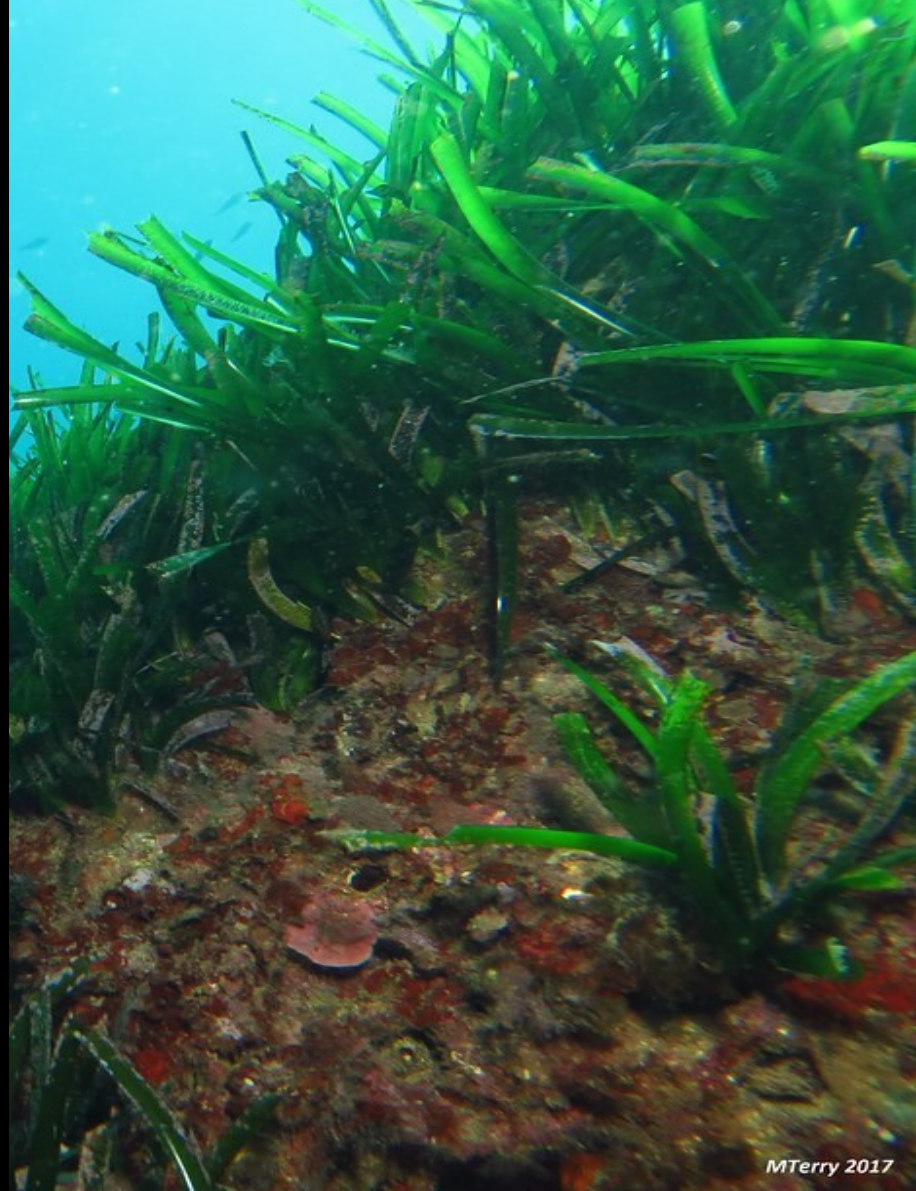
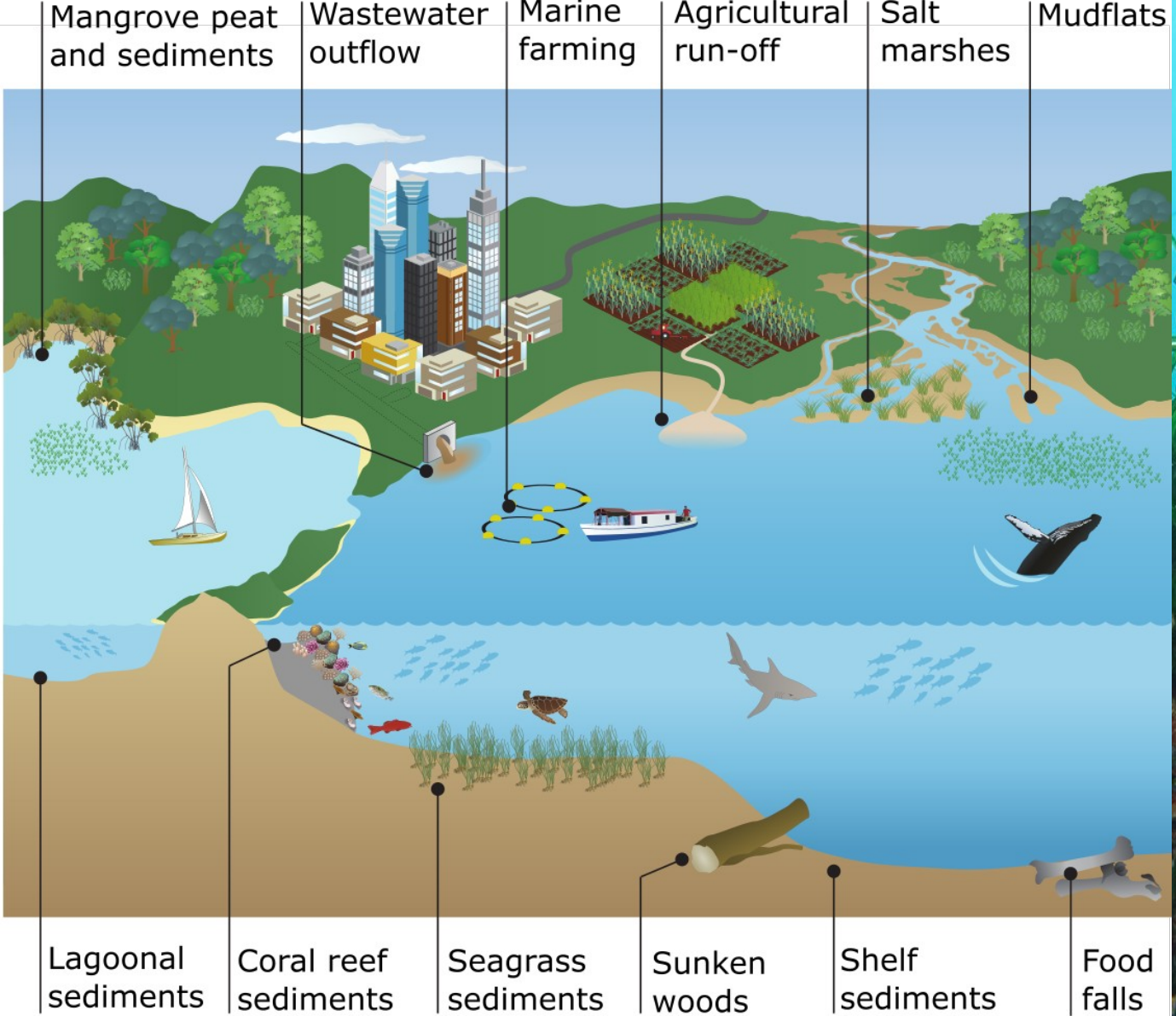
Building a network of invertebrate-microbe associations from seagrass beds

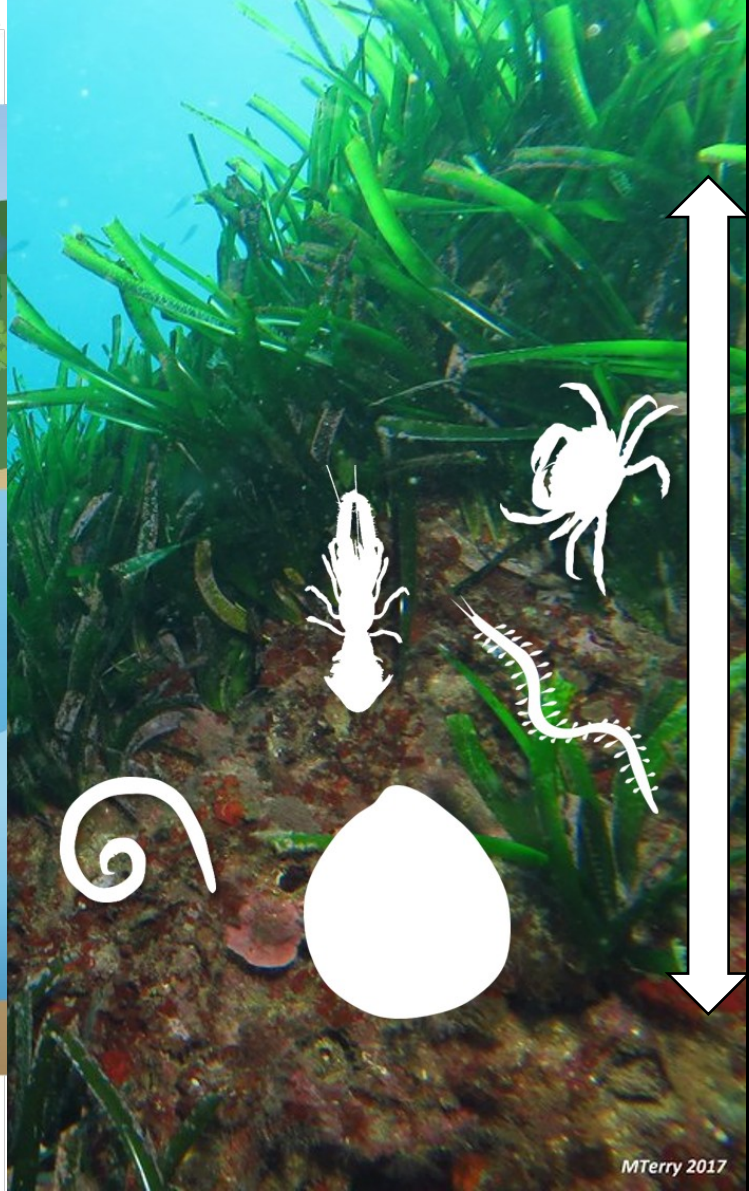
Luis Montilla

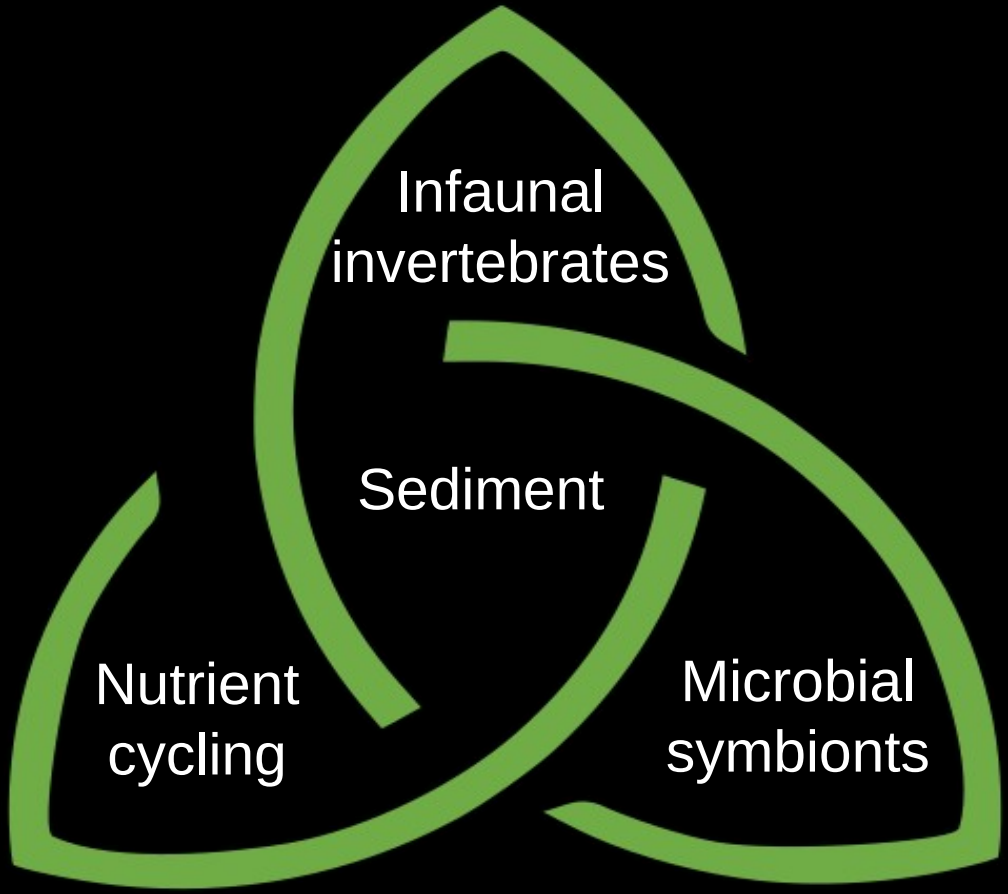
Roberta Piredda

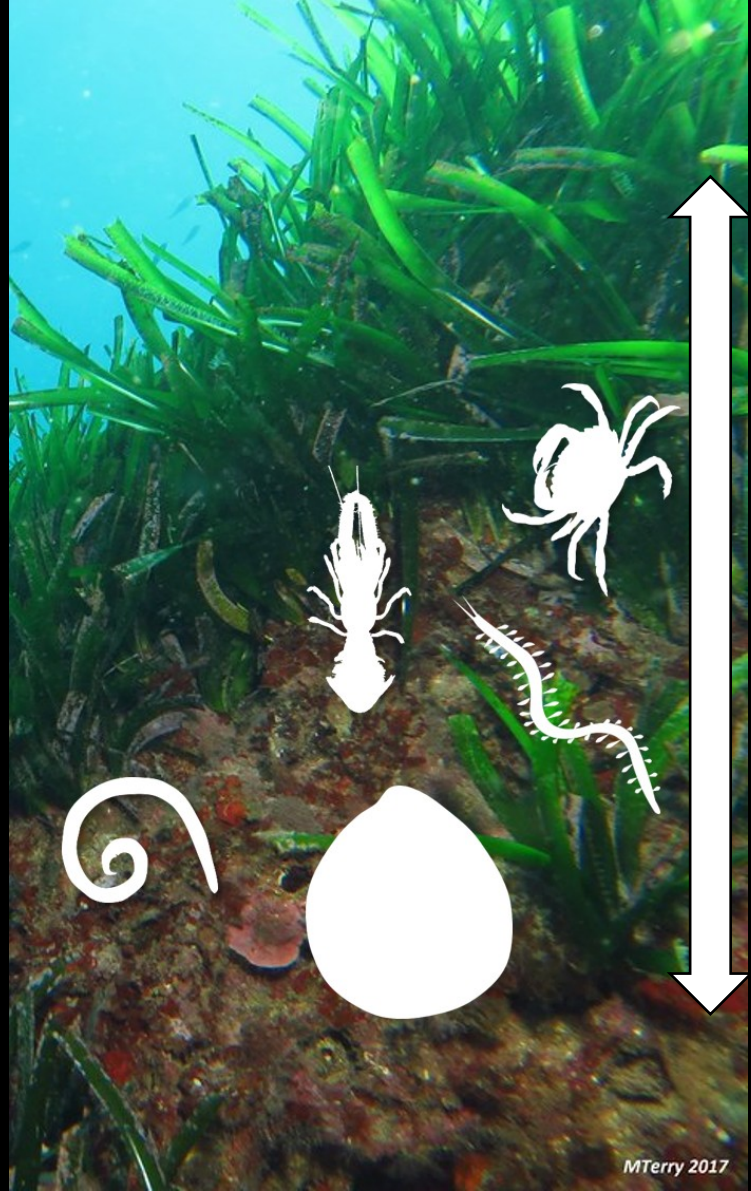
Ulisse Cardini

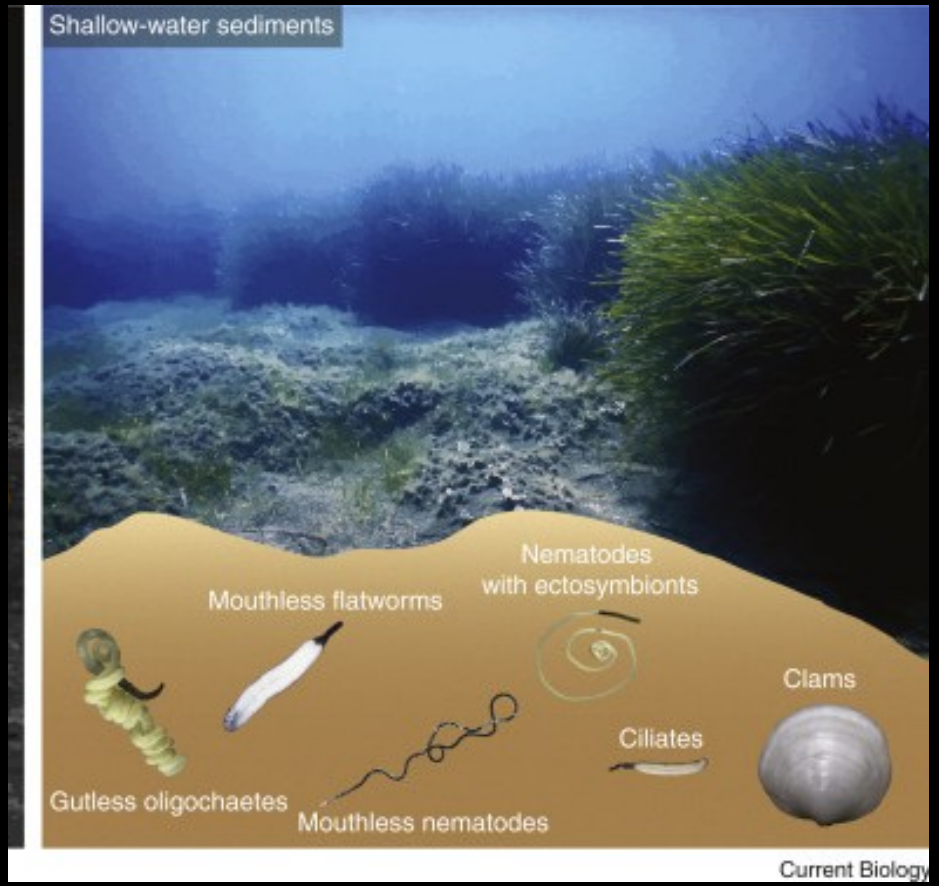



Mangrove peat and sediments
Wastewater outflow
Marine farming
Agricultural run-off
Salt marshes
Mudflats
Lagoonal sediments
Coral reef sediments
Seagrass sediments
Sunken woods
Shelf sediments
Food falls


MTerry 2017

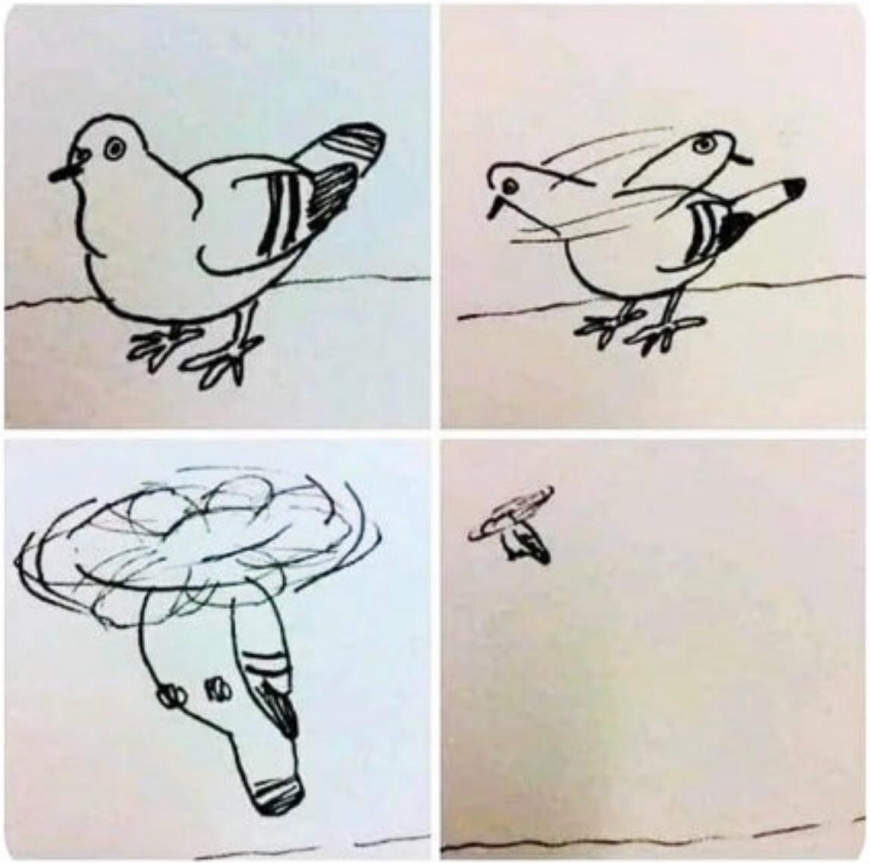

Infaunal
invertebrates
Sediment
Nutrient
cycling
Microbial
symbionts


MTerry 2017



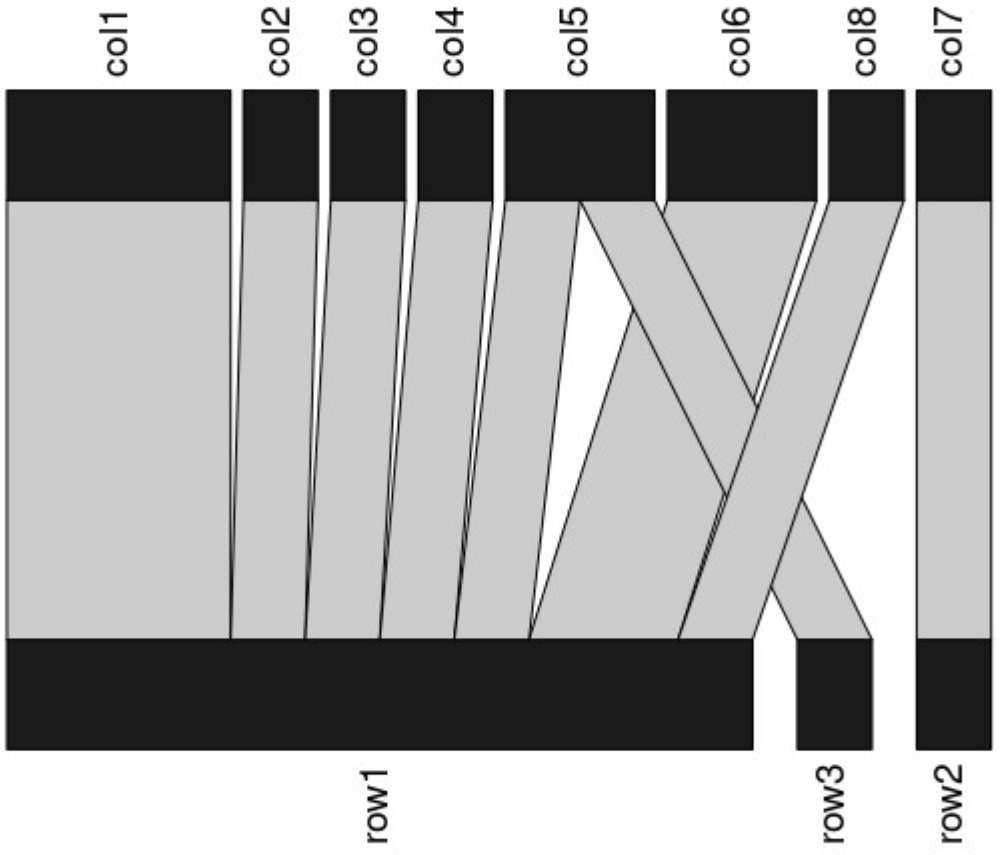
What patterns would we see in a more extensive data set?

When your program is a complete mess, but it does its job

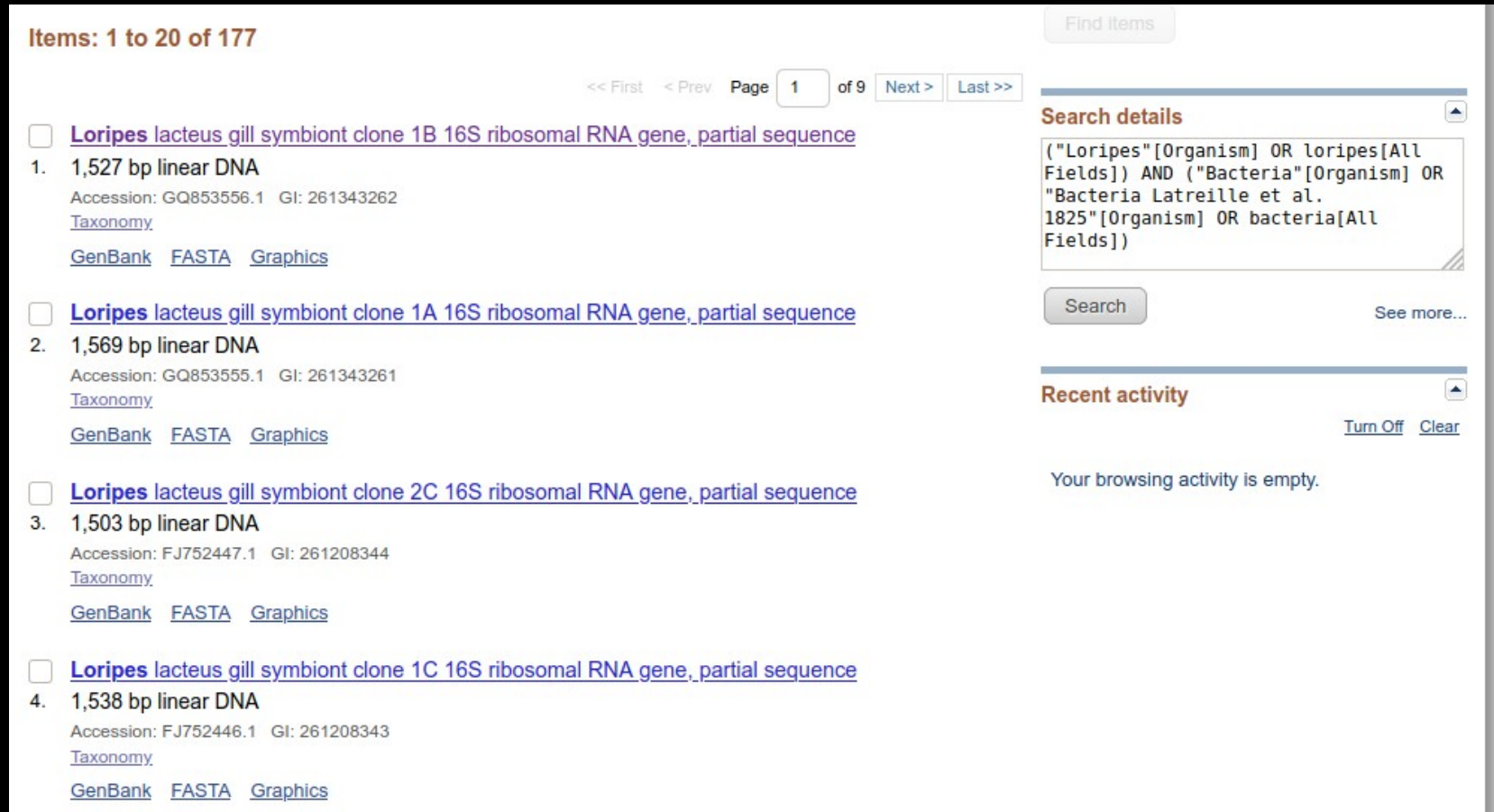


Network of associations between seagrass macrofauna and microbial symbionts

The first ingredient: a list of 291 dominant invertebrates from Mediterranean seagrasses



Search query:Genus + one group of symbionts (e.g. Bacteria)



Items: 1 to 20 of 177

<< First < Prev Page 1 of 9 Next > Last >>

Find items

1. **Loripes** lacteus gill symbiont clone 1B 16S ribosomal RNA gene, partial sequence
   1,527 bp linear DNA
   Accession: GQ853556.1 GI: 261343262
   Taxonomy
   GenBank FASTA Graphics

2. **Loripes** lacteus gill symbiont clone 1A 16S ribosomal RNA gene, partial sequence
   1,569 bp linear DNA
   Accession: GQ853555.1 GI: 261343261
   Taxonomy
   GenBank FASTA Graphics

3. **Loripes** lacteus gill symbiont clone 2C 16S ribosomal RNA gene, partial sequence
   1,503 bp linear DNA
   Accession: FJ752447.1 GI: 261208344
   Taxonomy
   GenBank FASTA Graphics

4. **Loripes** lacteus gill symbiont clone 1C 16S ribosomal RNA gene, partial sequence
   1,538 bp linear DNA
   Accession: FJ752446.1 GI: 261208343
   Taxonomy
   GenBank FASTA Graphics

Search details

("Loripes"[Organism] OR loripes[All Fields]) AND ("Bacteria"[Organism] OR "Bacteria Latreille et al. 1825"[Organism] OR bacteria[All Fields])

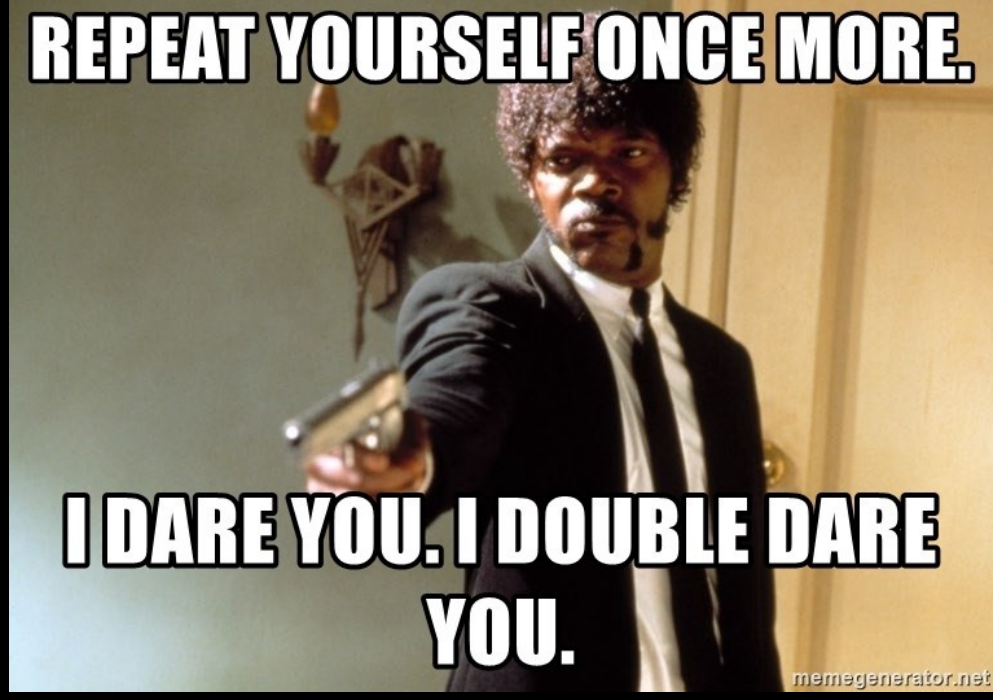
Search See more...

Recent activity

Turn Off Clear

Your browsing activity is empty.

Manually download the records for 291 genera of invertebrates (?)



rOpenSci: The *rentrez* package

`rentrez` provides functions that work with the NCBI Eutils API to search, download data from, and otherwise interact with NCBI databases.

The first setback:

We needed a specific qualifier

GenBank

Loripes lacteus gill symbiont clone 1B 16S ribosomal RNA gene,

GenBank: GQ853556.1

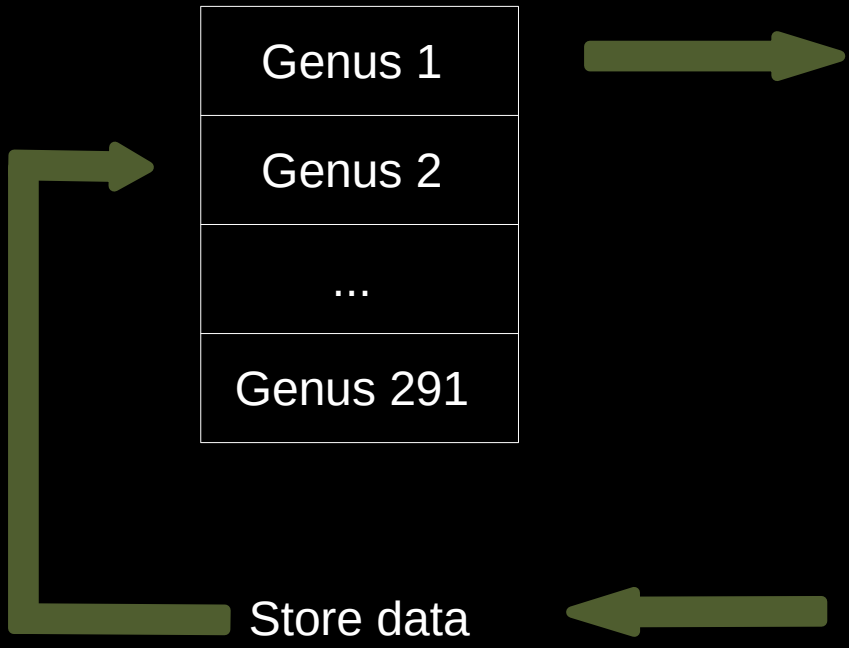
FASTA Graphics

Go to:

```
LOCUS       GQ853556               1527 bp    DNA     linear   BCT 31-AUG-2010
DEFINITION  Loripes lacteus gill symbiont clone 1B 16S ribosomal RNA gene,
            partial sequence.
ACCESSION   GQ853556
VERSION     GQ853556.1
KEYWORDS    .
SOURCE      Loripes lacteus gill symbiont
  ORGANISM  Loripes lacteus gill symbiont
            Bacteria; Proteobacteria; Gammaproteobacteria; sulfur-oxidizing
            symbionts.
REFERENCE   1  (bases 1 to 1527)
  AUTHORS   Mausz,M., Schmitz-Esser,S. and Steiner,G.
  TITLE     Identification and comparative analysis of the endosymbionts of
            Loripes lacteus and Anodontia fragilis (Bivalvia: Lucinidae)
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1527)
  AUTHORS   Mausz,M., Schmitz-Esser,S. and Steiner,G.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-AUG-2009) Department for Evolutionary Biology,
            University of Vienna, Institute for Organismal Systems Biology,
            Althanstrasse 14, Vienna A-1492, Austria
FEATURES             Location/Qualifiers
     source          1..1527
                     /organism="Loripes lacteus gill symbiont"
                     /mol_type="genomic DNA"
                     /host="Loripes lacteus isolate 52"
                     /db_xref="taxon:603670"
                     /clone="1B"
     rRNA            <1..>1527
                     /product="16S ribosomal RNA"
ORIGIN
        1 agagtttgat catggctcag attgaacgct ggcggaggcc taacacatgc aagtcgagcg
       61 gtaacagggg gagcttgctc tctgctgacg agcggcggac gggtgcgtaa cacgtaggaa
```

Retrieve genbank files

Write a loop - > Repeat this step for each 291 genera x 5 groups (bacteria, archaea, virus, fungi, eukaryota)




```
> gb_file
Object of class ‘efetch’
LOCUS       GQ853556                1527 bp    DNA     linear   BCT 31-AUG-2010
DEFINITION  Loripes lacteus gill symbiont clone 1B 16S ribosomal RNA gene,
            partial sequence.
ACCESSION   GQ853556
VERSION     GQ853556.1
KEYWORDS    .
SOURCE      Loripes lacteus gill symbiont
  ORGANISM  Loripes lacteus gill symbiont
            Bacteria; Proteobacteria; Gammaproteobacteria; sulfur-oxidizing
            symbionts.
REFERENCE   1  (bases 1 to 1527)
  AUTHORS   Mausz,M., Schmitz-Esser,S. and Steiner,G.
...
EFetch query using the ‘nuccore’ database.
Query url: ‘https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?=efetch...’
Retrieval type: ‘gbwithparts’, retrieval mode: ‘text’

>
```

Retrieve genbank files

Identify objects with 0 entries, and objects with >900 entries

210 genera didn't return any results, leaving us with only 28% of the original list.

Retrieve genbank files

Write files to disk

Also, manually downloading files with >900 entries

The specific element we wanted isn't easily retrievable

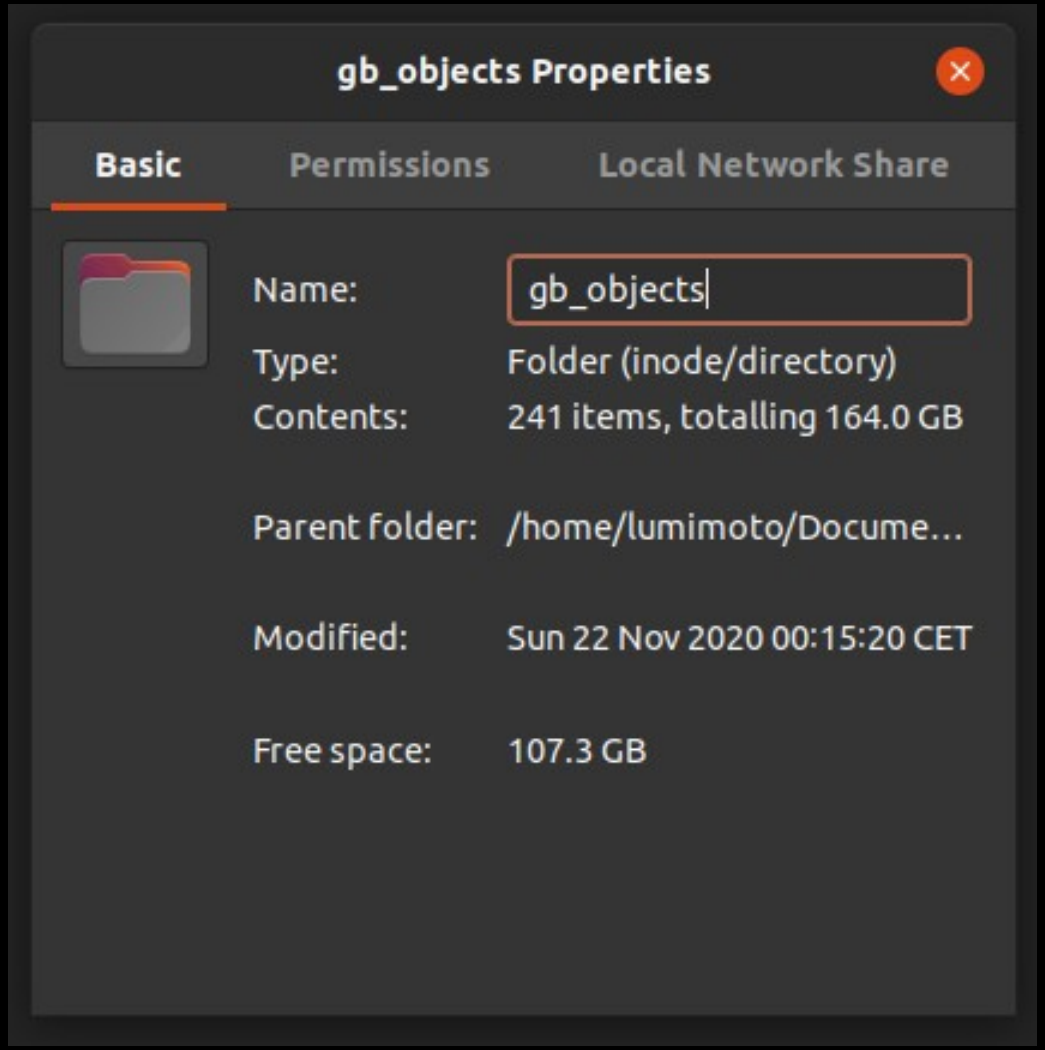
Also, the gb objects include genetic sequences

(maybe using regex)

```
> gb_file$content
[1] "LOCUS       GQ853556               1527 bp    DNA     linear   BCT 31-AUG-2010\nDEFINITION  Loripes lacteus gill symbiont clone 1B 16S ribosomal
 RNA gene,\n            partial sequence.\nACCESSION   GQ853556\nVERSION     GQ853556.1\nKEYWORDS    .\nSOURCE      Loripes lacteus gill symbiont\n  OR
GANISM  Loripes lacteus gill symbiont\n            Bacteria; Proteobacteria; Gammaproteobacteria; sulfur-oxidizing\n            symbionts.\nREFERENCE
 1  (bases 1 to 1527)\n  AUTHORS   Mausz,M., Schmitz-Esser,S. and Steiner,G.\n  TITLE     Identification and comparative analysis of the endosymbionts
 of\n            Loripes lacteus and Anodontia fragilis (Bivalvia: Lucinidae)\n  JOURNAL   Unpublished\nREFERENCE   2  (bases 1 to 1527)\n  AUTHORS   M
ausz,M., Schmitz-Esser,S. and Steiner,G.\n  TITLE     Direct Submission\n  JOURNAL   Submitted (26-AUG-2009) Department for Evolutionary Biology,\n
       University of Vienna, Institute for Organismal Systems Biology,\n            Althanstrasse 14, Vienna A-3492, Austria\nFEATURES             Loc
ation/Qualifiers\n     source          1..1527\n                     /organism=\"Loripes lacteus gill symbiont\"\n                     /mol_type=\"geno
mic DNA\"\n                     /host=\"Loripes lacteus isolate 52\"\n                     /db_xref=\"taxon:682670\"\n                     /clone=\"1B
\"\n     rRNA            <1..>1527\n                     /product=\"16S ribosomal RNA\"\nORIGIN      \n        1 agagtttgat catggctcag attgaacgct ggcgg
aggcc taacacatgc aagtcgagcg\n       61 gtaacagggg gagcttgctc tctgctgacg agcggcggac gggtgcgtaa cacgtaggaa\n      121 tctgcccggt agtgggggat agcccggaga aa
tccggatt aataccgcat acgccccaag\n      181 ggggaaagca ggggatccgt tcttttcgga gagtggacct tgcgctattg gatgagcctg\n      241 cgtcggatta gcttgttggt ggggtaaagg
 cctaccaagg caacgatccg tagctggtct\n      301 gagaggatga tcagccacac tgggactgag acacggccca gactcctacg ggaggcagca\n      361 gtggggaata ttgcacaatg ggggaaa
ccc tgatgcagcc atgccgcgtg tgtgaagaag\n      421 gctctagggt tgtaaagcac tttcagcgag gaggaaaggt tgttgattaa tactcaacag\n      481 ctgtgacgtt actcgcagaa gaag
caccgg ctaactccgt gccagcagcc gcggtaatac\n      541 ggagggtgcg agcgttaatc ggaattactg ggcgtaaagc gtgcgtaggc ggctgcctaa\n      601 gttggatgtg aaagccccgg g
cttaacctg ggaactgcat ccaaaactgg gcggctagag\n      661 tgcggaagag gagtgtggaa tttcctgtgt agcggtgaaa tgcgtagata taggaaggaa\n      721 caccagtggc gaaggcgac
a ctctggtctg acactgacgc tgaggtacga aagcgtgggg\n      781 agcaaacagg attagatacc ctggtagtcc acgccgtaaa cgatgtctac tagttgtccg\n      841 gagtcttgta ttctgg
gtaa cgaagctaac gcgataagta gaccgcctgg ggagtacggc\n      901 cgcaaggtta aaactcaaat gaattgacgg gggcccgcac aagcggtgga gcatgtggtt\n      961 taattcgaag caa
cgcgaag aaccttacct ggccttgaca tcctgcgaat cctttagaga\n     1021 tagaggagtg ccttcgggaa cgcagagaca ggtgctgcat ggctgtcgtc agctcgtgtc\n     1081 gtgagatgtt
 gggttaagtc ccgcaacgag cgcaaccctt gtcctcagtt accagcacgt\n     1141 tatggtgggc actctgggga gactgccggt gacaaaccgg aggaaggtgg ggacgacgtc\n     1201 aagtcat
cat ggcccttacg gccagggcta cacacgtgct acaatggtgc atacagacgg\n     1261 ttgccaagcc gcgaggtgga gctaatctga gaaagtgcat cgtagtccgg attggagtct\n     1321 gcaa
ctcgac tccatgaagt cggaatcgct agtaatcgtg aatcagaatg tcacggtgaa\n     1381 tacgttcccg ggccttgtac acaccgcccg tcacaccatg ggagtgggtt gctccagaag\n     1441 t
ggttagcct aacctctctt ttcgaagggg ggagggcgat caccacggag tgattcatga\n     1501 ctggggtgaa gtcgtaacaa ggtagcc\n//\n\n"
```

Retrieve genbank files

Write files to disk



Retrieve genbank files

Write files to disk

Merge & transform

```
> gb_file
Object of class 'efetch'
LOCUS       GQ853556               1527 bp    DNA     linear   BCT 31-AUG-2010
DEFINITION  Loripes lacteus gill symbiont clone 1B 16S ribosomal RNA gene,
            partial sequence.
ACCESSION   GQ853556
VERSION     GQ853556.1
KEYWORDS    .
SOURCE      Loripes lacteus gill symbiont
  ORGANISM  Loripes lacteus gill symbiont
            Bacteria; Proteobacteria; Gammaproteobacteria; sulfur-oxidizing
            symbionts.
REFERENCE   1  (bases 1 to 1527)
  AUTHORS   Mausz,M., Schmitz-Esser,S. and Steiner,G.
...
EFetch query using the 'nuccore' database.
Query url: 'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?=efetch...'
Retrieval type: 'gbwithparts', retrieval mode: 'text'

>
```

	Taxon 1	Taxon 2	Taxon 3
Sample 1			
Sample 2			

Retrieve genbank files

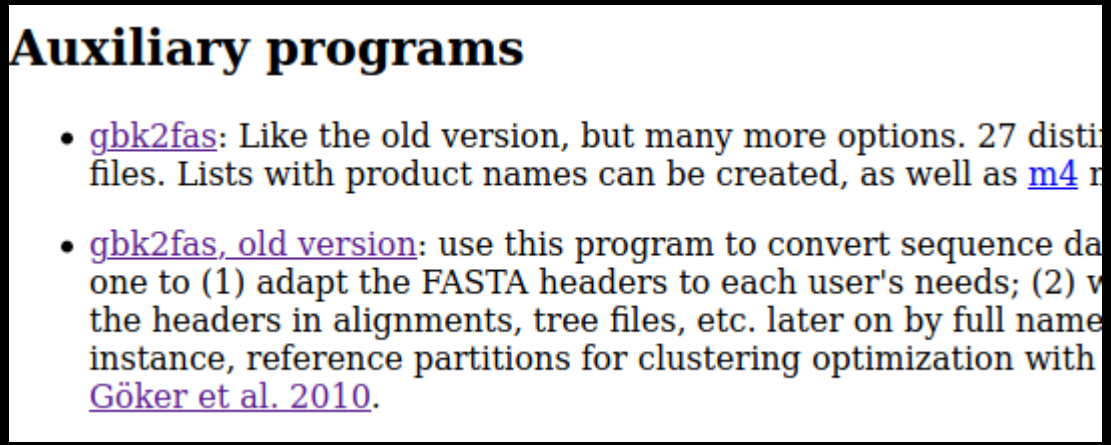
Write files to disk

Merge & transform

The easiest choice was to use a third-party software

Gbk2fas (Göker et al. 2010)

http://www.goeker.org/mg/clustering/



Auxiliary programs

- gbk2fas: Like the old version, but many more options. 27 disti
  files. Lists with product names can be created, as well as m4 r
- gbk2fas, old version: use this program to convert sequence da
  one to (1) adapt the FASTA headers to each user's needs; (2) v
  the headers in alignments, tree files, etc. later on by full name
  instance, reference partitions for clustering optimization with
  Göker et al. 2010.

Retrieve genbank files

Write files to disk

Merge & transform

Most recent version (Linux)

Old version (Linux)

New version (Windows)

Auxiliary programs

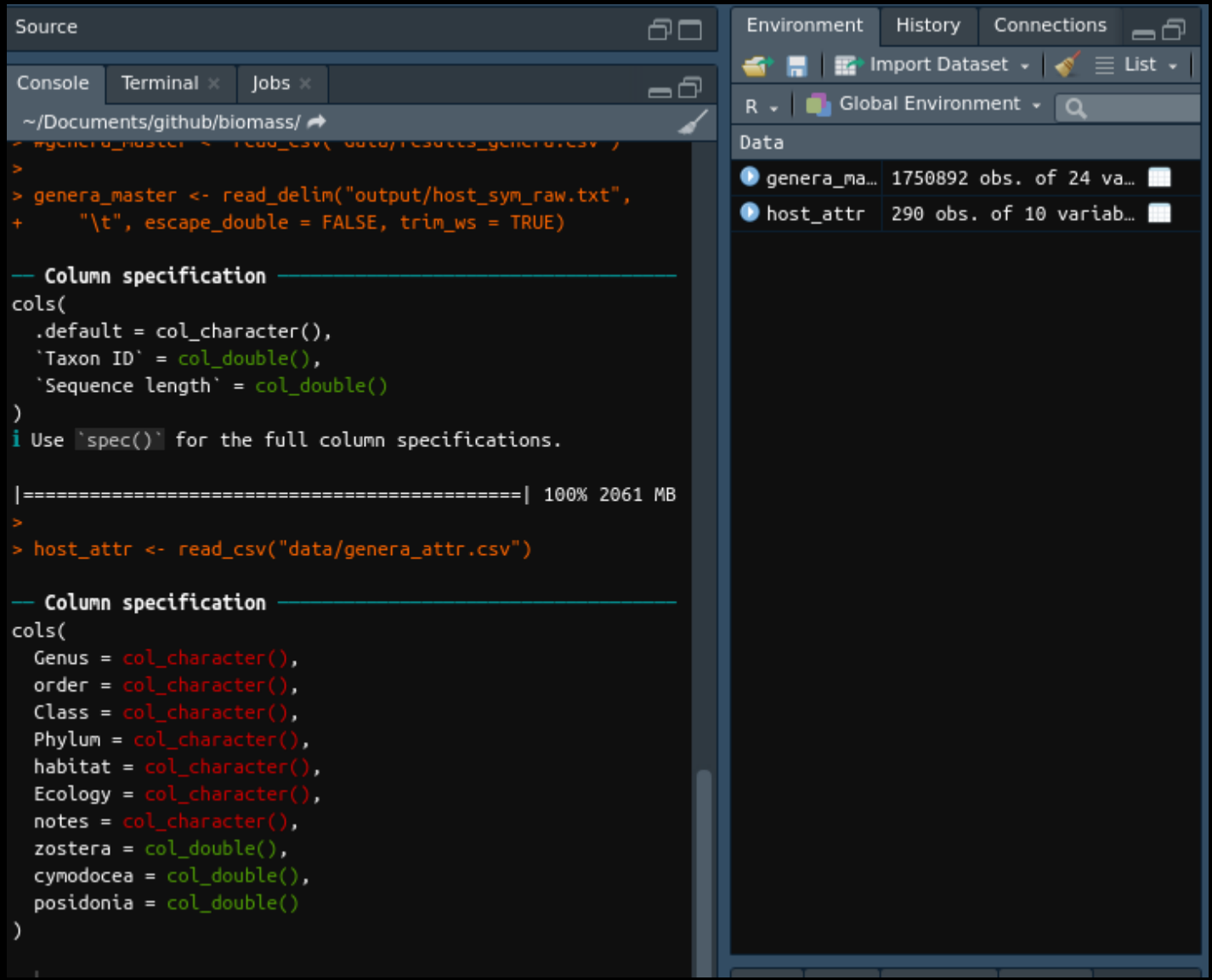
- gbk2fas: Like the old version, but many more options. 27 disti
  files. Lists with product names can be created, as well as m4 r
- gbk2fas, old version: use this program to convert sequence da
  one to (1) adapt the FASTA headers to each user's needs; (2) w
  the headers in alignments, tree files, etc. later on by full name
  instance, reference partitions for clustering optimization with
  Göker et al. 2010.

Retrieve genbank files

Write files to disk

Merge & transform

Reimport to R



Retrieve genbank files

Write files to disk

Merge & transform

Reimport to R

Data cleaning (hosts)

1) Non-marine organisms
2) Corals, sponges, algae

worrms

repo status Active | CRAN OK | R-check passing | codecov 84% | downloads 2854/month | CRAN 0.4.2

`worrms` is a R client for the World Register of Marine Species

- World Register of Marine Species (WoRMS) http://www.marinespecies.org/
- WoRMS REST API docs: http://www.marinespecies.org/rest/

See the taxize book (https://taxize.dev) for taxonomically focused work in this and similar packages.

“Host unknown” = plants and vertebrates

Retrieve genbank files

Write files to disk

Merge & transform

Reimport to R

Data cleaning (hosts)

Data cleaning (symbionts)

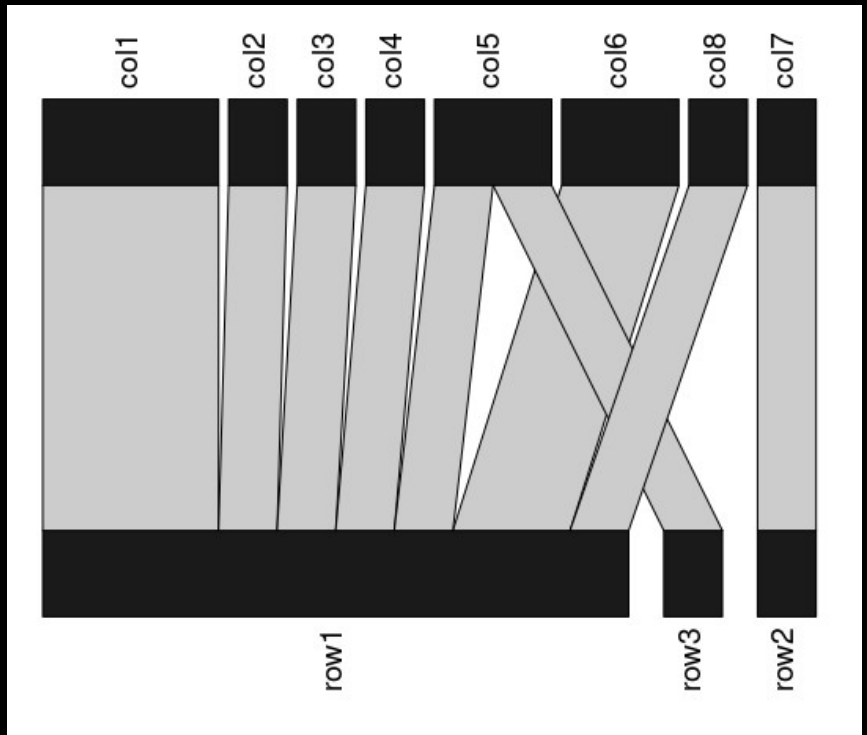
Lowest common taxonomic resolution

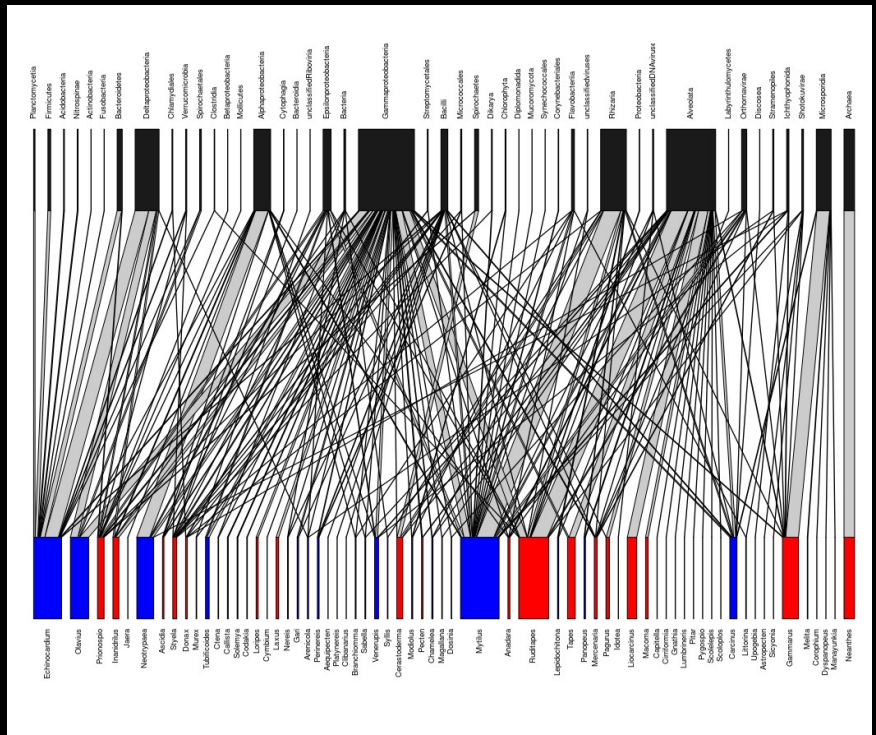
Raw file: 1,750,892 rows
Cleaned data set: 39,036 rows (2,2%)

igraph – The network analysis package

igraph is a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. igraph is open source and free. igraph can be programmed in R, Python, Mathematica and C/C++.

igraph R package | python-igraph | IGraph/M | igraph C library



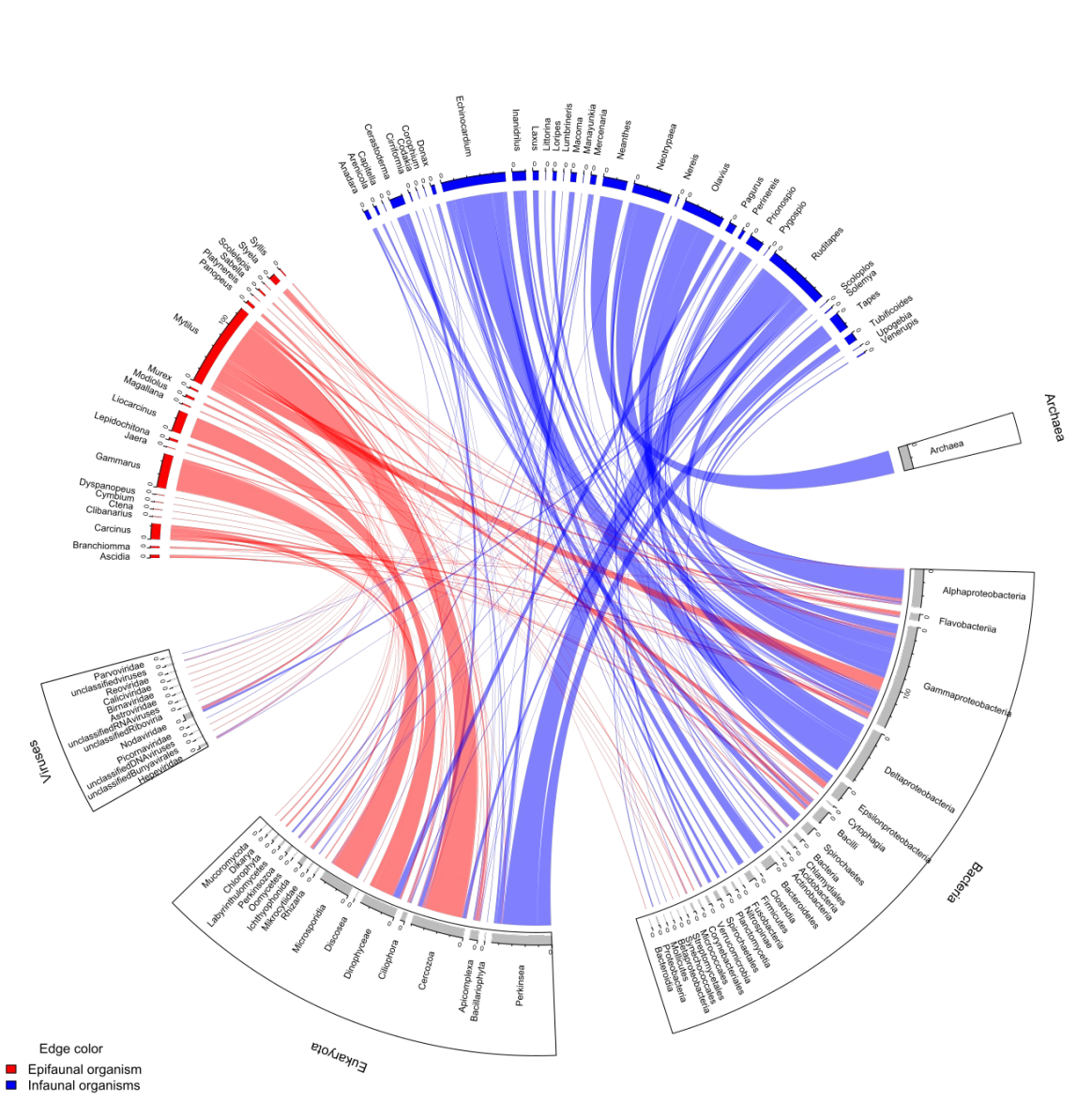


112 nodes	53 invertebrate genera	21 epifaunal
		32 infaunal
	59 microbial groups	28 Bacteria
		17 Eukaryota
		13 viruses
		1 Archea

circlize: circular visualization in R

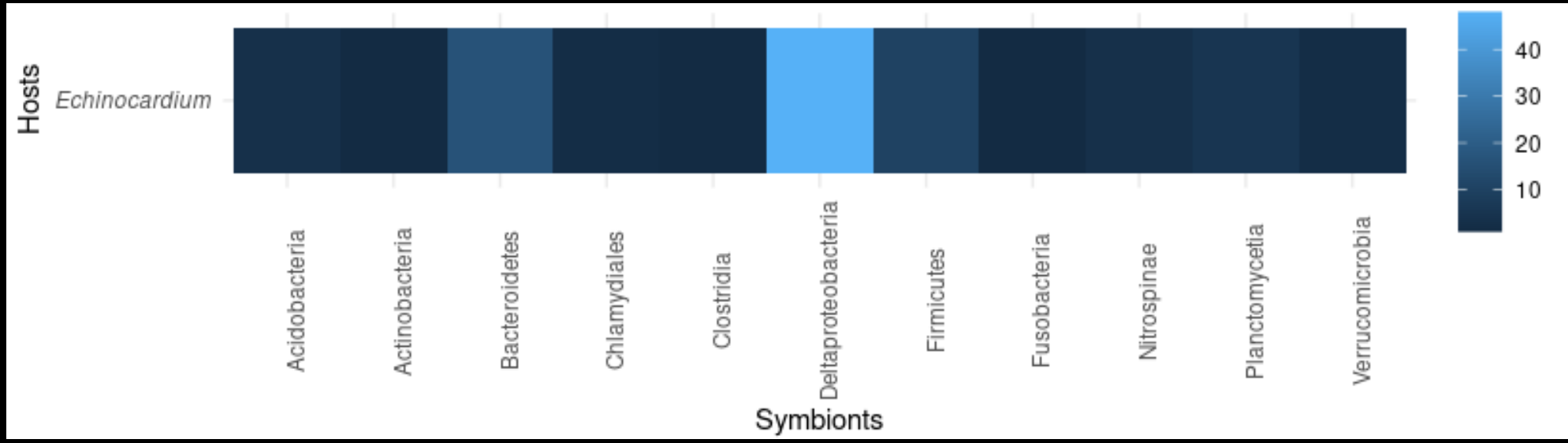
Circular layout is an efficient way for the visualization of huge amounts of information. Here the circlize package provides an implementation of circular layout generation in R as well as an enhancement of available software. The flexibility of this package is based on the usage of low-level graphics functions such that self-defined high-level graphics can be easily implemented by users for specific purposes. Together with the seamless connection between the powerful computational and visual environment in R, circlize gives users more convenience and freedom to design figures for better understanding complex patterns behind multi-dimensional data.

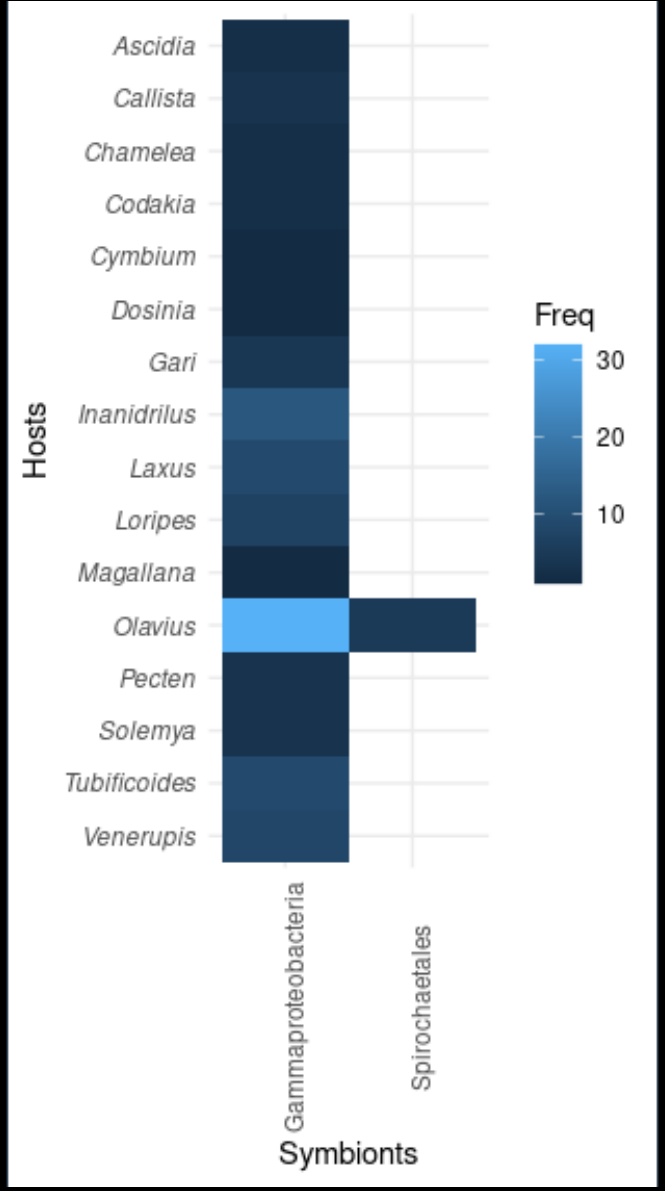
112 nodes	53 invertebrate genera	21 epifaunal
		32 infaunal
	59 microbial groups	28 Bacteria
		17 Eukaryota
		13 viruses
		1 Archea



Examples of obtained communities

© Hans Hillewaert




Ascidia
Callista
Chamelea
Codakia
Cymbium
Dosinia
Gari
Inanidrilus
Laxus
Loripes
Magallana
Olavius
Pecten
Solemya
Tubificoides
Venerupis
Hosts
Freq
30
20
10
Gammaproteobacteria
Spirochaetales
Symbionts

Next steps

Expand list to include macrofauna from other marine ecoregions

Add additional attributes if possible

Optimize code

build error | build failing | downloads 103/month | CRAN not published

biofiles - an interface to GenBank/GenPept files in R

biofiles provides interfacing to GenBank/GenPept or Embl flat file records. It includes utilities for reading and writing GenBank files, and methods for interacting with annotation and sequence data.

The Open University

Thanks for your attention

Questions?

Luismmontilla.com
luismiguel.montilla@szn.it