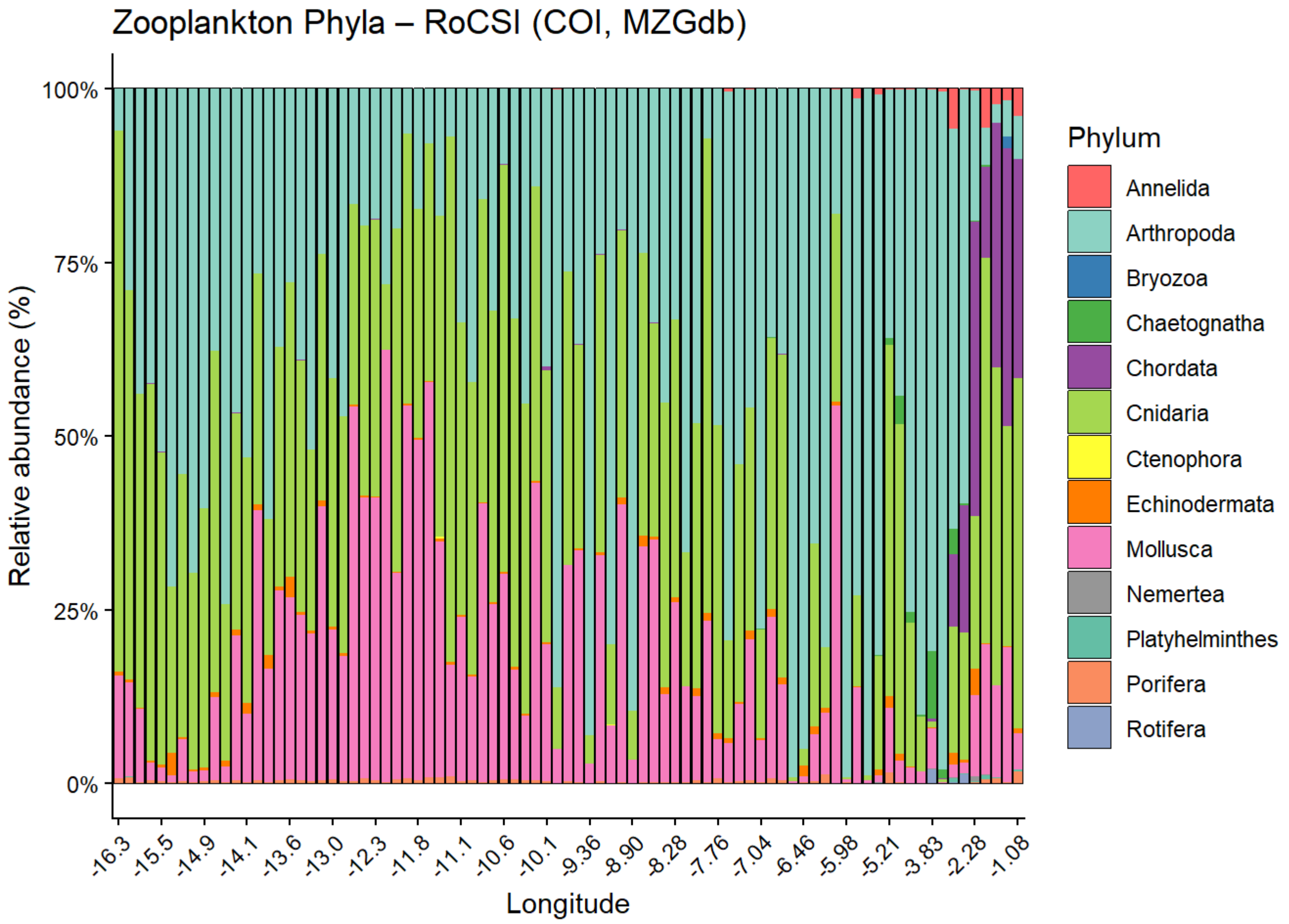
Alpha diversity


RoCSI (COI, MZGdb)
Observed
Shannon
Alpha Diversity Measure
5000
4000
3000
2000
1000
6
5
4
3
2
-16
-12
-8
-4
lon
Water_masses
Open Ocean
Shelf

Ordination


Ordination – RoCSI (COI, MZGdb)
[9.9%]
0.3
0.0
-0.3
-0.6
-0.75
-0.50
-0.25
0.00
0.25
[24.8%]
Water_masses
Open Ocean
Shelf

Phylum composition


Zooplankton Phyla – RoCSI (COI, MZGdb)
Relative abundance (%)
Longitude
Phylum
Annelida
Arthropoda
Bryozoa
Chaetognatha
Chordata
Cnidaria
Ctenophora
Echinodermata
Mollusca
Nemertea
Platyhelminthes
Porifera
Rotifera