The POSEIDON Project: Protistian diversity as revealed by Ribosomal Barcoding along the TARA OCEANS' circumnavigation.

- Global Presentation of The Tara ocean's circumnavigation: Frederic Mahé
- Bioinformatics analyses of Barcode data: Stéphane Audic

Equipe “Evolution du Plancton et PaléOcéans”,
UMR7144 “Adaptation et Diversité en Milieu Marin”,
Station Biologique de Roscoff.

Frederic.Mahe@sb-roscoff.fr
Stephane.Audic@sb-roscoff.fr

CNRS UPMC INSU
Station Biologique
Roscoff

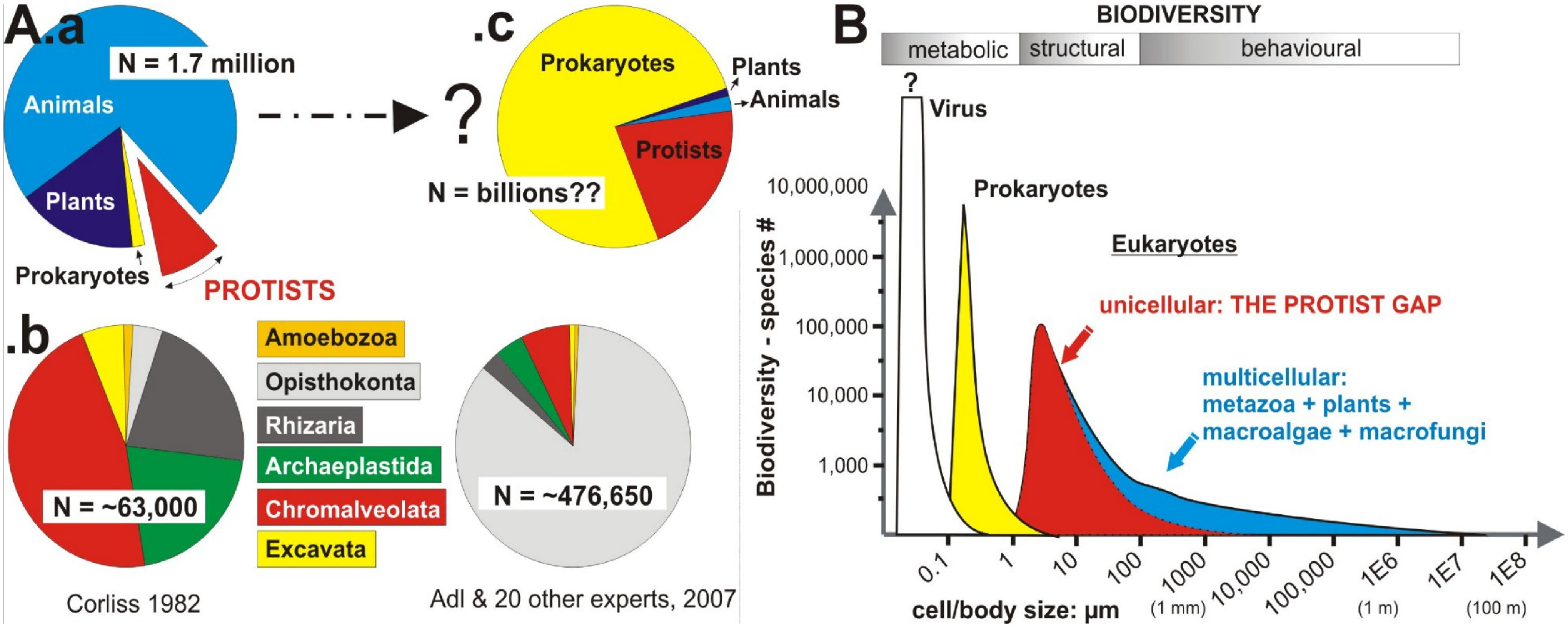


Figure 2: The marine protist biodiversity gap. A. a: Total number of living phenotypically described species on Earth (N=1.7 million; from International Union for Conservation of Nature). The relatively lower numbers of unicellular (pro- and eu-karyotic) species principally reflects the history of biological science, dominated by de visu and microscopy observations since the 1600s. b: 1982 (single author) and 2007 (consortium) surveys of described protist species and their classification among the higher level protist super-groups. These censuses are highly dependant on the number (and character!) of experts working on particular groups. The 2007 picture is for instance largely skewed toward fungi and diatoms; estimation of the "real" biodiversity by the same expert consortium, based on preliminary molecular datasets and gut-feeling, reached ~18 million species. Environmental (e)-rDNA clone library based studies over the last decade, have shown that eukaryotic microbial diversity is far greater than previously catalogued. Recent group-specific PCR of shorter e-DNA fragments (which reduces classical PCR biases) have further shown that protist biodiversity remains largely undersampled even using standard molecular approaches (Fig. 3). Overall, the molecular exploration of biodiversity tends to support the hypothetical distributions shown in Ac. and B. Hypothetical number of species (a species being defined as a distinct genome/phenotype couple at a given time) of viruses, prokaryotes, and eukaryotes in relation to cell/body size. The red area depicts the "protist gap", where most eukaryotic species still remain to be discovered, despite their fundamental biogeochemical importance and their key position as a bridge between prokaryote and animal/plant biodiversities.

Some key figures

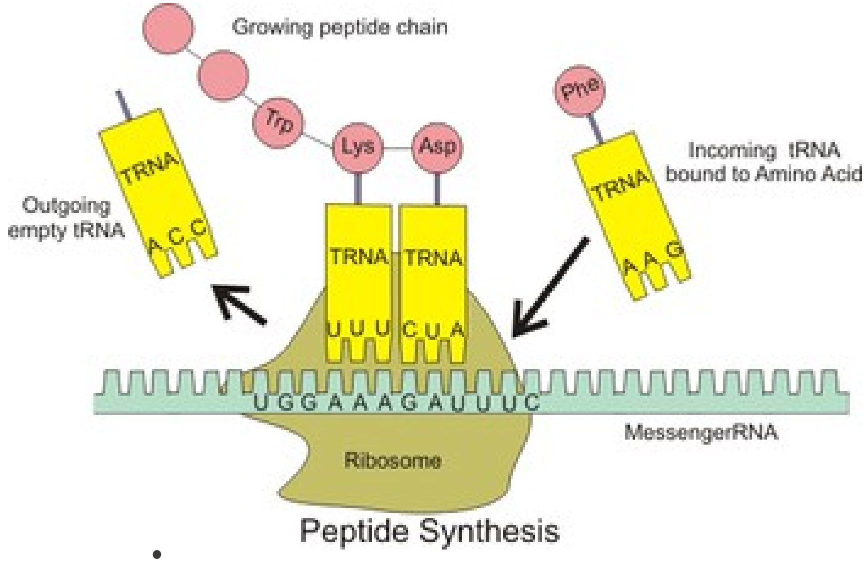
- 1 one ml of water, you have approximately
  - 10 millions viruses
  - 1 million prokaryots
  - 0.1 millions eukaryots
- In Tara, each samples is approximately 100 liters
  - 10^{12} viruses
  - 10^{11} prokaryots
  - 10^{10} eukaryots

DNA Barcoding for metagenomics

- Find a sequence region with very conserved parts and very variable parts: this is not a trivial problem, but this is a prerequisite because you want
    - To design primers that are very specific
    - But give you information ->variability
- Why not directly sequence all DNA? It depends of your problem.
    - If you just want to know who is here:
        - Most of it is of no use to you!
        - So you save at many levels: time/cost/..

What are good candidates for barcoding?

- It depends on what you are interested on:
  - All living organisms have ribosomes with which they do protein assembly
  - Among ribosome constituents ribosomal RNA is the more conserved between different species



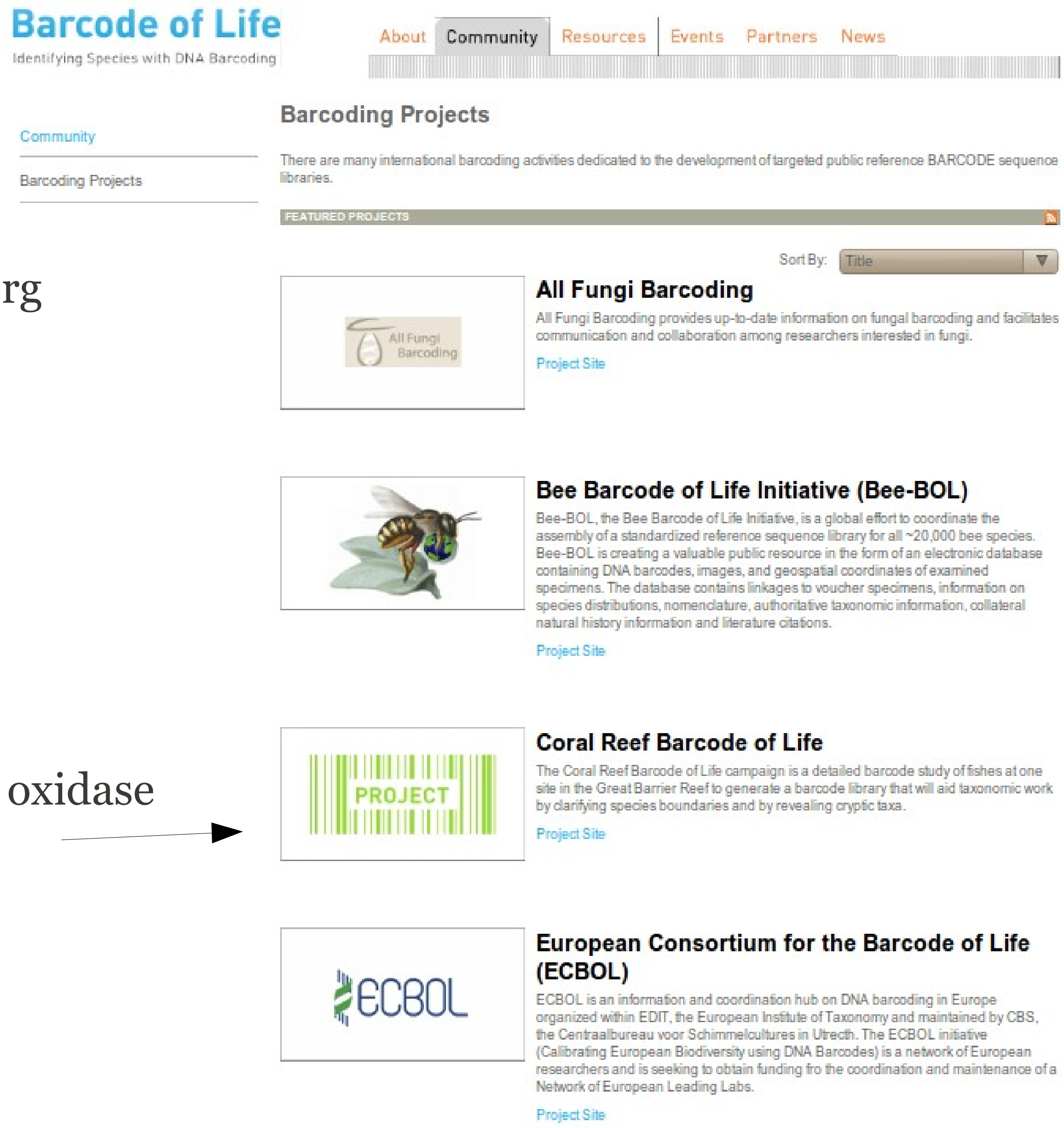
Peptide Synthesis

- More targeted studies in general use other markers:
  - See Barcode of Life project

Many barcoding initiatives

http://www.barcodeoflife.org

mitochondrial cytochome c oxidase
subunit I (CO1)



Barcoding Projects

There are many international barcoding activities dedicated to the development of targeted public reference BARCODE sequence libraries.

All Fungi Barcoding

All Fungi Barcoding provides up-to-date information on fungal barcoding and facilitates communication and collaboration among researchers interested in fungi.

Project Site

Bee Barcode of Life Initiative (Bee-BOL)

Bee-BOL, the Bee Barcode of Life Initiative, is a global effort to coordinate the assembly of a standardized reference sequence library for all ~20,000 bee species. Bee-BOL is creating a valuable public resource in the form of an electronic database containing DNA barcodes, images, and geospatial coordinates of examined specimens. The database contains linkages to voucher specimens, information on species distributions, nomenclature, authoritative taxonomic information, collateral natural history information and literature citations.

Project Site

Coral Reef Barcode of Life

The Coral Reef Barcode of Life campaign is a detailed barcode study of fishes at one site in the Great Barrier Reef to generate a barcode library that will aid taxonomic work by clarifying species boundaries and by revealing cryptic taxa.

Project Site

European Consortium for the Barcode of Life (ECBOL)

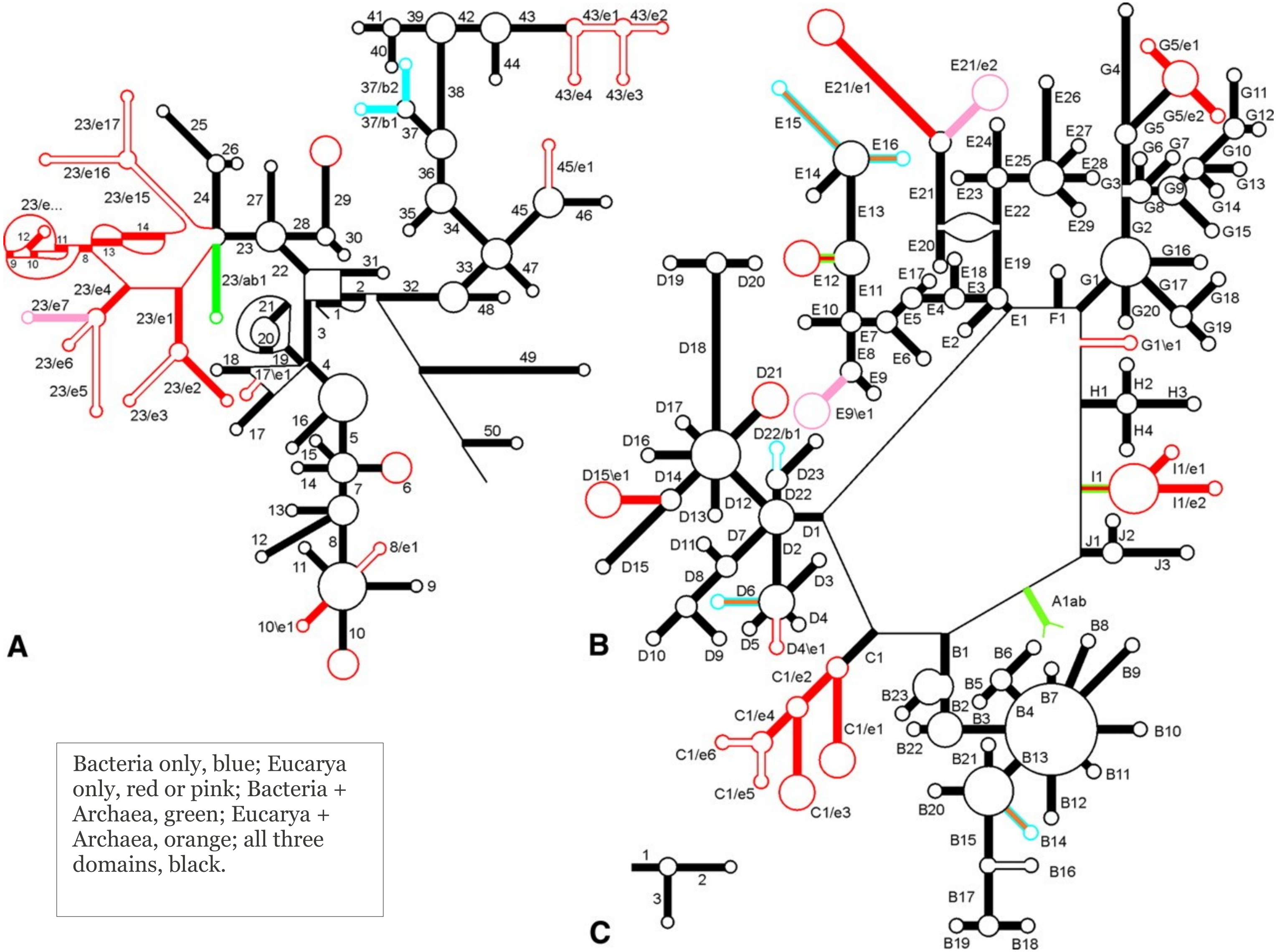
ECBOL is an information and coordination hub on DNA barcoding in Europe organized within EDIT, the European Institute of Taxonomy and maintained by CBS, the Centraalbureau voor Schimmelcultures in Utrech. The ECBOL initiative (Calibrating European Biodiversity using DNA Barcodes) is a network of European researchers and is seeking to obtain funding fro the coordination and maintenance of a Network of European Leading Labs.

Project Site

Potential problem with ribosomal RNA? May be too conserved

- Ribosomal RNA is very conserved even between Bacteria and Eukaryots.
    - Is there a region in Ribosomal RNA that can be used
    - Yes: V4 and V9 regions are relatively commonly present in Eukaryots and absent in Bacteria.

Figure 1. Secondary structure and helix numbering of SSU rRNA (A), LSU rRNA (B) and 5S rRNA (C).



Bacteria only, blue; Eucarya only, red or pink; Bacteria + Archaea, green; Eucarya + Archaea, orange; all three domains, black.

Wuyts J et al. Nucl. Acids Res. 2001;29:5017-5028

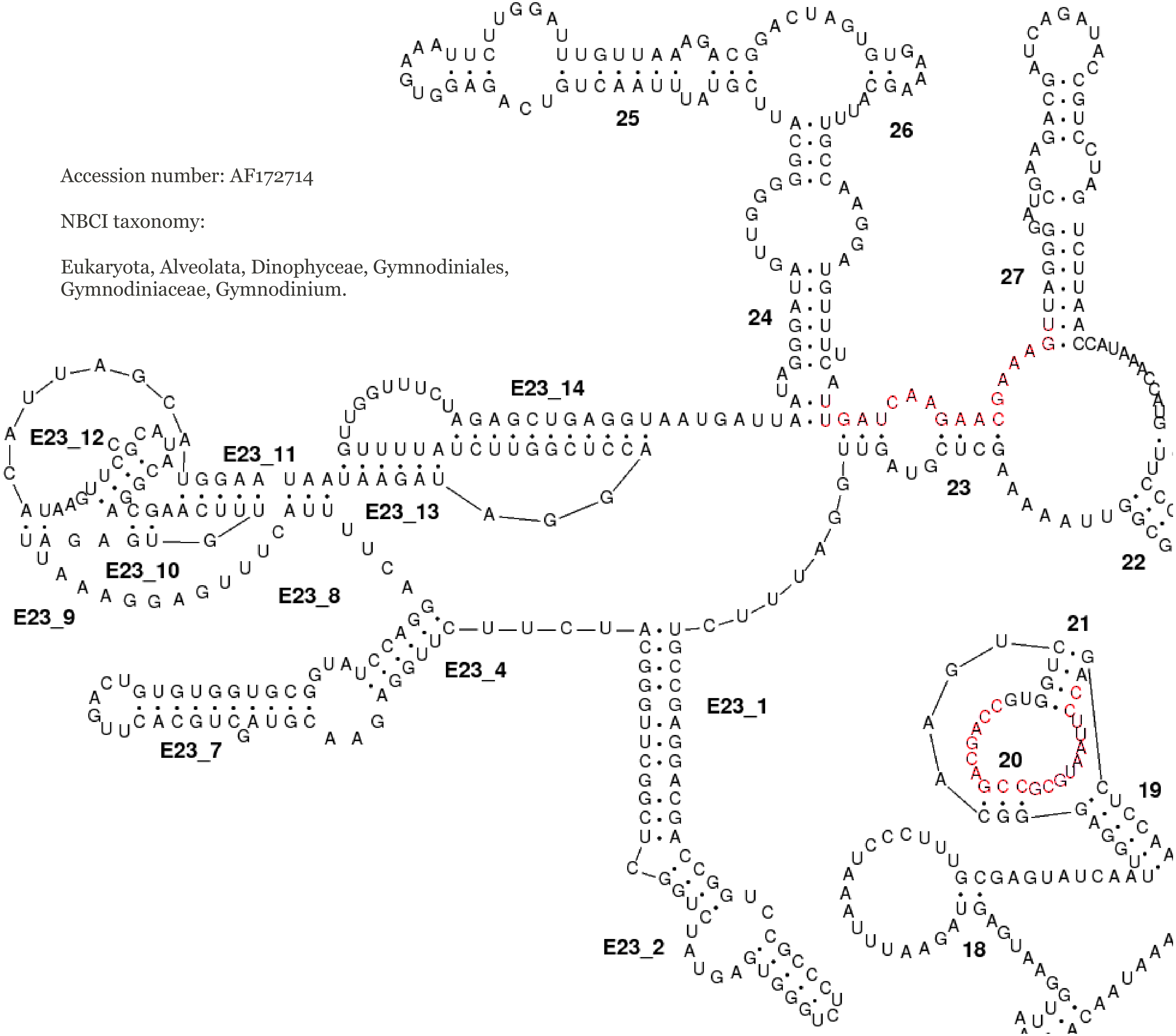
Nucleic Acids Research

Gymnodinium breve SSU rRNA secondary structure model

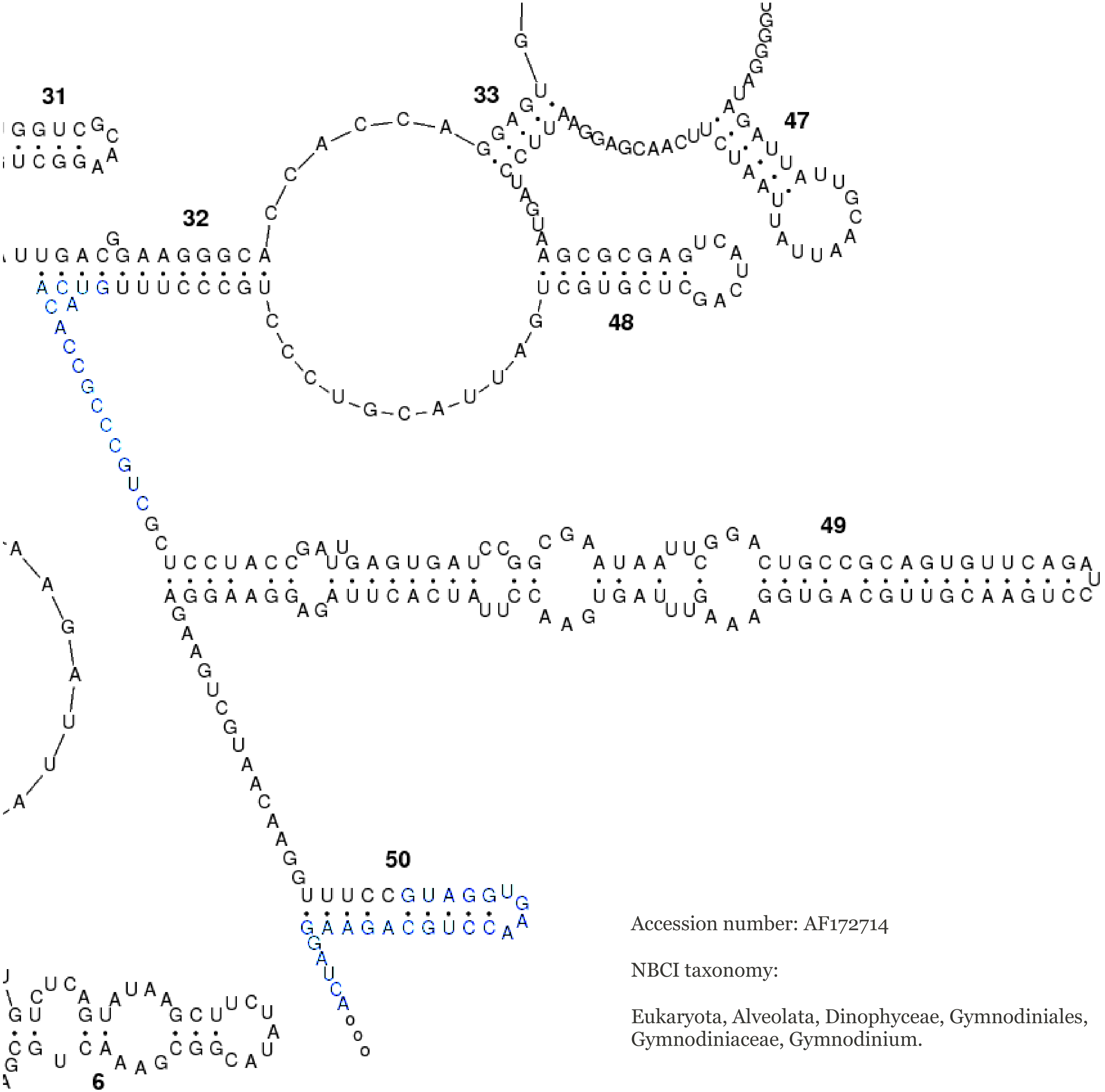
Accession number: AF172714

NBCI taxonomy:

Eukaryota, Alveolata, Dinophyceae, Gymnodiniales, Gymnodiniaceae, Gymnodinium.



Gymnodinium breve SSU rRNA secondary structure model



Accession number: AF172714

NBCI taxonomy:

Eukaryota, Alveolata, Dinophyceae, Gymnodiniales, Gymnodiniaceae, Gymnodinium.

Red: v4 region
Avg. 382 nt

Blue: v9 region
Avg. 126 nt

```
>ENA|AF172714|AF172714.1 Gymnodinium breve strain CCMP718
18S small subunit ribosomal RNA gene, partial sequence.
aacctggttgatcctgccagtagtcatatgcttgtctcaaagattaagccatgcatgtct
cagtataagcttctatacggcgaaactgcgaatggctcattaaaacagttatagtttatt
tgatggtcattcattacatggataactttggaaattctagagctaatacatgcgcccaaa
cccgacttcttggaagggttgtgtttattagatacagaaccaacccaggccctgccgtgt
cttgtggtgattcrtaataaccgaacgaatcgcattgcatcagctggcgatagatcattc
aagtttctgacctatcagtttccgacggtagggtattggcctaccgtggcattgacgggt
aacggagaattagggttcgattccggagagggagcctgagaaacggctaccacatctaag
gaaggcagcaggcgcgcaaattacccaatcctgacacagggaggtagtgacaagaaataa
caatacagggcatccatgtcttgtaattggaatgagtagaatttaaatccctttgcgagt
atcaattggagggcaagtctggtgccagcagccgcgtaattccagctccaatagcgtata
ttaaagttgttgcggttaaaaagctcgtagttggatttctgccgaggacgaccggtccgc
cctctgggtgagtatctggctcggcttgggcatcttcttggagaacgtagctgcacttga
ctgtgtggtgcggtatccaggacttttactttgaggaaattagagtgtttcaagcaggca
tacgccttgaatacattagcatggaataataagataggacctcggttctattttgttggt
ttctagagctgaggtaatgattaatagggatagttgggggcattcgtatttaactgtcag
aggtgaaattcttggatttgttaaagacggactagtgtgaaagcatttgccaaggatgtt
ttcattgatcaagaacgaaagttaggggatcgaagacgatcagataccgtcctagtctta
accataaaccatgccgactagagattggaggtcgttacttatacgactccttcagcacct
tatgagaaatcaaagtctttgggttccgggggagtatggtcgcaaggctgaaacttaaa
ggaattgacggaagggcaccaccaggagtggagcctgcgcttaatttgactcaacacggg
gaaacttaccaggtccagacatagtaaggattgacagattgatagctctttcttgattct
atgggtggtggtgcatggccgttcttagttggtggagtgatttgtctggttaattccgtt
aacgaacgagaccttaacctgctaaatagttacacgtaacttcggttacgtgggcaactt
cttagagggactttgcgtgtctaacgcaaggaagtttgaggcaatagcaggtctgtgatg
cccttagatgttctgggctgcacgcgcgctacactgatgcgctcaacgagttttcgacct
tgcccggaaggttgggtaatctttttaaacgcatcgtgatggggatagattattgcaat
tattaatcttcaacgaggaattcctagtaagcgcgagtcatcagctcgtgctgattacgt
ccctgccctttgtacacaccgcccgtcgctcctaccgattgagtgatccggcgaataatt
cggactgccgcagtgttcagatcctgaacgttgcagtggaaagtttagtgaaccttatca
cttagaggaaggagaagtcgtaacaaggtttccgtaggtgaacctgcagaaggatca
```

Some statistics on Silva

silva
comprehensive ribosomal RNA databases

http://www.arb-silva.de/

V4: CCAGCA[GC]C[CT]GCGGTAATTCC T[CT][AG]ATCAAGAACGAAAGT

Up to 4 mismatches per primer:

DB 262092 seq. // 219921 Bacteria // 10302 Archae // 31809 Eukaryota

->> 29303 seq. // 944 Bacteria // 16 Archae // 28296 Eukaryota

V9: GTACACACCGCCCGTC GTAGGTGAACCTGCAGAAGGATCA

->> 11632 seq. // 5447 Bacteria // 144 Archae // 6025 Eukaryota

Number of mismatches on SSU RNA

V9

Bacteria

	15	16	20	21	22	23
12	0	0	6	1	0	0
13	0	0	6	1	0	0
14	0	0	10	0	0	0
15	6	0	227	42	1	0
16	0	46	5054	35	0	2

Eukaryota

	15	16	20	21	22	23	24
12	0	0	0	1	3	15	3
13	0	0	2	5	2	7	3
14	0	0	1	10	5	7	3
15	6	0	6	3	17	70	49
16	0	159	295	156	370	2729	2084

V4

Bacteria

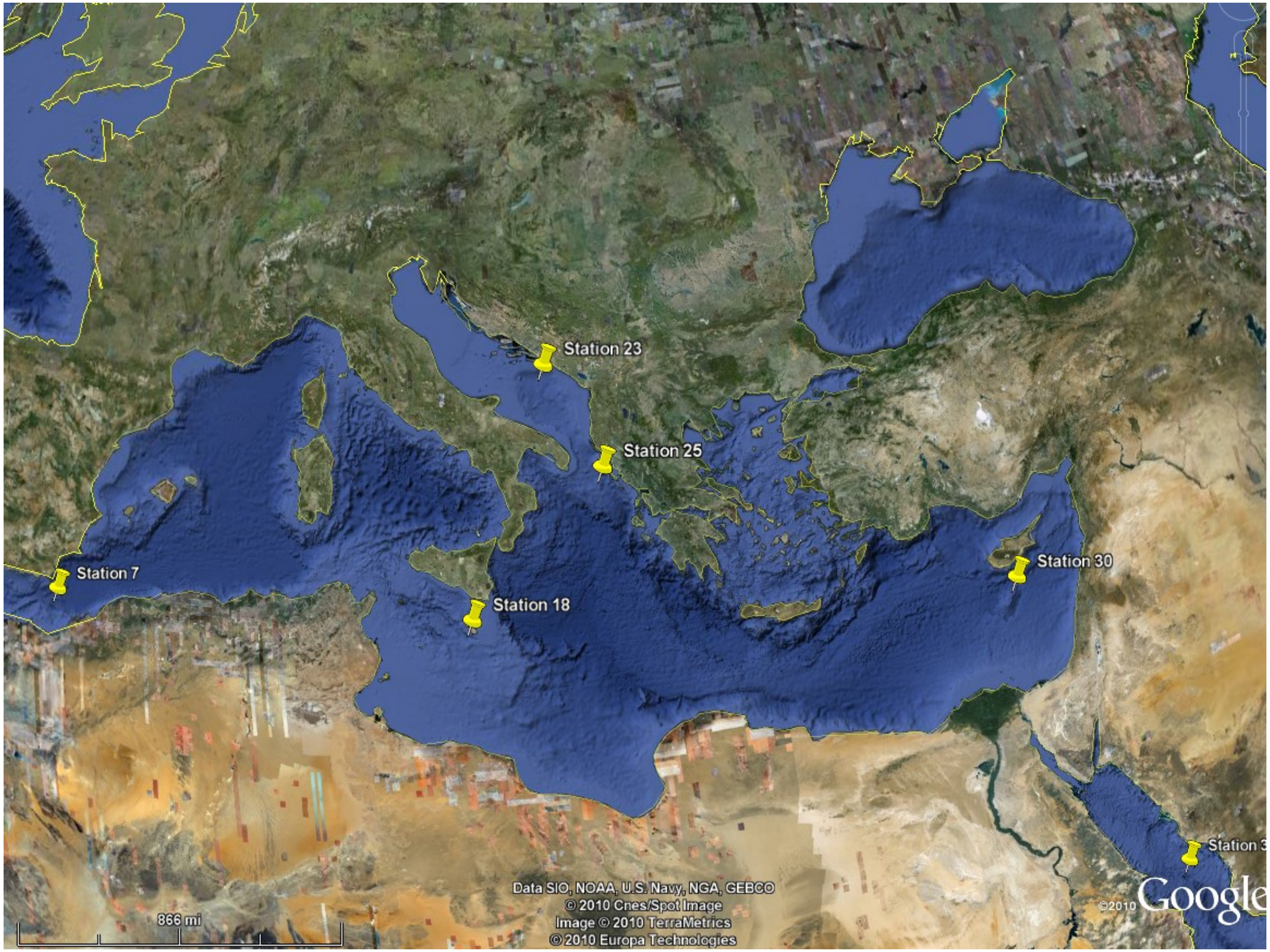
	14	15	18
14	3	0	0
16	17	0	0
17	67	7	0
18	835	9	2
19	1	0	0
20	0	0	3

Eukaryota

	14	15	16	17	18
14	0	0	0	0	2
15	0	2	0	0	3
16	3	1	13	13	49
17	0	8	8	35	99
18	10	33	26	51	333
19	211	416	142	163	870
20	106	486	790	4918	19373

Available Barcode Data

- V9 loop, DNA
- 6 stations
- 2 depths
  - Subsurface
  - DCM
- 4 size fractions:
  - 0.8-5 µm
  - 5-20 µm
  - 20-180 µm
  - 180-2000 µm



Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2010 Cnes/Spot Image
Image © 2010 TerraMetrics
© 2010 Europa Technologies

Sequence processing chain: from raw sff files to assigned sequences

```
Common Header:
  Magic Number:  0x2E736666
  Version:       0001
  Index Offset:  116227640
  Index Length:  1017430
  # of Reads:    50845
  Header Length: 840
  Key Length:    4
  # of Flows:    800
  Flowgram Code: 1
  Flow Chars:    TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
etc...
  Key Sequence:  TCAG

>GJA7C3N03FIHFN
  Run Prefix:   R_2010_06_22_11_54_29_
  Region #:     3
  XY Location:  2143_2865

  Read Header Len:  32
  Name Length:      14
  # of Bases:       234
  Clip Qual Left:   15
  Clip Qual Right:  68
  Clip Adap Left:   0
  Clip Adap Right:  0

Flowgram:       1.01    0.05    1.03    0.06    0.03    1.02    0.05    1.00    0.06    1.08    etc.
Flow Indexes:   1       3       6       8       10      12      13      16      19      21      etc.
Bases:
tcagagtgctacgaTTTGTACACACCGCCGGTCGTTCCTACCGATTGAATGATCCGGTGAGCCGCCCGatggagtcactgcgtccggttttcccgacaatgagaccttccgtgaagctt
cgtaccctaaacctttatctattttagaggaaggagaagtcgtaacaaggtccttcccgtaggtgaacctgcagaaggatcacctgagactgccaaggcacacgaggggataggn
Quality Scores: 37      37      37      37      37      37      37      37      37      37      37      37      37
37      37      37      37      39      39      37      37      39      39      37      35      35      37      etc.
```

Sequence processing chain: from raw sff files to assigned sequences

```
Common Header:
  Magic Number:  0x2E736666
  Version:       0001
  Index Offset:  116227640
  Index Length:  1017430
  # of Reads:    50845
  Header Length: 840
  Key Length:    4
  # of Flows:    800
  Flowgram Code: 1
  Flow Chars:    TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
etc...
  Key Sequence:  TCAG

>GJA7C3N03FIHFN
  Run Prefix:   R_2010_06_22_11_54_29_
  Region #:     3
  XY Location:  2143_2865

  Read Header Len:  32
  Name Length:      14
  # of Bases:       234
  Clip Qual Left:   15
  Clip Qual Right:  68
  Clip Adap Left:   0
  Clip Adap Right:  0

Flowgram:       1.01    0.05    1.03    0.06    0.03    1.02    0.05    1.00    0.06    1.08    etc.
Flow Indexes:   1       3       6       8       10      12      13      16      19      21      etc.
Bases:
tcagagtgctacgaTTTGTACACACCGCCGGTCGTTCCTACCGATTGAATGATCCGGTGAGCCGCCCGatggagtcactgcgtccggttttcccgacaatgagaccttccgtgaagctt
cgtaccctaaacctttatctattttagaggaaggagaagtcgtaacaaggtccttcccgtaggtgaacctgcagaaggatcacctgagactgccaaggcacacgaggggataggn
Quality Scores: 37      37      37      37      37      37      37      37      37      37      37      37      37
37      37      37      37      39      39      37      37      39      39      37      35      35      37      etc.
```

1. Find primers

```
GTACACACCGCCGGTCGTTCCTACCGATTGAATGATCCGGTGAGCCGCCCGatggagtcactgcgtccggttttcccgacaatgagaccttccgtgaagcttcgtaccctaaaccttta
tctattttagaggaaggagaagtcgtaacaaggtccttcccgtaggtgaacctgcagaagg
```

2. Remove primers

```
GTTCCTACCGATTGAATGATCCGGTGAGCCGCCCGatggagtcactgcgtccggttttcccgacaatgagaccttccgtgaagcttcgtaccctaaacctttatctattttagaggaag
gagaagtcgtaacaaggtccttccc
```

3. Extract corresponding qualities

```
37 39 39 37 37 39 39 37 35 35 37 37 37 37 37 39 37 37 37 37 37 37 37 37 37 37 32 25 25 25 25 28 32 33 32 32 32 28 28 19
19 19 23 24 18 17 17 23 16 16 11 14 14 14 12 12 14 11 14 19 19 17 19 25 19 17 19 26 27 17 17 13 13 13 13 21 21 16 16 18
13 16 20 22 22 13 12 12 12 12 12 12 11 11 13 16 17 12 12 12 18 12 12 12 18 22 28 20 16 11 12 11 11 11 11 11 11 11 11 16
17 28 20 18 11 11 11 11 17 13 12 12 12 20 20 22 22 20 16 16 17 18 19 19 18
```

4. Convert quality values to Expected number of errors.

$$EE = \sum_i 10^{-Q_i/10}$$

5. Dereplicate sequences

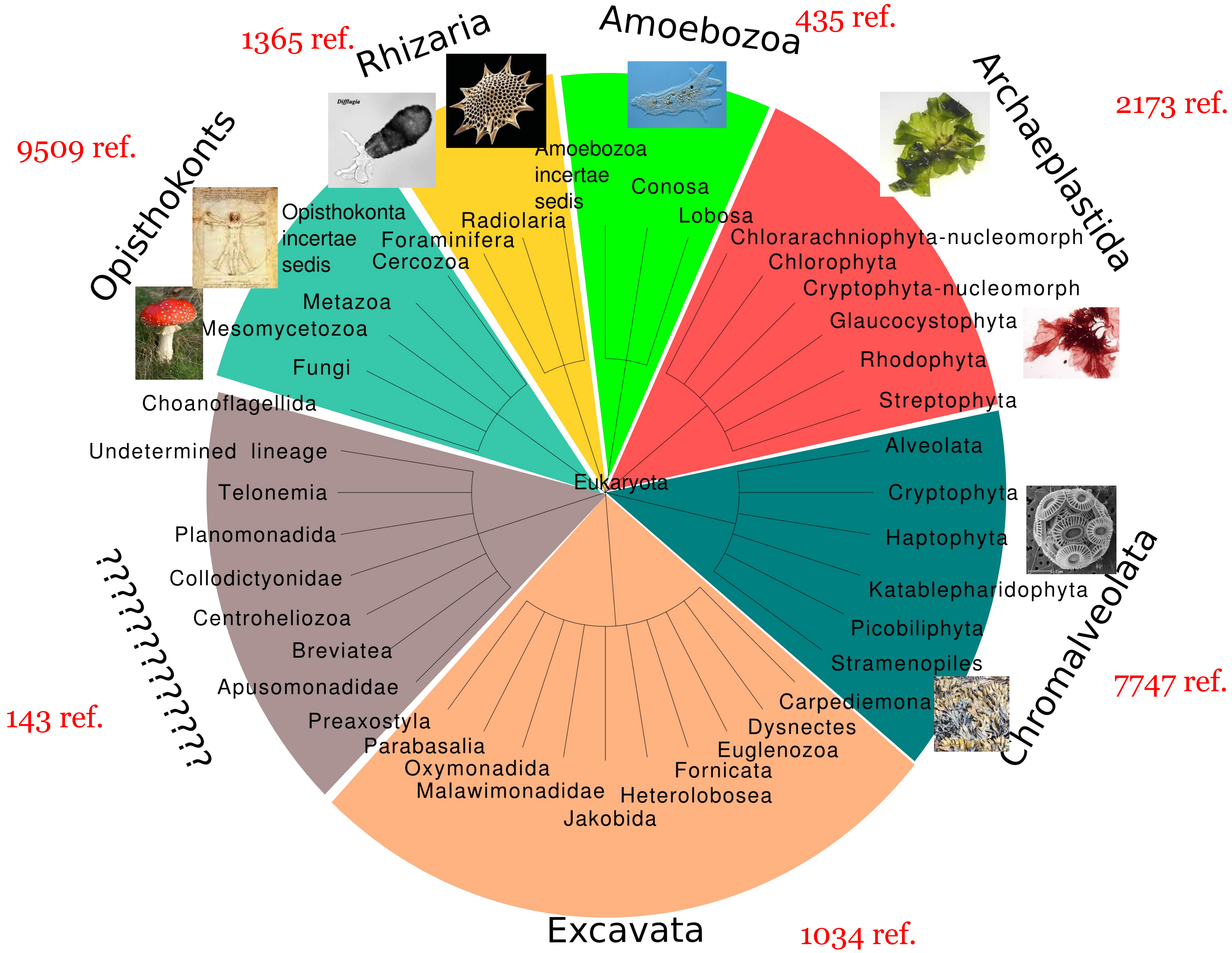
```
>NPGJ8MOYR06G3N03_724   EE=0.015525     MD5=4e38e8ced9070952b314e1880bede1ca    L=127   S=T0099
GCTACTACCGATTGAACGTTTTAGTGAGGTCCTCGGACTGTTTGGTAGTCGGATCACTCTGACTGCCTGGCGGGAAGACG
ACCAAACTGTAGCGTTTAGAGGAAGTAAAAGTCGTAACAAGGTTTCC
```

5. Filter sequences:

Sequences which are unique (within or between samples), with EE/L > 1% (more than 1 % of error) are disgarded.
In the above case EE = 3.93, sequence is disgarded.

Sequence assignation: giving a name to those sequences

- Sequence assignation is performed from a dedicated Reference database:
  - Built from a curated database: we use Laure Guillou's 18S database, soon to be incorporated in Silva.
  - We extract (from primer search with degeneracy), parts of the sequence corresponding to the targeted fragment.
    - One primer set = one reference database
    - guided search: less false positives, faster processing


1365 ref.
Rhizaria
Amoebozoa
435 ref.
Archaeplastida
2173 ref.
9509 ref.
Opisthokonts
Difflugia
Amoebozoa incertae sedis
Conosa
Lobosa
Opisthokonta incertae sedis
Radiolaria
Foraminifera
Cercozoa
Chlorarachniophyta-nucleomorph
Chlorophyta
Cryptophyta-nucleomorph
Glaucocystophyta
Rhodophyta
Streptophyta
Metazoa
Mesomycetozoa
Fungi
Choanoflagellida
Undetermined lineage
Telonemia
Planomonadida
Collodictyonidae
Centroheliozoa
Breviatea
Apusomonadidae
Preaxostyla
Parabasalia
Oxymonadida
Malawimonadidae
Jakobida
Heterolobosea
Fornicata
Euglenozoa
Dysnectes
Carpediemonas
Eukaryota
Alveolata
Cryptophyta
Haptophyta
Katablepharidophyta
Picobiliphyta
Stramenopiles
Chromalveolata
7747 ref.
?????????????
143 ref.
Excavata
1034 ref.

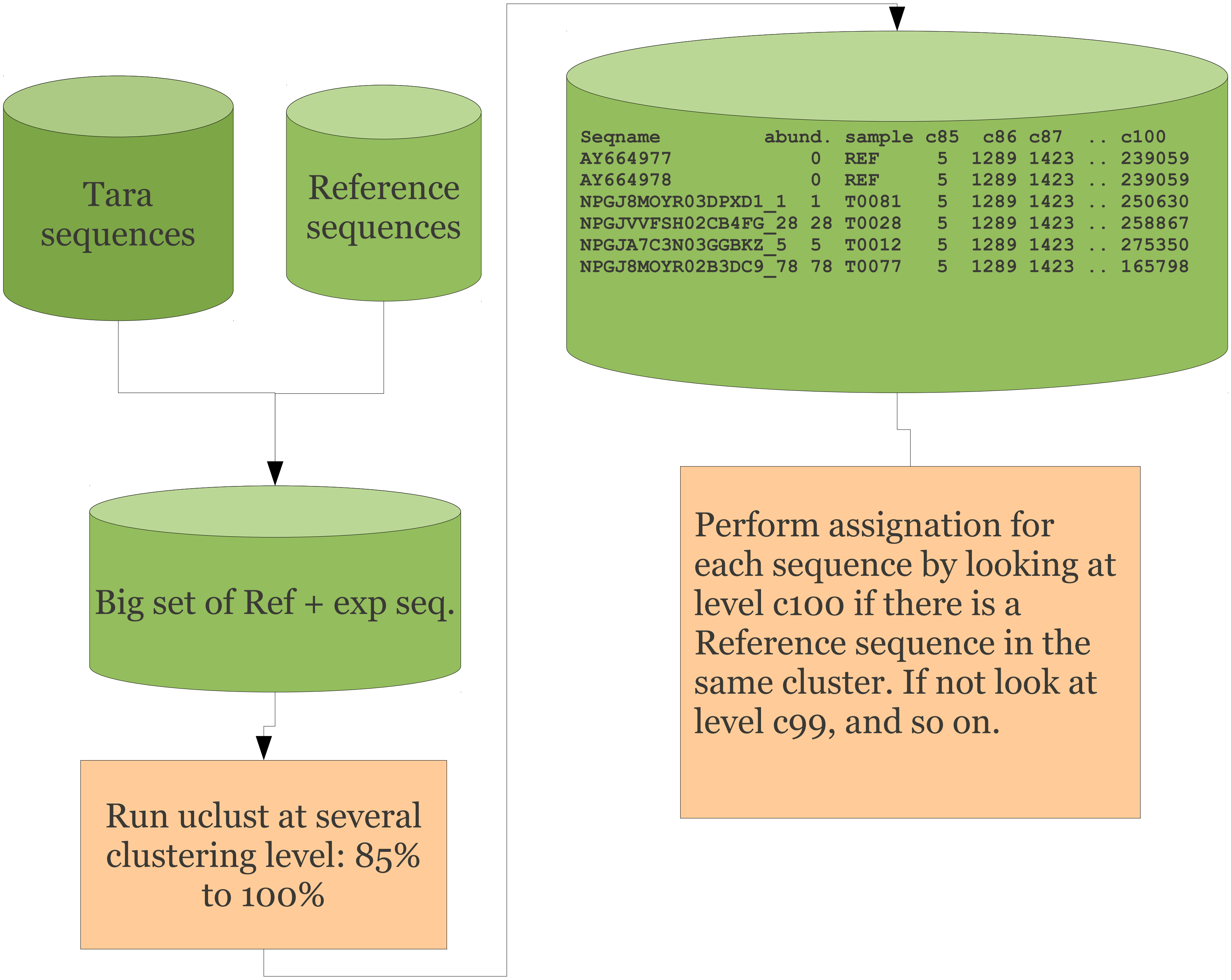
Example reference database

Identifier

Sequence

Lineage

```
X53904|
gctcctaccgattgggtgtgctggtgaagtgttcggattgactttgactggggcaacctggtcatggtcgagaagttcattaaaccctcccacctagaggaaggagaagtcgtaacaagg
ttaccgtaggtgaacctgcggaaggatcattg|Eukaryota|Archaeplastida|Chlorophyta|Chlorophyceae|Chlorophyceae|CW_Chlamydomonadales|
Volvox|carteri|*
X73991|
gctcctaccgattggatgtgctggtgaagcgttcggattggagtcactcggcggctcgtccgccggctggcaccgagaagttcgttaaaccctcccatctagaggaaggagaagtcgtaa
caaggtctcc|Eukaryota|Archaeplastida|Chlorophyta|Trebouxiophyceae|Trebouxiophyceae|Watanabea Clade|Watanabea|reniformis|*
U13985|
gctcctaccgattgggtgtgctggtgaagtgttcggattgacttcagcgggggcaactctgctgttgttgagaagatcattaaaccctcccacctagaggaaggagaagtcgtaacaagg
tttccgtagtgaacc|Eukaryota|Archaeplastida|Chlorophyta|Chlorophyceae|Chlorophyceae|CW_Chlamydomonadales|Chlamydomonas|
dysosmos|*
X62441|
gctcctaccgattgggtgtgctggtgaagtgttcggattggcgaccgggggcggtctccgctctcggccgccgagaagttcattaaaccctcccacctagaggaaggagaagtcgtaaca
aggtttcc|Eukaryota|Archaeplastida|Chlorophyta|Trebouxiophyceae|Trebouxiophyceae|Chlorellales|Chlorella|sorokiniana|*
U70784|gctcctaccgattgggtgtgctggtgaagtgttcggattggcttcaggtgatggcaacatcgcttgttgctgagaagttcattaaaccctcccacctagaggaagg|
Eukaryota|Archaeplastida|Chlorophyta|Chlorophyceae|Chlorophyceae|CW_Chlamydomonadales|Lobochlamys|segnis|*
AJ132619|
gctcctaccgattggattggctggtgaagcgttcggattgtgggttcggagcggttcgccgttctgggcttgcgagaagttcgttaaaccctcctatctagaggaaggagaagtcgtaac
aaggtttccgt|Eukaryota|Archaeplastida|Chlorophyta|Prasinophyceae|Prasino-Clade V|Pycnococcaceae|Pseudoscourfieldia|
marina|*
AB017123|
gctcctaccgattgaatggtccggtgaaatgttcggaccgcggcgagacggacggttcgctgtcagcctcgctgcgggaagttcattaaaccttatcatttagaggaaggagaagtcgta
acaaggtttcc|Eukaryota|Archaeplastida|Chlorophyta|Prasinophyceae|Pyramimonadales|*|Pyramimonas|propulsa|*
X73993|
gctcctaccgattgggtgtgctggtgaagtgttcggattggcgaccgggggcggtctccgctctcggccgccgagaagttcattaaaccctcccacctagaggaaggagaagtcgtaaca
aggtttcc|Eukaryota|Archaeplastida|Chlorophyta|Trebouxiophyceae|Trebouxiophyceae|Chlorellales|Chlorella|sorokiniana|*
U13984|
gctcctaccgattgggtgtgctggtgaagtgttcggattgacttcagcgggggcaactctgctgttgttgagaagatcattaaaccctcccacctagaggaaggagaagtcgtaacaagg
tttccgtagtgaacc|Eukaryota|Archaeplastida|Chlorophyta|Chlorophyceae|Chlorophyceae|CW_Chlamydomonadales|Chlamydomonas|
applanata|*
X74001|
gctcctaccgattgggtgtgctggtgaagtgttcggattggcgaccgggggcggtctccgctctcggccgccgagaagttcattaaaccctcccacctagaggaaggagaagtcgtaaca
aggtttcc|Eukaryota|Archaeplastida|Chlorophyta|Trebouxiophyceae|Trebouxiophyceae|Chlorellales|Chlorella|sorokiniana|*
```



Tara sequences

Reference sequences

Big set of Ref + exp seq.

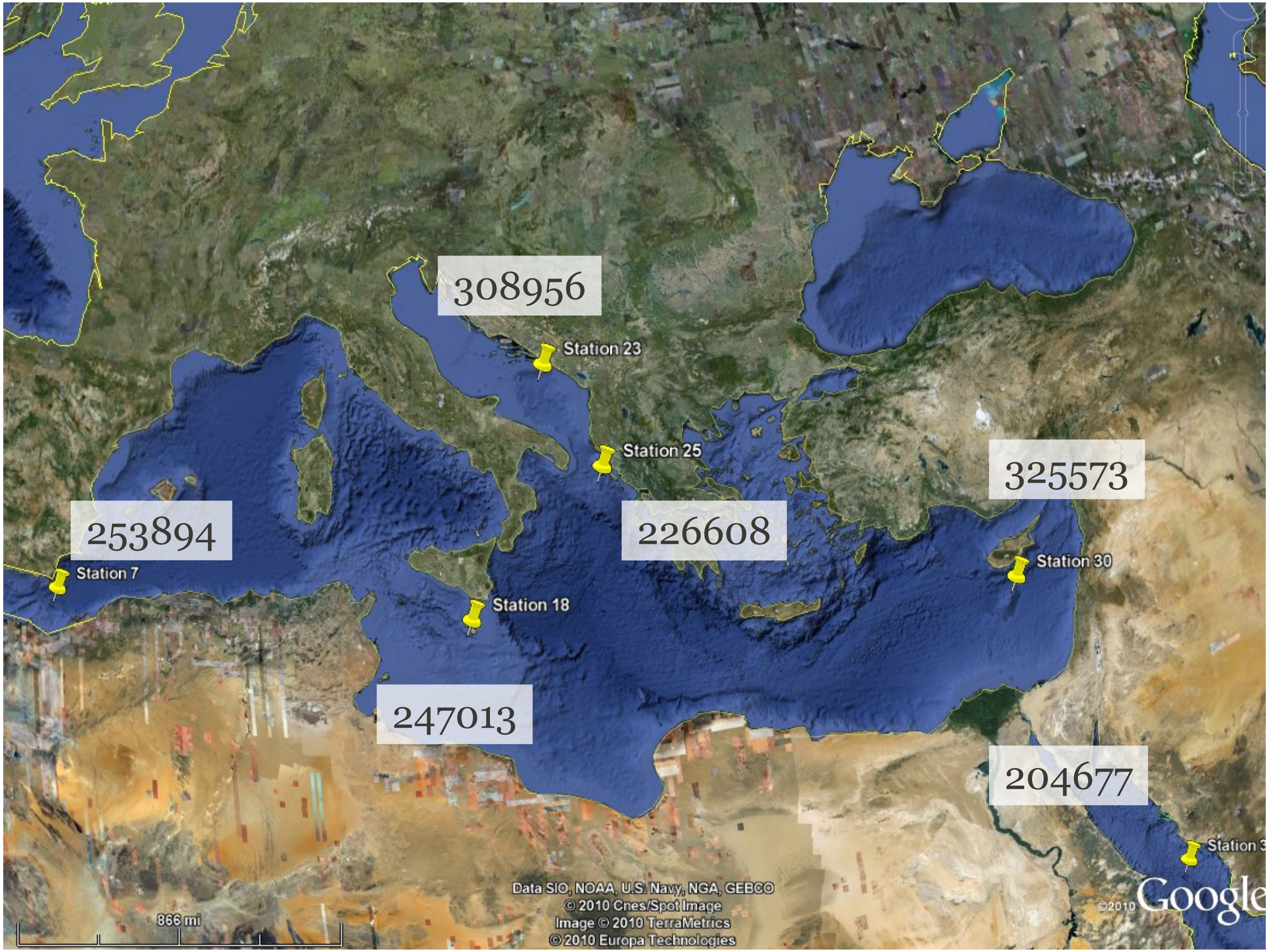
Run uclust at several clustering level: 85% to 100%

```
Seqname            abund. sample c85  c86 c87  .. c100
AY664977                0  REF      5  1289 1423 .. 239059
AY664978                0  REF      5  1289 1423 .. 239059
NPGJ8MOYR03DPXD1_1  1  T0081    5  1289 1423 .. 250630
NPGJVVFSH02CB4FG_28 28 T0028    5  1289 1423 .. 258867
NPGJA7C3N03GGBKZ_5  5  T0012    5  1289 1423 .. 275350
NPGJ8MOYR02B3DC9_78 78 T0077    5  1289 1423 .. 165798
```

Perform assignation for each sequence by looking at level c100 if there is a Reference sequence in the same cluster. If not look at level c99, and so on.

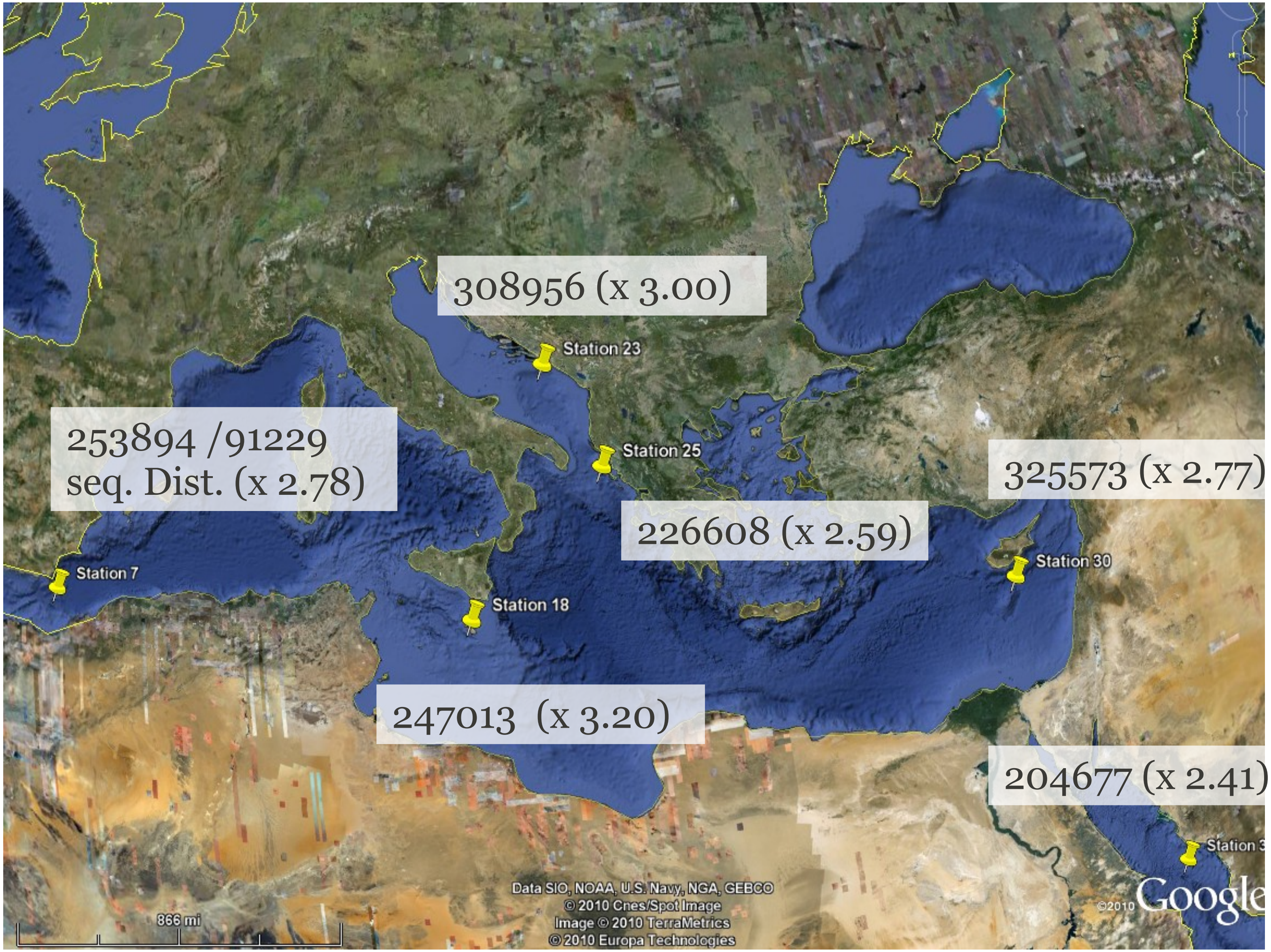
Preliminary analysis of the Barcode datasets

Number of valid sequences



Preliminary analysis of the Barcode datasets

Multiplicity of sequences



308956 (x 3.00)

253894 /91229
seq. Dist. (x 2.78)

325573 (x 2.77)

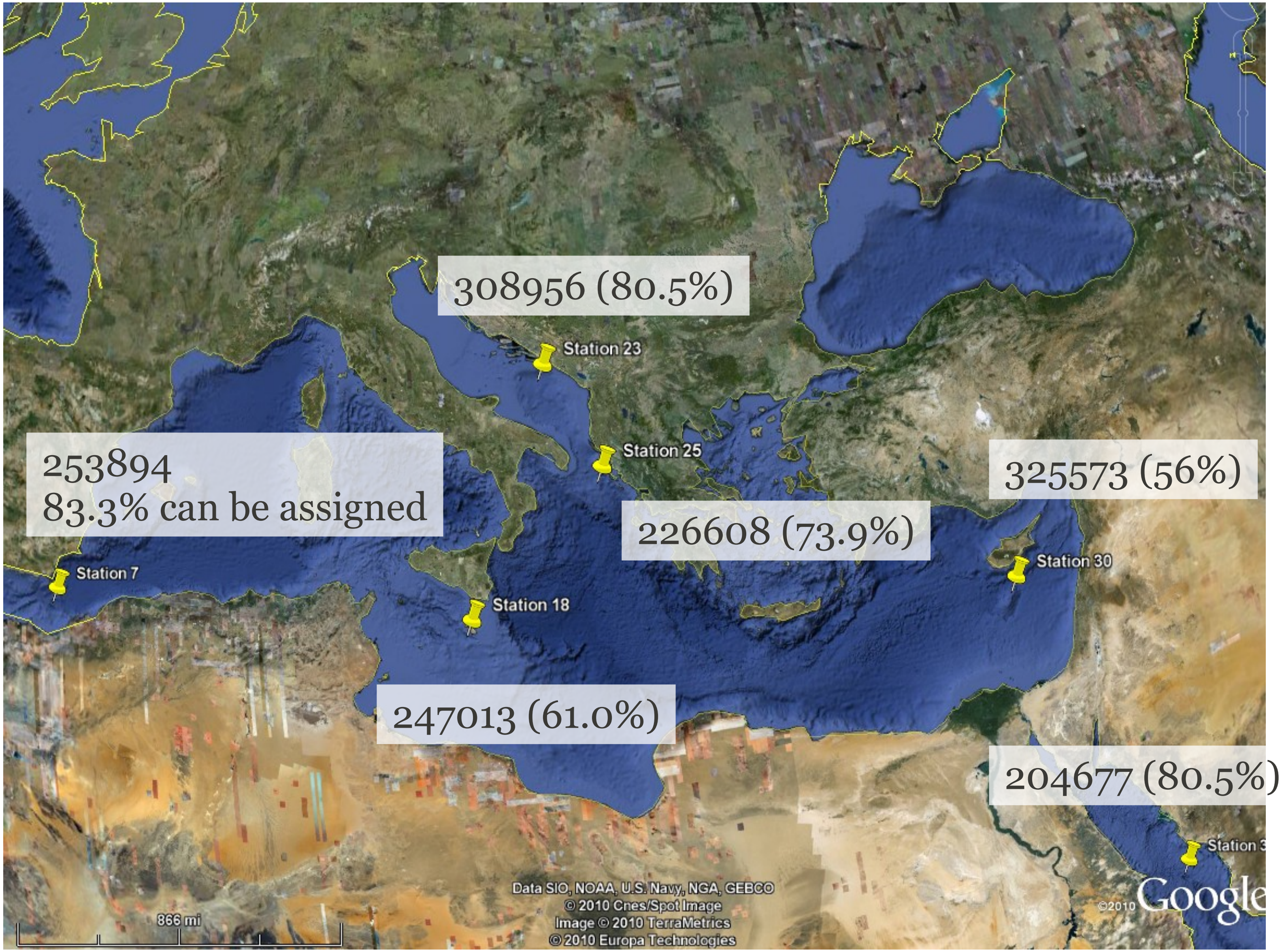
226608 (x 2.59)

247013 (x 3.20)

204677 (x 2.41)

Preliminary analysis of the Barcode datasets

Fraction of assigned sequences



308956 (80.5%)

253894
83.3% can be assigned

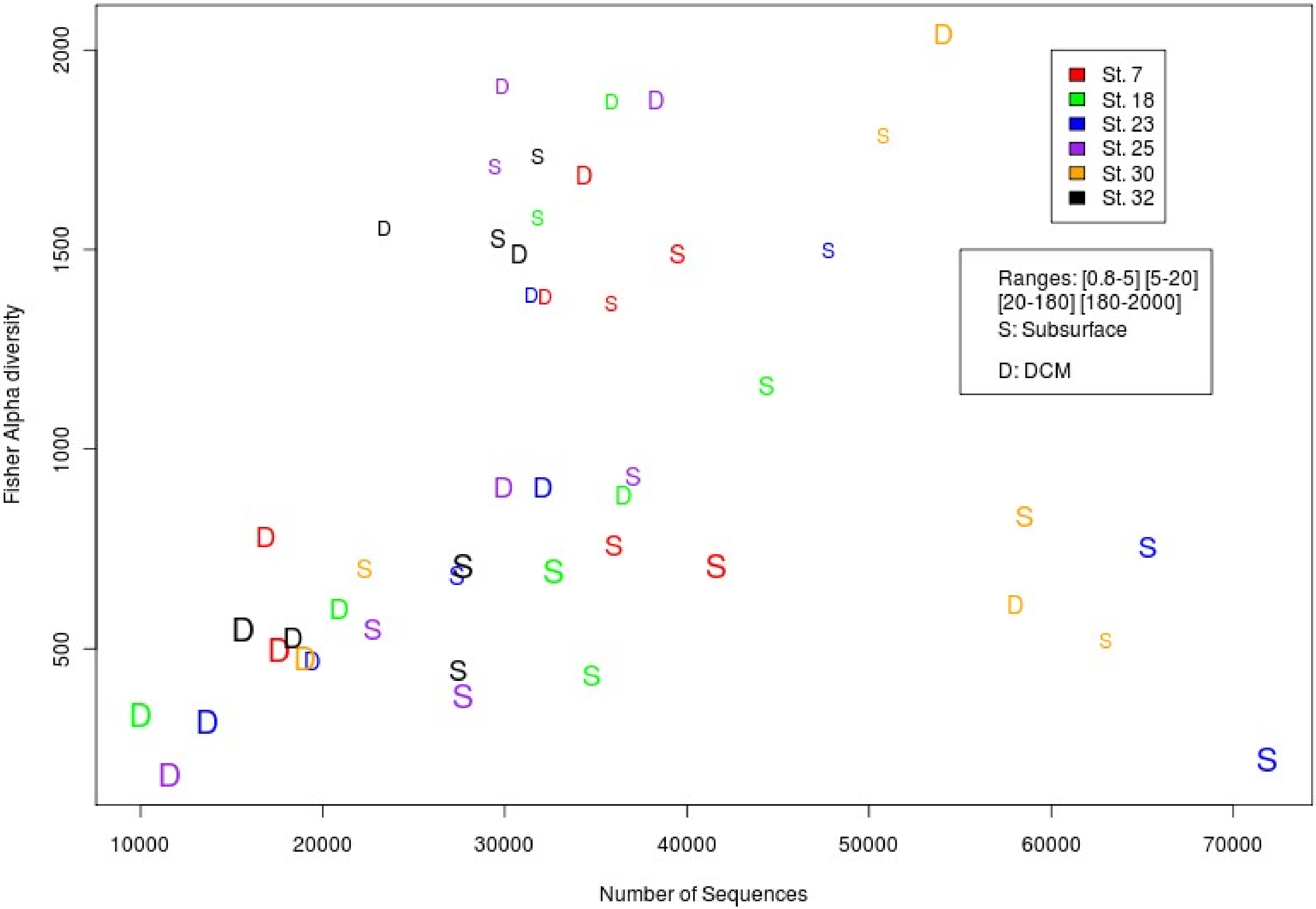
325573 (56%)

226608 (73.9%)

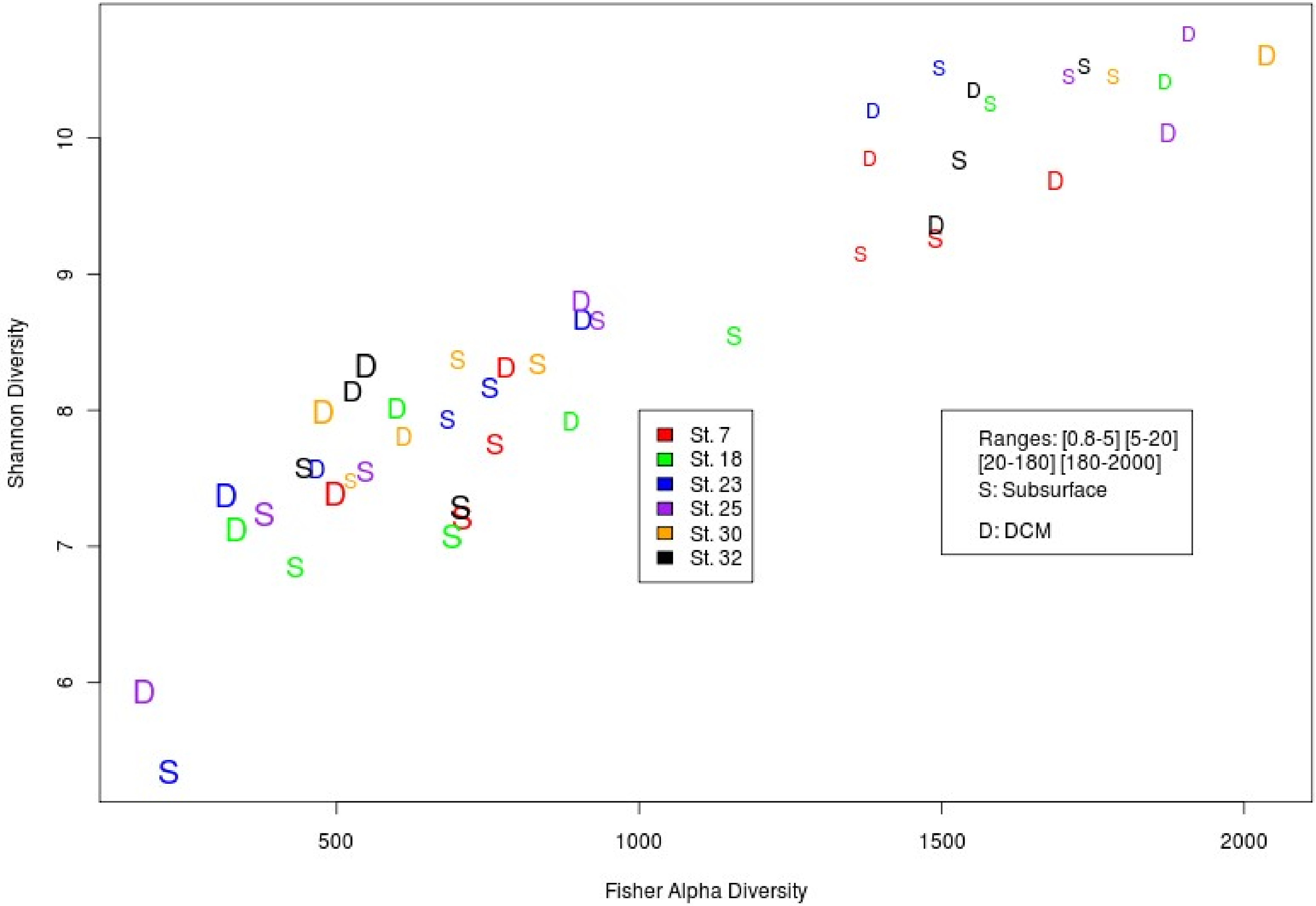
247013 (61.0%)

204677 (80.5%)

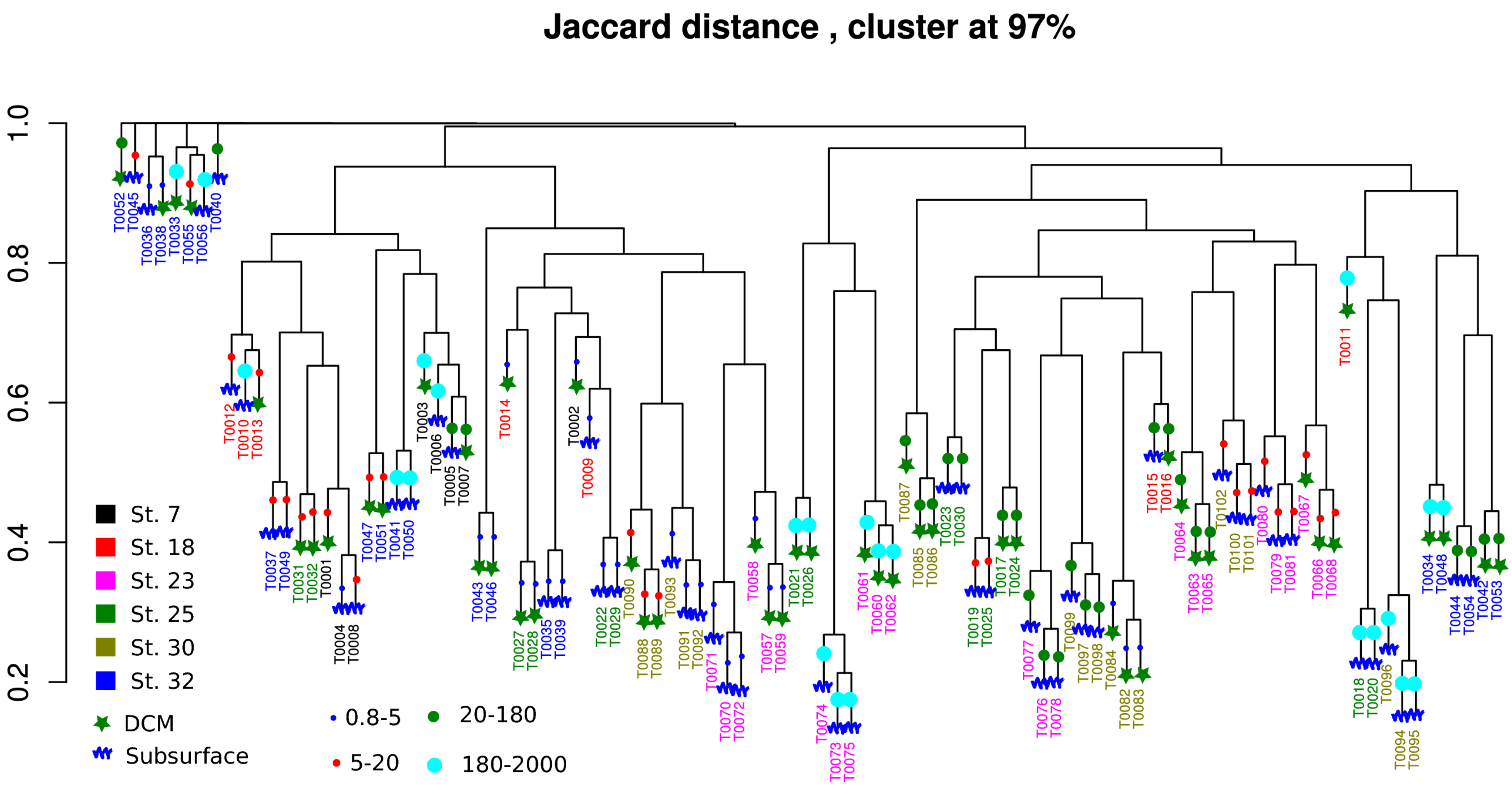
Small size fractions are more diverse



Small size fractions are more diverse



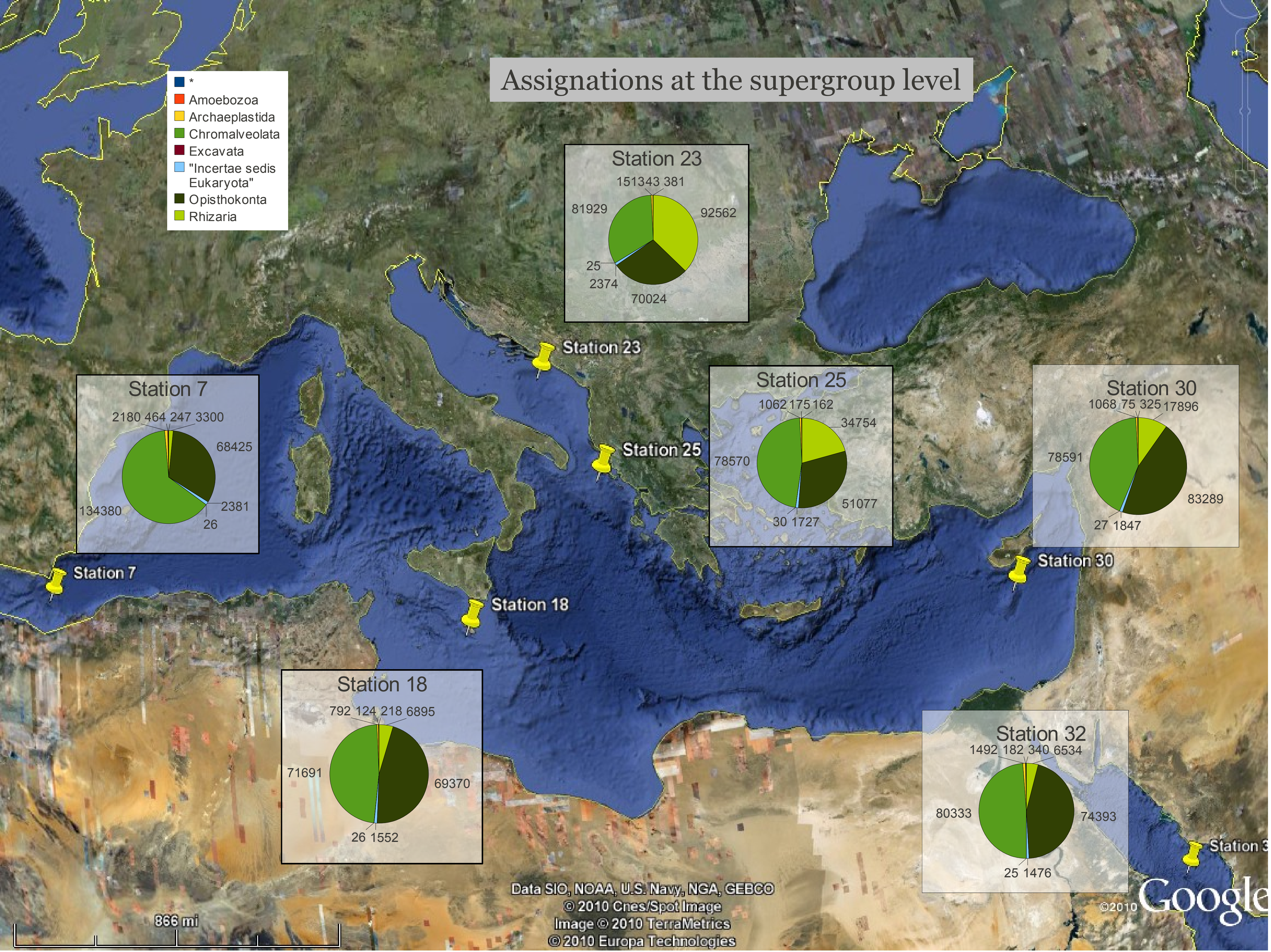
Clustering of samples



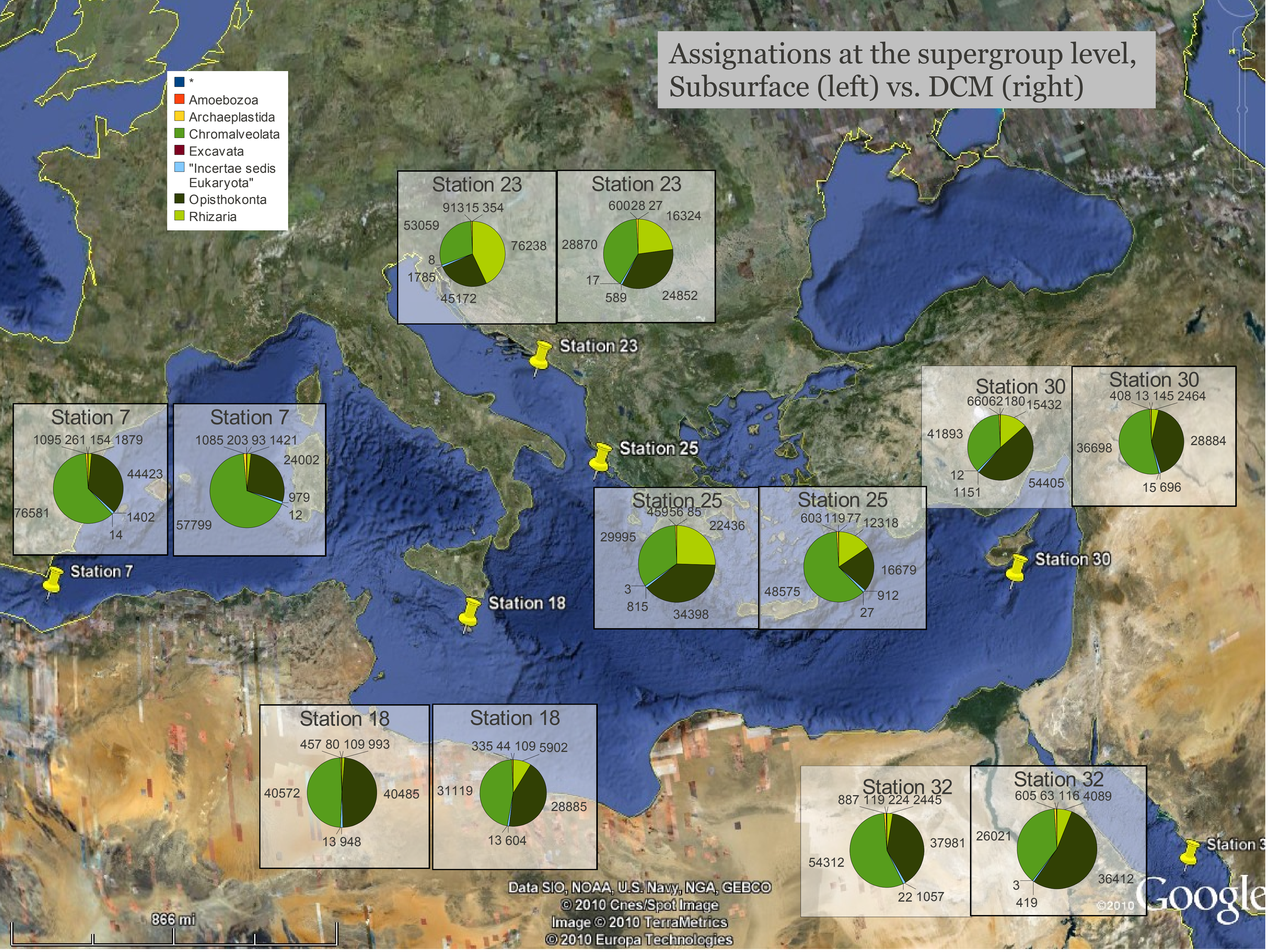
Distance between sample is calculated for sequence clusters at 97% identity (uclust). We used Jaccard distance: J(A,B)=1 - |A∩B|/|A∪B|.
3 sets of seq. data for St 32 (1 run of sequencing failed-> cluster on the top left)

Small size fraction of all stations cluster together (one exception, T004?)

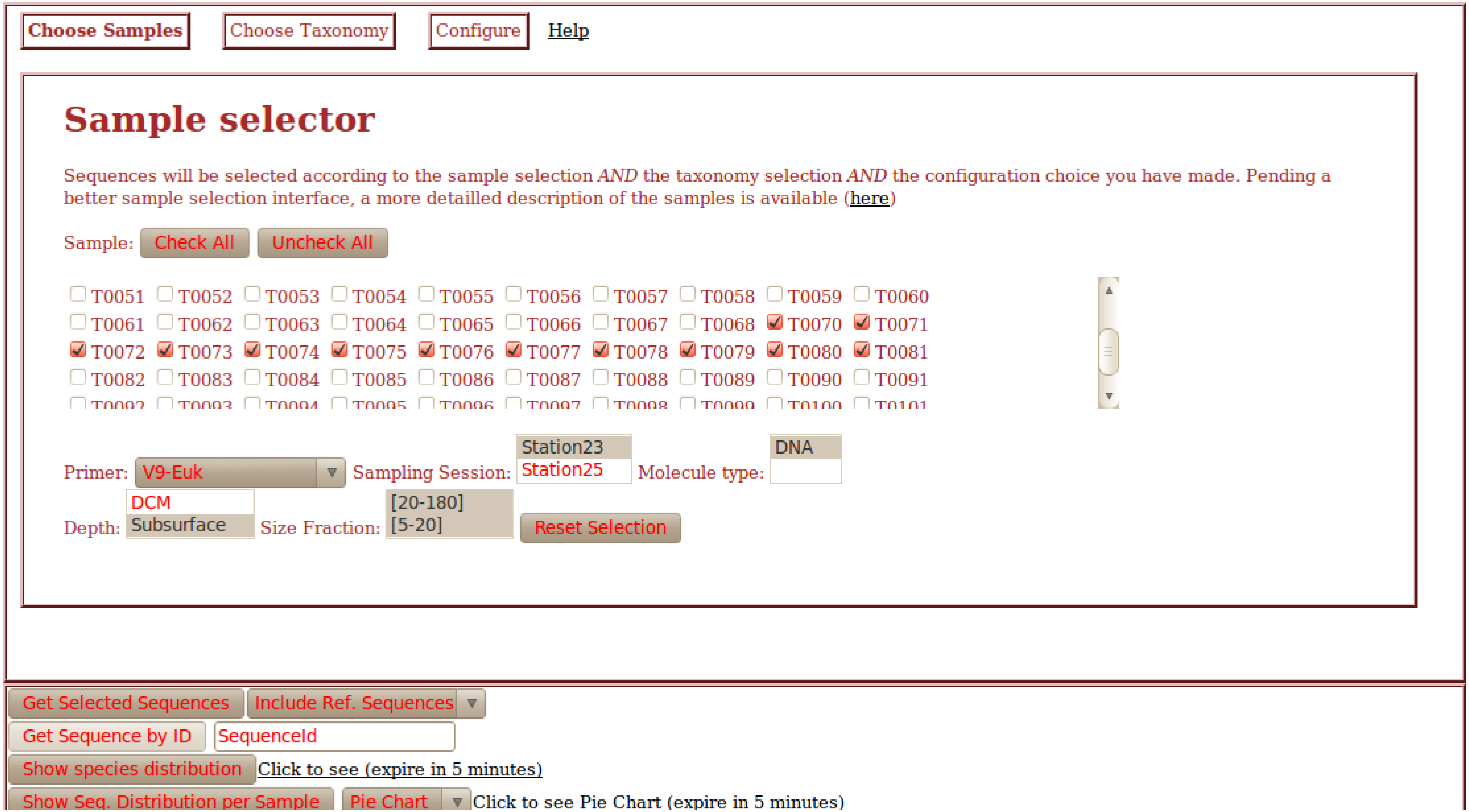
Assignations at the supergroup level



Assignations at the supergroup level, Subsurface (left) vs. DCM (right)



Interactive Web Interface



Choose Samples | Choose Taxonomy | Configure | Help

Sample selector

Sequences will be selected according to the sample selection *AND* the taxonomy selection *AND* the configuration choice you have made. Pending a better sample selection interface, a more detailled description of the samples is available (here)

Sample: Check All | Uncheck All

☐ T0051 ☐ T0052 ☐ T0053 ☐ T0054 ☐ T0055 ☐ T0056 ☐ T0057 ☐ T0058 ☐ T0059 ☐ T0060
☐ T0061 ☐ T0062 ☐ T0063 ☐ T0064 ☐ T0065 ☐ T0066 ☐ T0067 ☐ T0068 ☑ T0070 ☑ T0071
☑ T0072 ☑ T0073 ☑ T0074 ☑ T0075 ☑ T0076 ☑ T0077 ☑ T0078 ☑ T0079 ☑ T0080 ☑ T0081
☐ T0082 ☐ T0083 ☐ T0084 ☐ T0085 ☐ T0086 ☐ T0087 ☐ T0088 ☐ T0089 ☐ T0090 ☐ T0091
☐ T0092 ☐ T0093 ☐ T0094 ☐ T0095 ☐ T0096 ☐ T0097 ☐ T0098 ☐ T0099 ☐ T0100 ☐ T0101

Primer: V9-Euk
Sampling Session: Station23, Station25
Molecule type: DNA

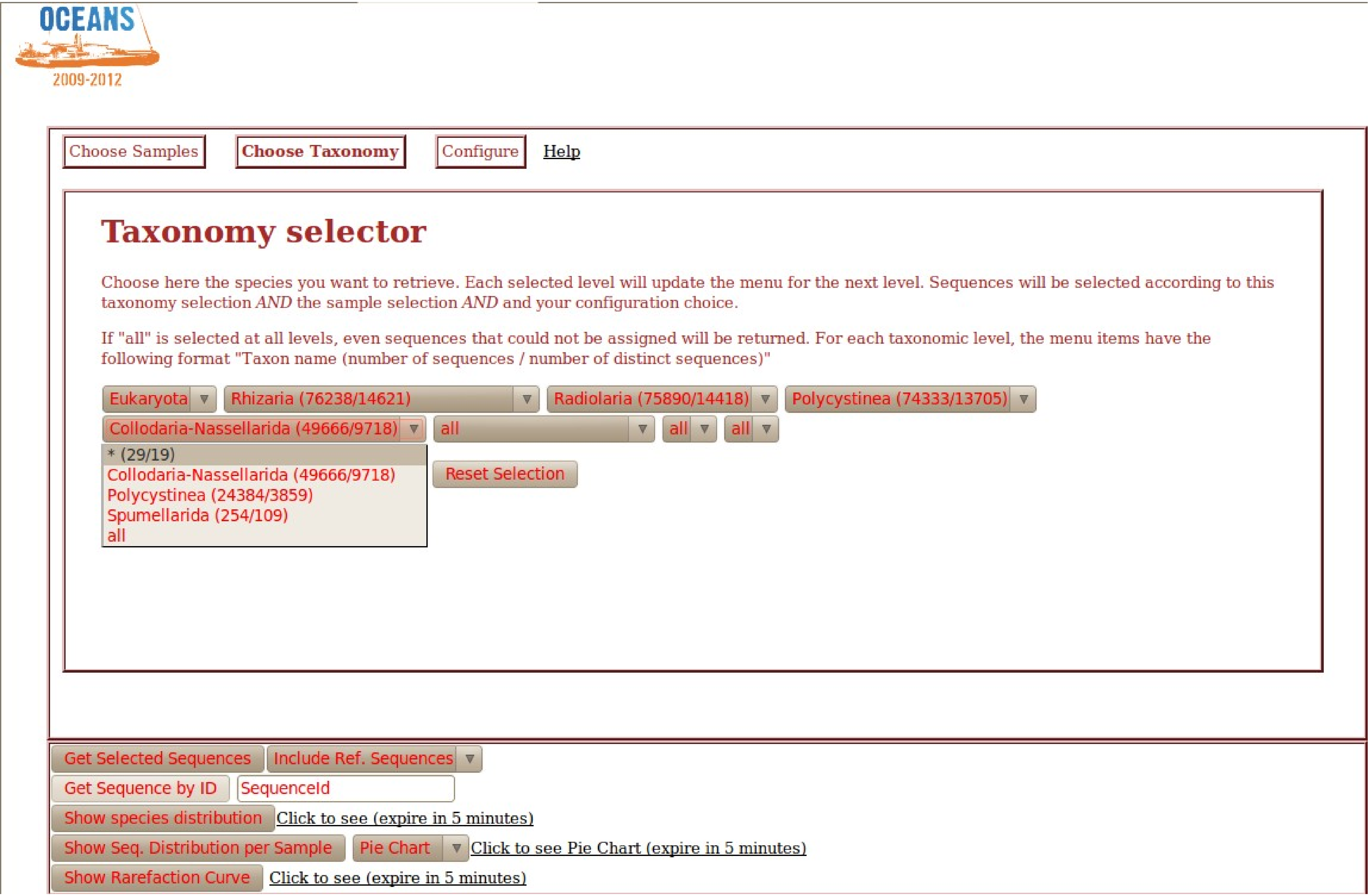
Depth: DCM, Subsurface
Size Fraction: [20-180], [5-20]
Reset Selection

Get Selected Sequences | Include Ref. Sequences
Get Sequence by ID | SequenceId
Show species distribution | Click to see (expire in 5 minutes)
Show Seq. Distribution per Sample | Pie Chart | Click to see Pie Chart (expire in 5 minutes)

Taxonomy selector

Choose here the species you want to retrieve. Each selected level will update the menu for the next level. Sequences will be selected according to this taxonomy selection *AND* the sample selection *AND* and your configuration choice.

If "all" is selected at all levels, even sequences that could not be assigned will be returned. For each taxonomic level, the menu items have the following format "Taxon name (number of sequences / number of distinct sequences)"



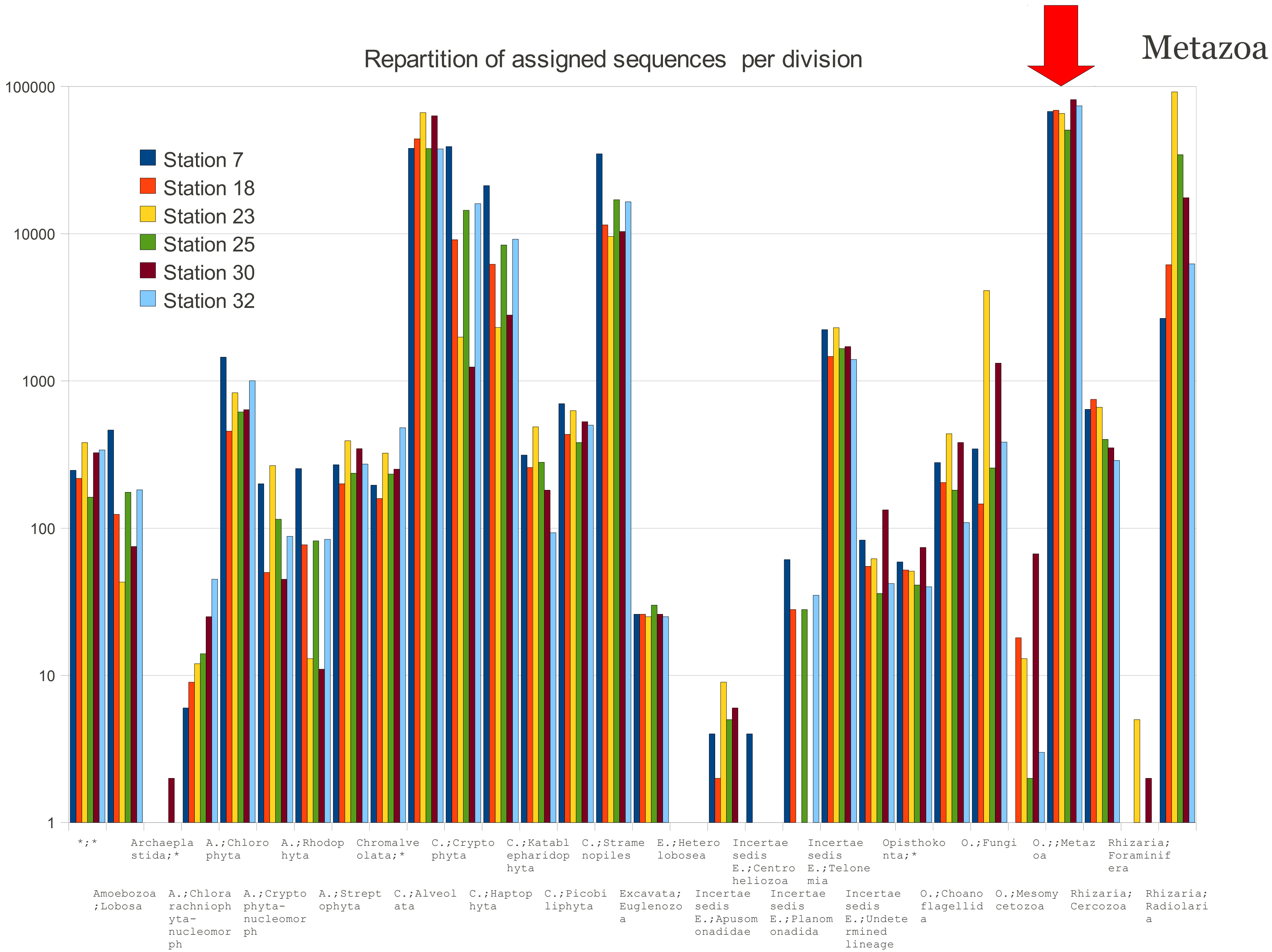
Taxonomy selector

Choose here the species you want to retrieve. Each selected level will update the menu for the next level. Sequences will be selected according to this taxonomy selection *AND* the sample selection *AND* and your configuration choice.

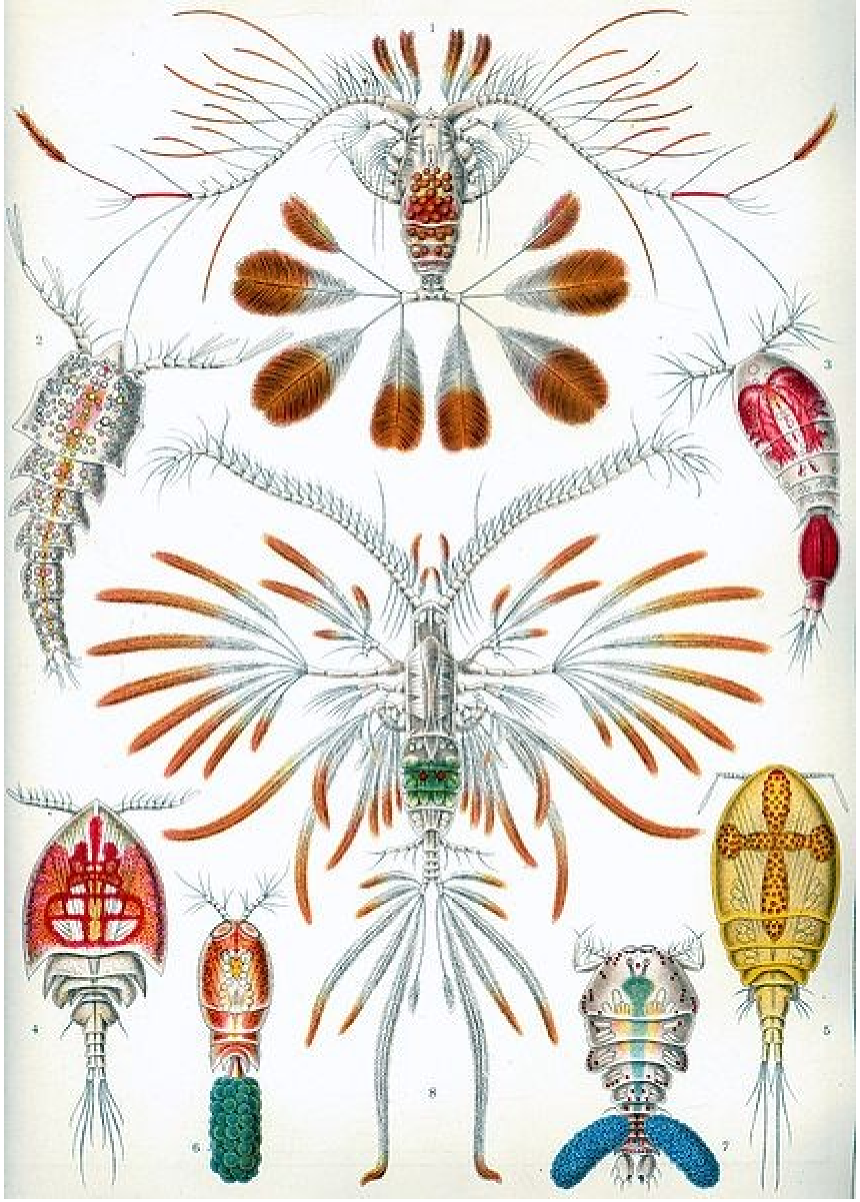
If "all" is selected at all levels, even sequences that could not be assigned will be returned. For each taxonomic level, the menu items have the following format "Taxon name (number of sequences / number of distinct sequences)"

Eukaryota
Rhizaria (76238/14621)
Radiolaria (75890/14418)
Polycystinea (74333/13705)
Collodaria-Nassellarida (49666/9718)
all
all
all

* (29/19)
Collodaria-Nassellarida (49666/9718)
Polycystinea (24384/3859)
Spumellarida (254/109)
all


Metazoa
Repartition of assigned sequences per division
100000
10000
1000
100
10
1
Station 7
Station 18
Station 23
Station 25
Station 30
Station 32
;
Amoebozoa;Lobosa
Archaeplastida;*
A.;Chlorarachniophyta-nucleomorph
A.;Chlorophyta
A.;Cryptophyta-nucleomorph
A.;Rhodophyta
A.;Streptophyta
Chromalveolata;*
C.;Alveolata
C.;Cryptophyta
C.;Haptophyta
C.;Katablepharidophyta
C.;Picobiliphyta
C.;Stramenopiles
Excavata;Euglenozoa
E.;Heterolobosea
Incertae sedis E.;Apusomonadidae
Incertae sedis E.;Centroheliozoa
Incertae sedis E.;Planomonadida
Incertae sedis E.;Telonemia
Incertae sedis E.;Undetermined lineage
Opisthokonta;*
O.;Choanoflagellida
O.;Fungi
O.;Mesomycetozoa
O.;;Metazoa
Rhizaria;Cercozoa
Rhizaria;Foraminifera
Rhizaria;Radiolaria

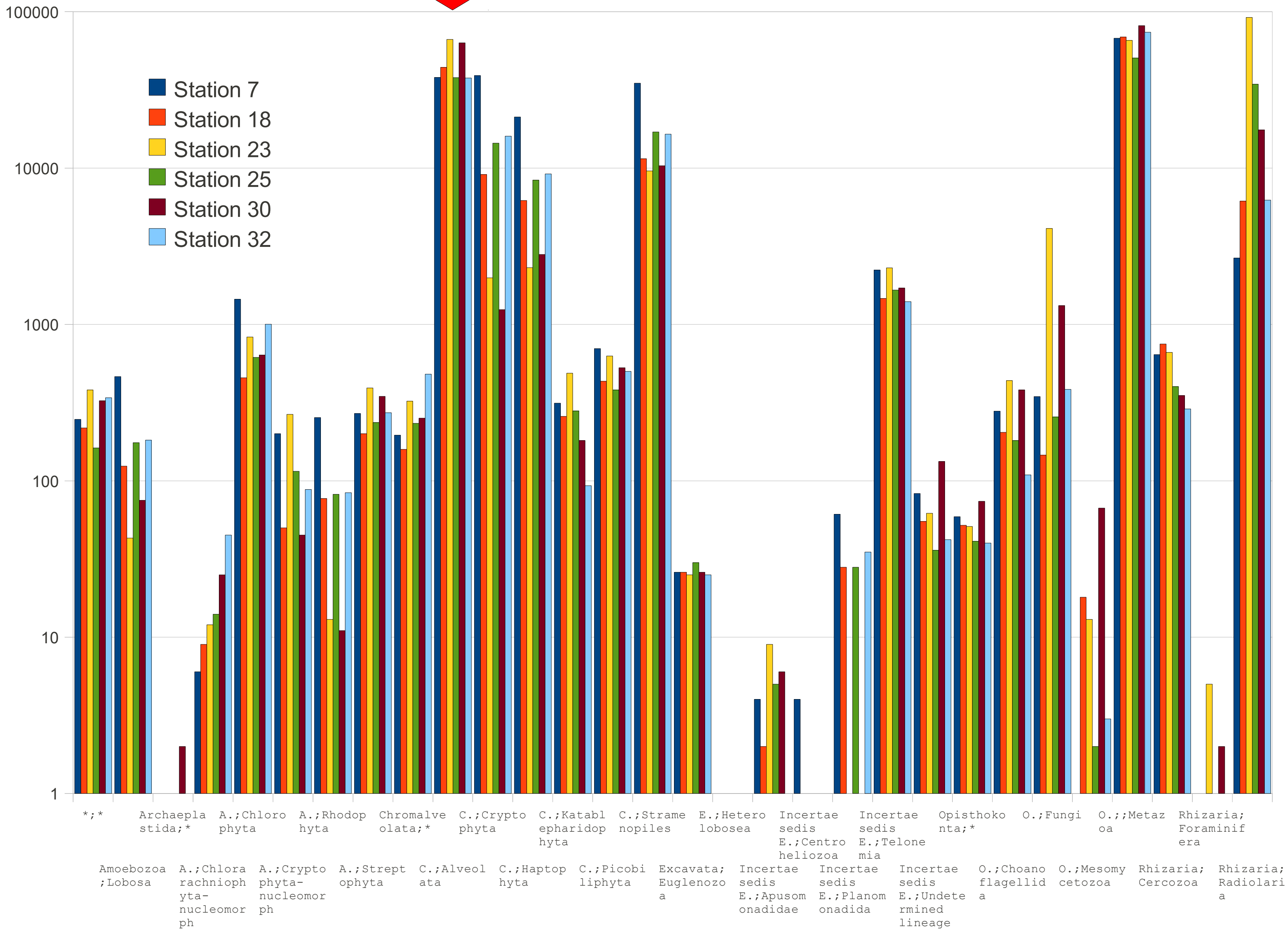
294714	Arthropoda Crustacea
54281	Cnidaria Cnidaria
15201	Urochordata Urochordata
11263	* *
10044	Annelida Annelida
8888	Chaetognatha Aphragmophora
4148	Ctenophora Ctenophora
3522	Mollusca Gastropoda
1452	Platyhelminthes Turbellaria
673	Craniata Craniata



Kunstformen der Natur (1900), Ernst Haeckel

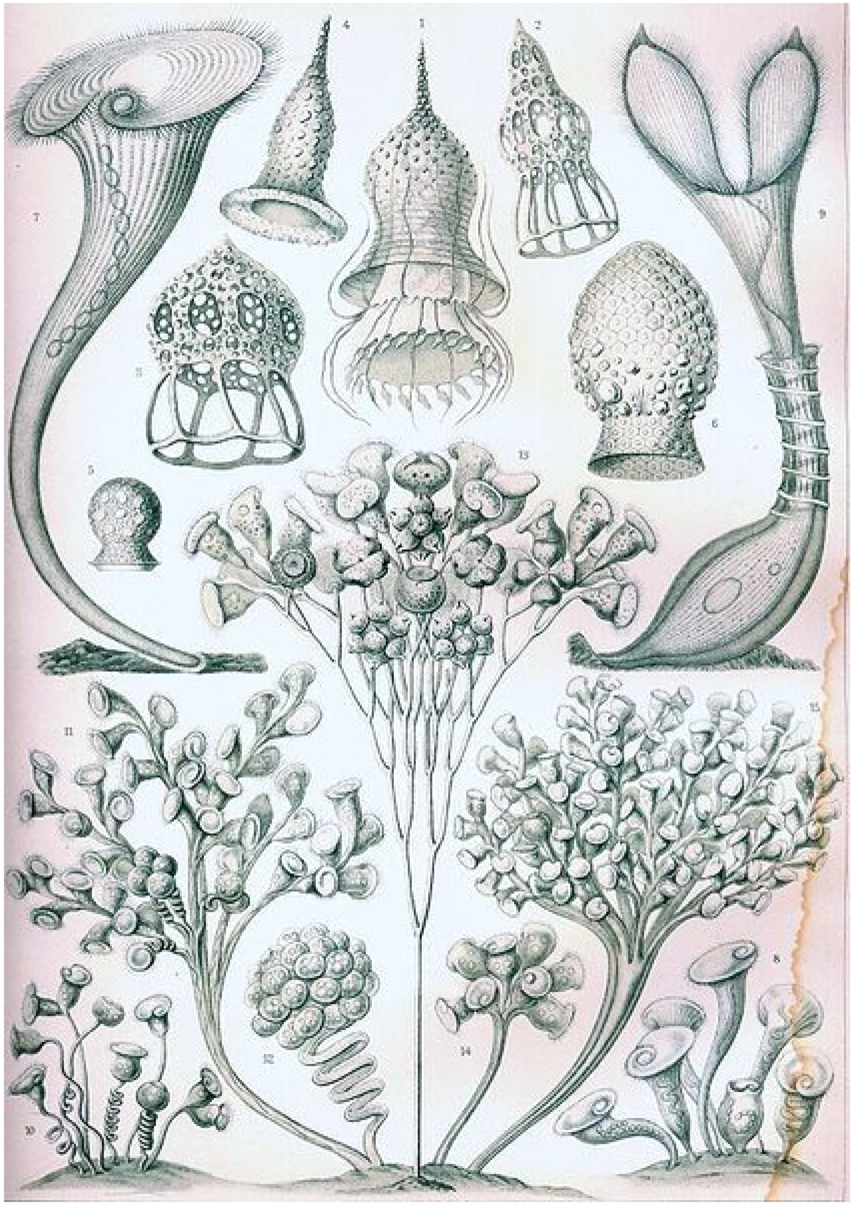
Alveolata

Rep tion of assigned sequences per division

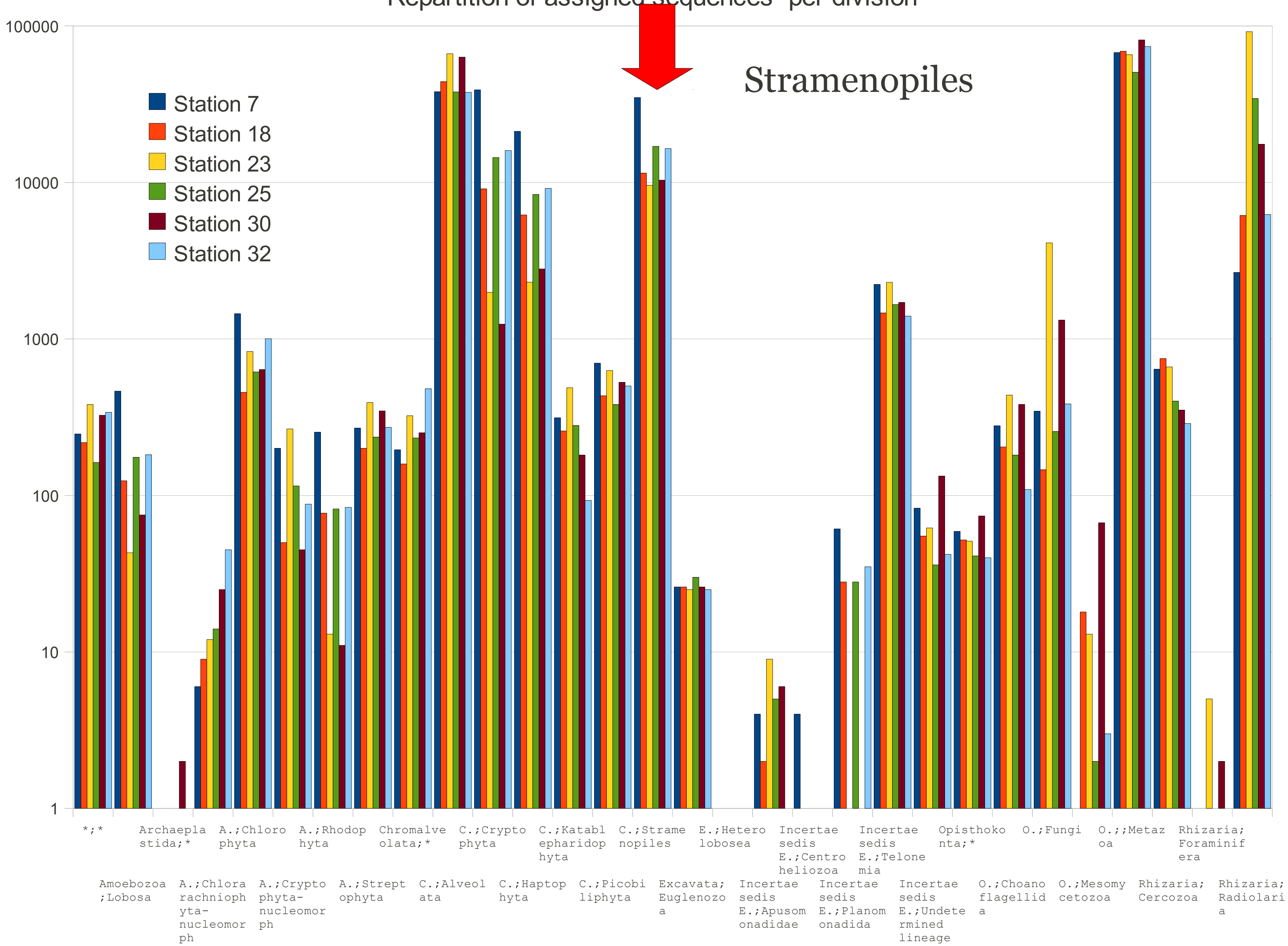


Kunstformen der Natur (1900), Ernst Haeckel

140728 Dinophyceae Dinophyceae
111073 Dinophyceae Syndiniales
12998 Dinophyceae *
8937 Ciliophora Oligotrichea
6916 * *
1453 Perkinsea Perkinsida
1386 Ciliophora Oligohymenophorea
929 Apicomplexa Coccidia
564 Ciliophora Colpodea and Nassophorea
519 Apicomplexa Cryptosporidium

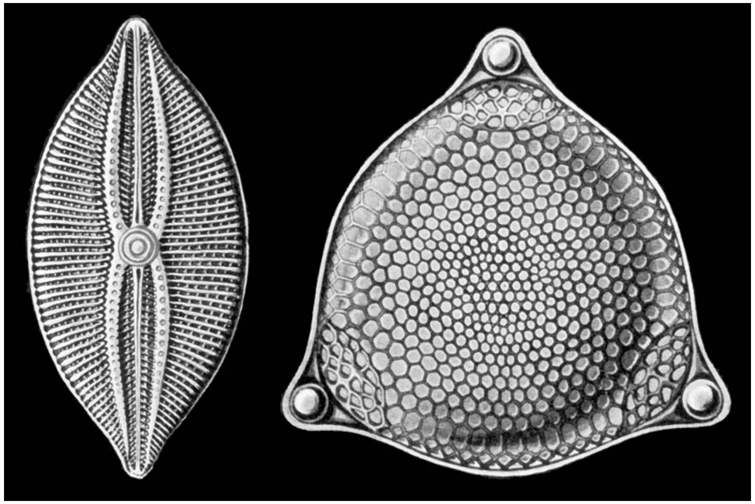


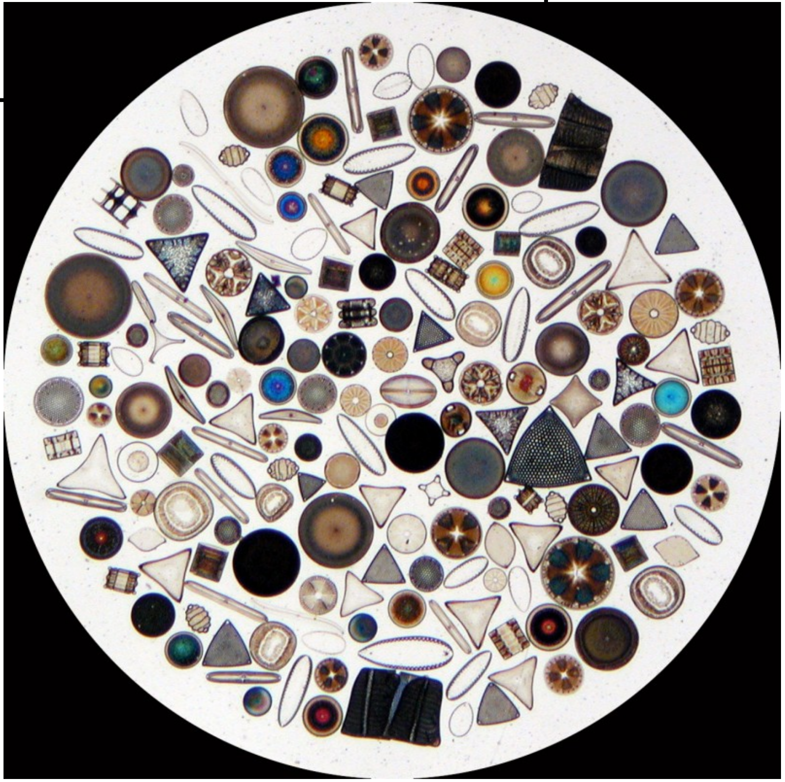
Repartition of assigned sequences per division



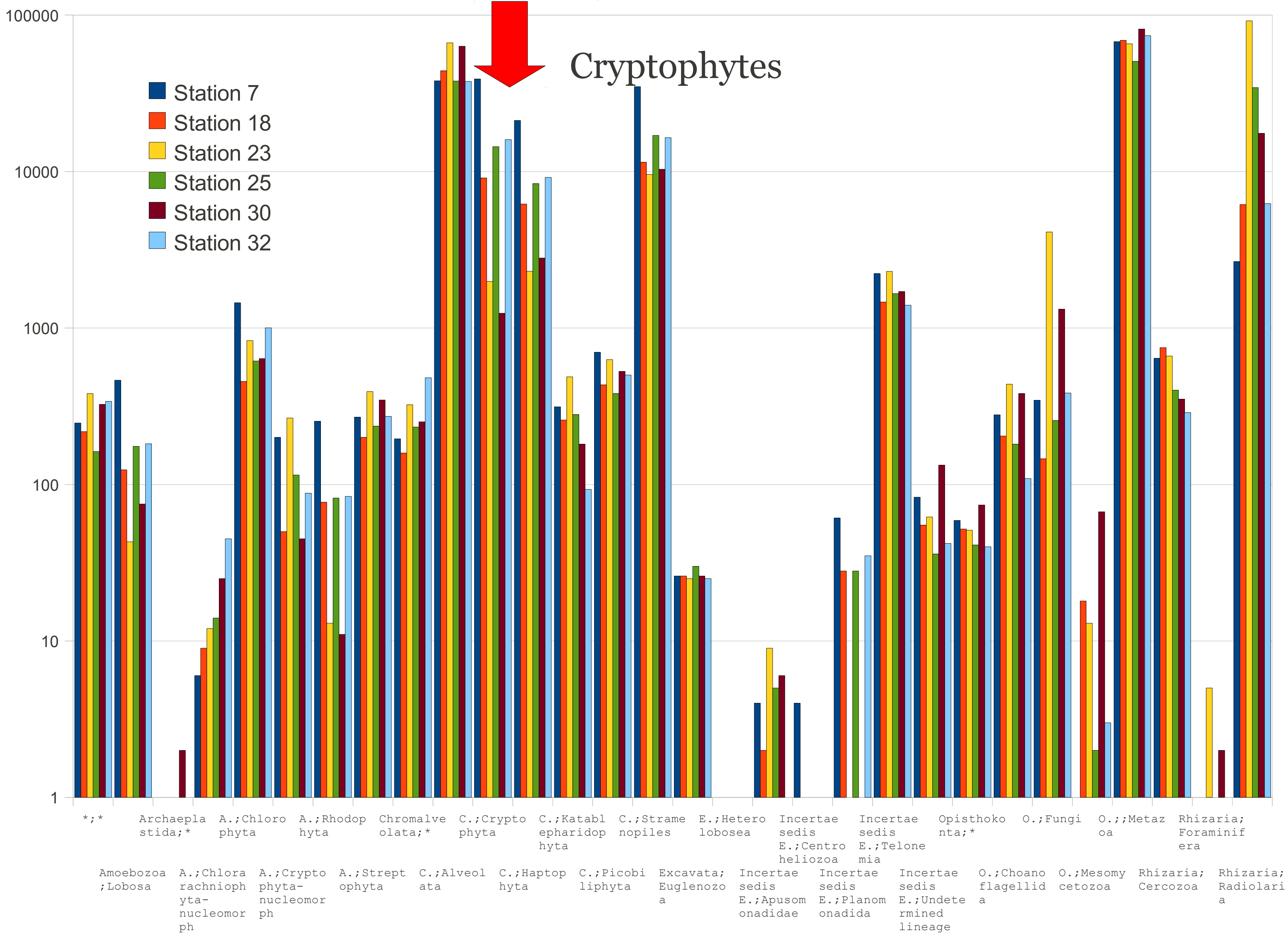
32264 Bacillariophyta Centric diatoms_Coscinodiscophyceae
16746 Pinguiophyceae Pinguiophyceae
9992 MAST-1 MAST-1
9934 Bacillariophyta Pennate diatoms
6232 MAST-3 MAST-3
4662 Chrysophyceae and Synurophyceae Chrysophyceae and Synurophyceae
4605 MAST-4 MAST-4
4229 * *
2545 Pelagophyceae Pelagophyceae
1295 MAST-9 MAST-9

Kunstformen der Natur (1900), Ernst Haeckel

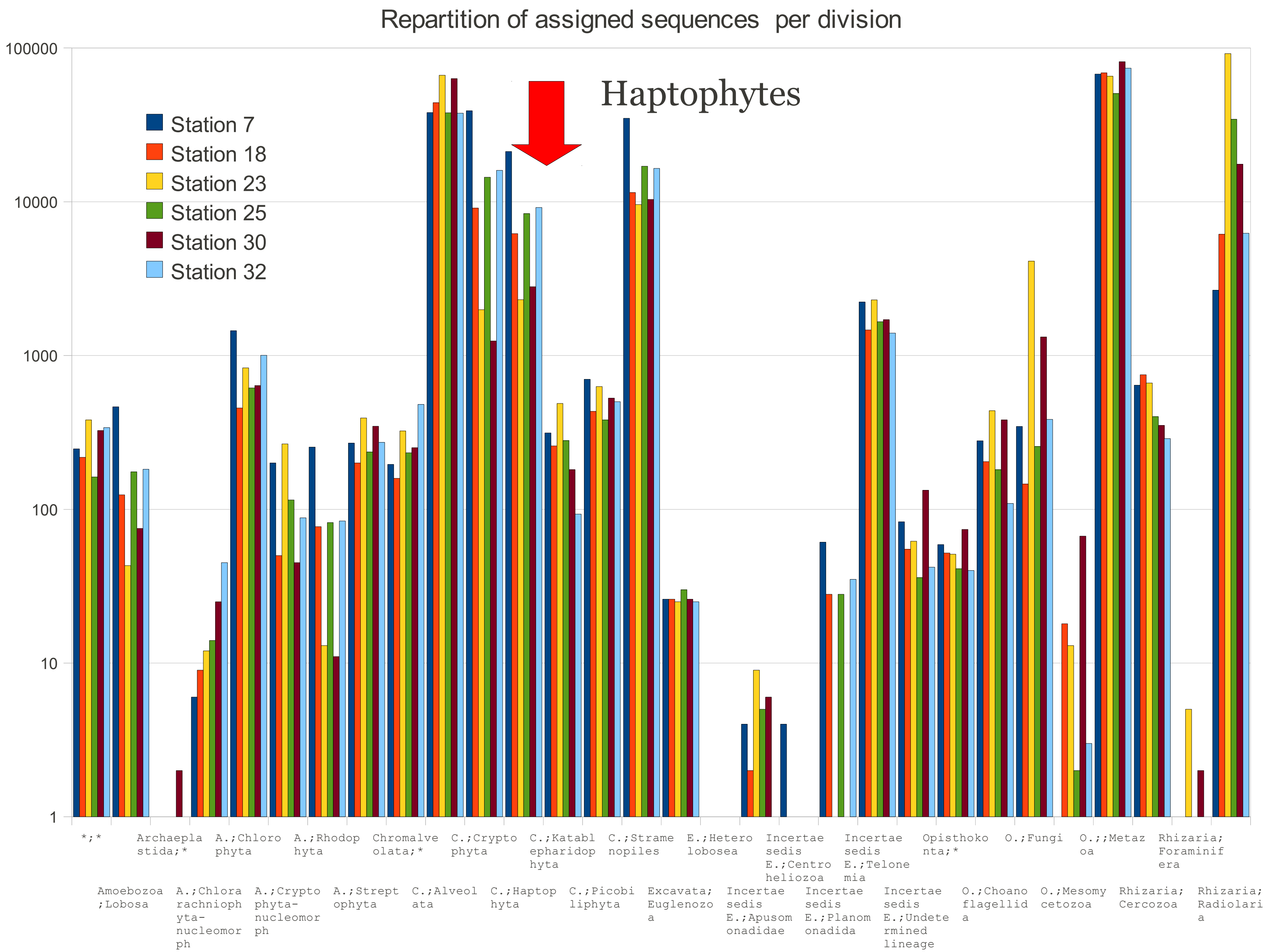




Repartition of assigned sequences per division

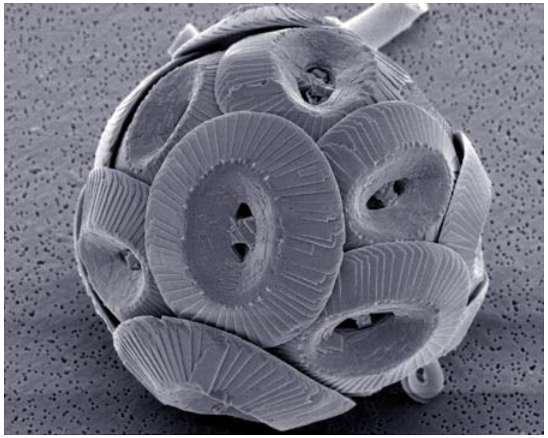


Cryptophytes

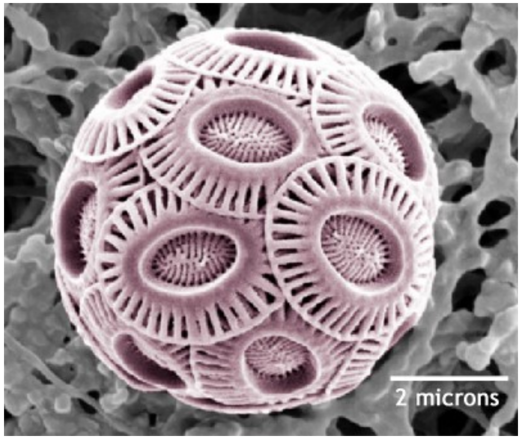


Haptophytes

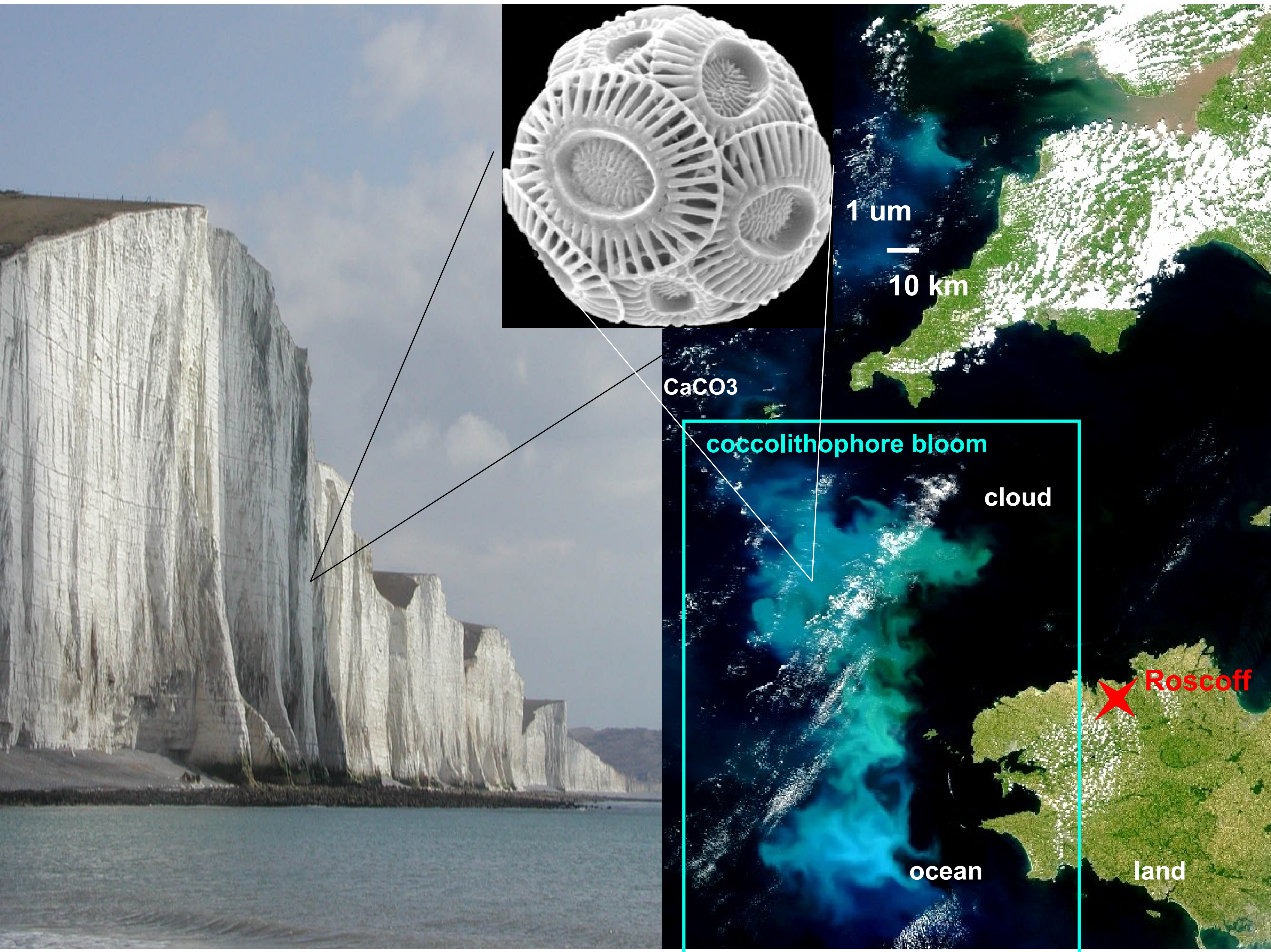
30052	Prymnesiophyceae Prymnesiophyceae *
12775	Pavlovophyceae Pavlovophyceae *
4253	Prymnesiophyceae Prymnesiophyceae Chrysochromulina group
1147	Prymnesiophyceae Prymnesiophyceae Pleurochrysis group
840	Prymnesiophyceae Prymnesiophyceae Emiliania and Gephyrocapsa group
245	Prymnesiophyceae Prymnesiophyceae Imantonia rotundata group
236	Prymnesiophyceae Prymnesiophyceae Haptophyta env. 3
169	Prymnesiophyceae Prymnesiophyceae Isochrysis group
122	Prymnesiophyceae Prymnesiophyceae Haptophyta env. 5
115	Prymnesiophyceae Prymnesiophyceae Braarudosphaera group

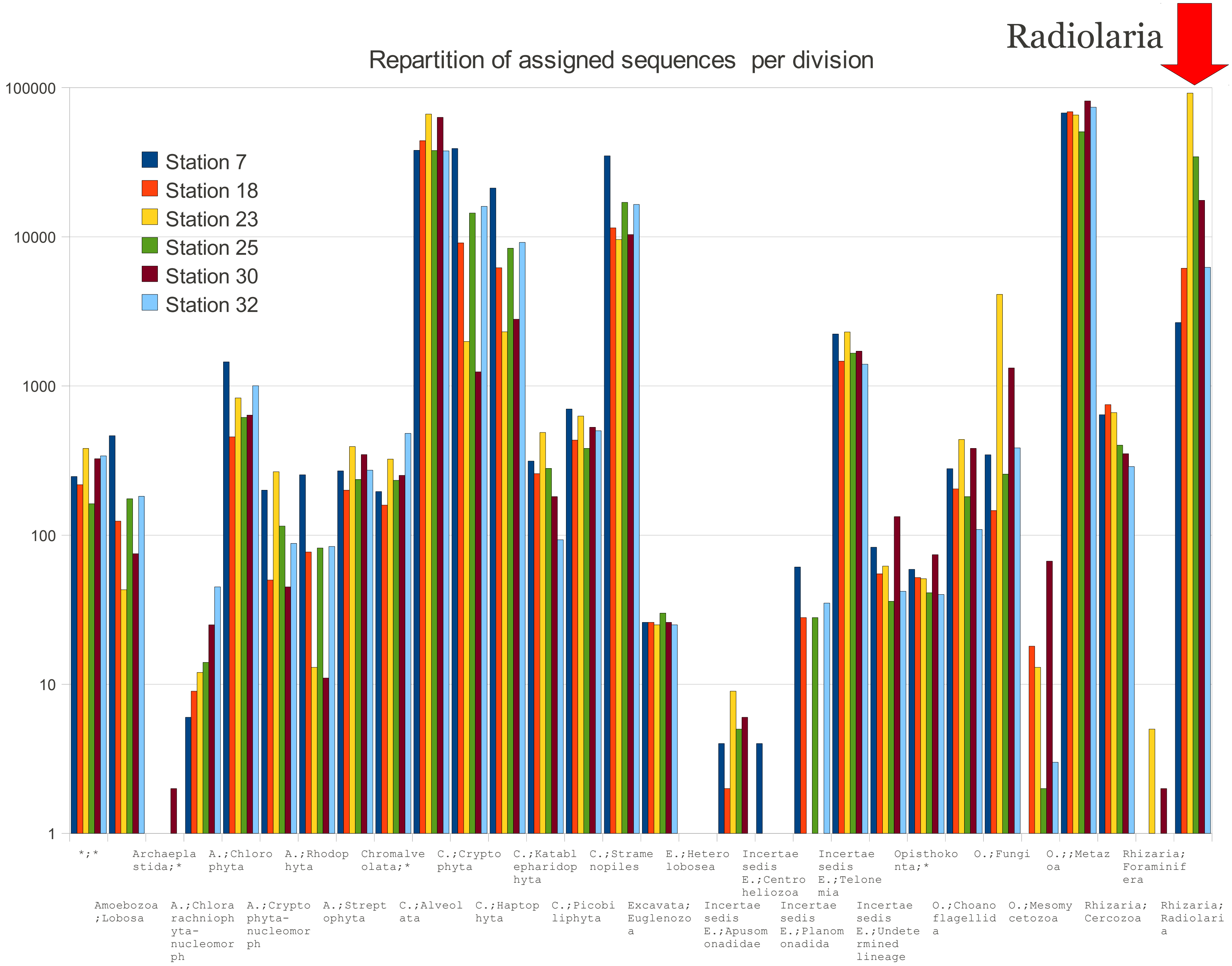


Coccolithus pelagicus



Emiliania huxleyi

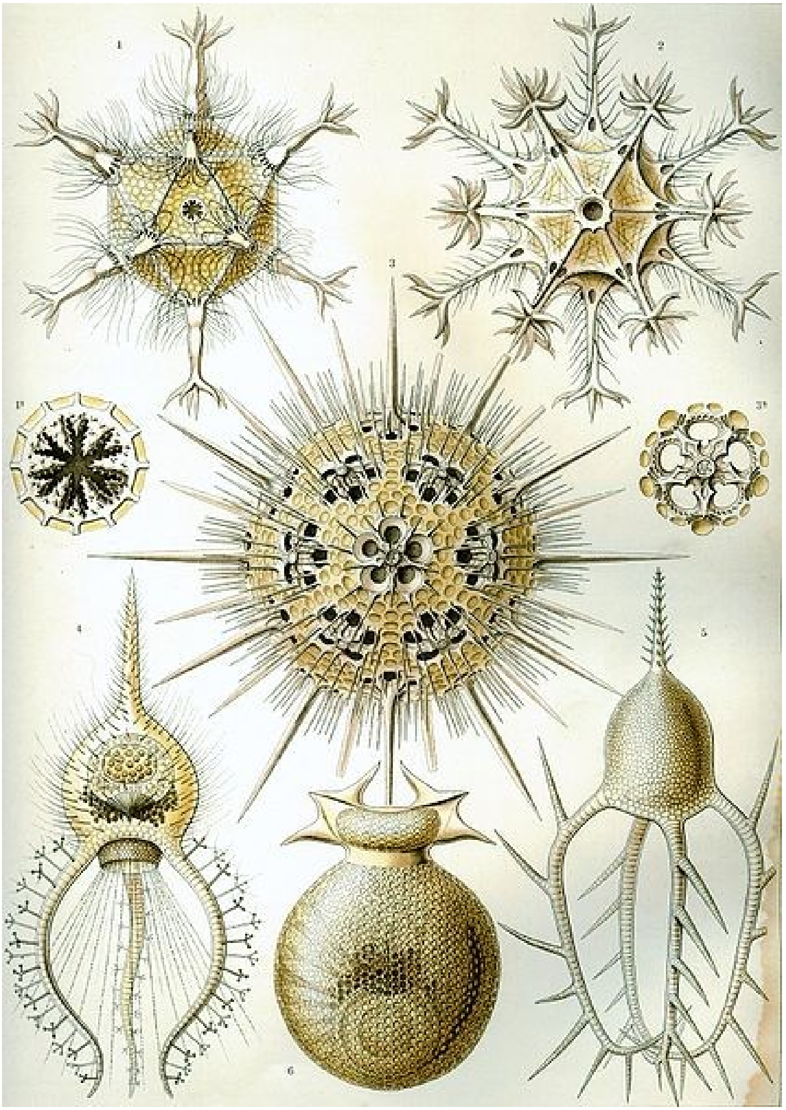

1 um
10 km
CaCO3
coccolithophore bloom
cloud
Roscoff
ocean
land


Repartition of assigned sequences per division
100000
10000
1000
100
10
1
Station 7
Station 18
Station 23
Station 25
Station 30
Station 32
;
Amoebozoa;Lobosa
Archaeplastida;*
A.;Chlorarachniophyta-nucleomorph
A.;Chlorophyta
A.;Cryptophyta-nucleomorph
A.;Rhodophyta
A.;Streptophyta
Chromalveolata;*
C.;Alveolata
C.;Cryptophyta
C.;Haptophyta
C.;Katablepharidophyta
C.;Picobiliphyta
C.;Stramenopiles
Excavata;Euglenozoa
E.;Heterolobosea
Incertae sedis E.;Apusomonadidae
Incertae sedis E.;Centroheliozoa
Incertae sedis E.;Planomonadida
Incertae sedis E.;Telonemia
Incertae sedis E.;Undetermined lineage
Opisthokonta;*
O.;Choanoflagellida
O.;Fungi
O.;Mesomycetozoa
O.;;Metazoa
Rhizaria;Cercozoa
Rhizaria;Foraminifera
Rhizaria;Radiolaria

Radiolaria

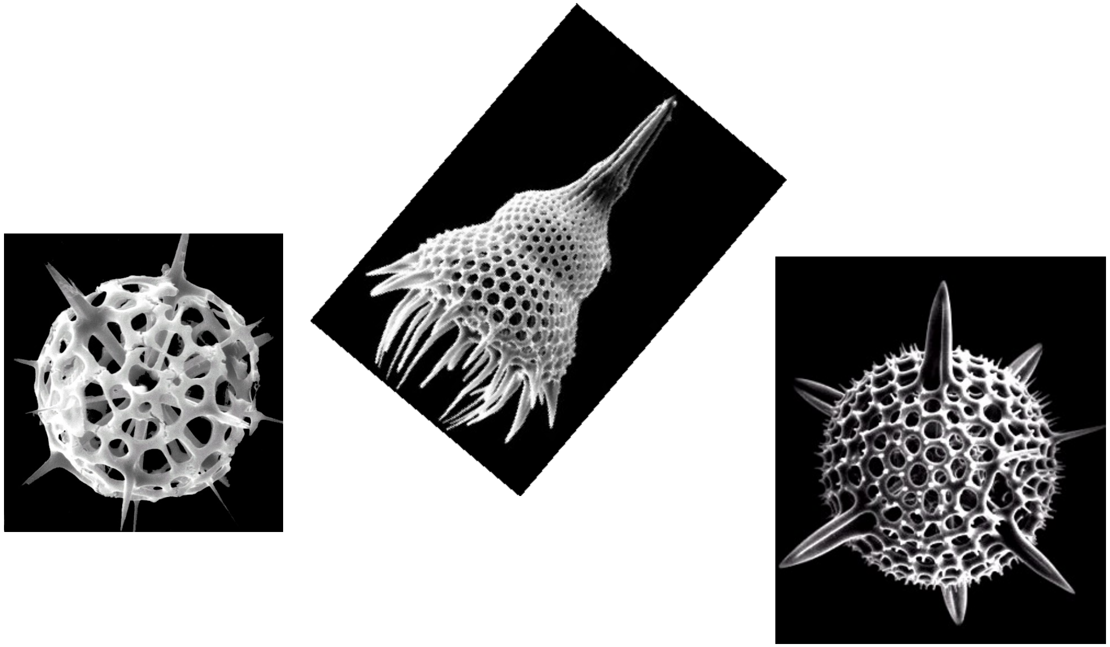
102088	Polycystinea Collodaria-Nassellarida
37543	Polycystinea Polycystinea
7022	RAD B RAD B
5094	Acantharea Acantharea
3778	RAD A RAD A
2912	Polycystinea Spumellarida
318	RAD C RAD C
82	Polycystinea *

Kunstformen der Natur (1900), Ernst Haeckel



Polycystinea -> siliceous skeleton

Acantharea->strontium sulfate skeleton

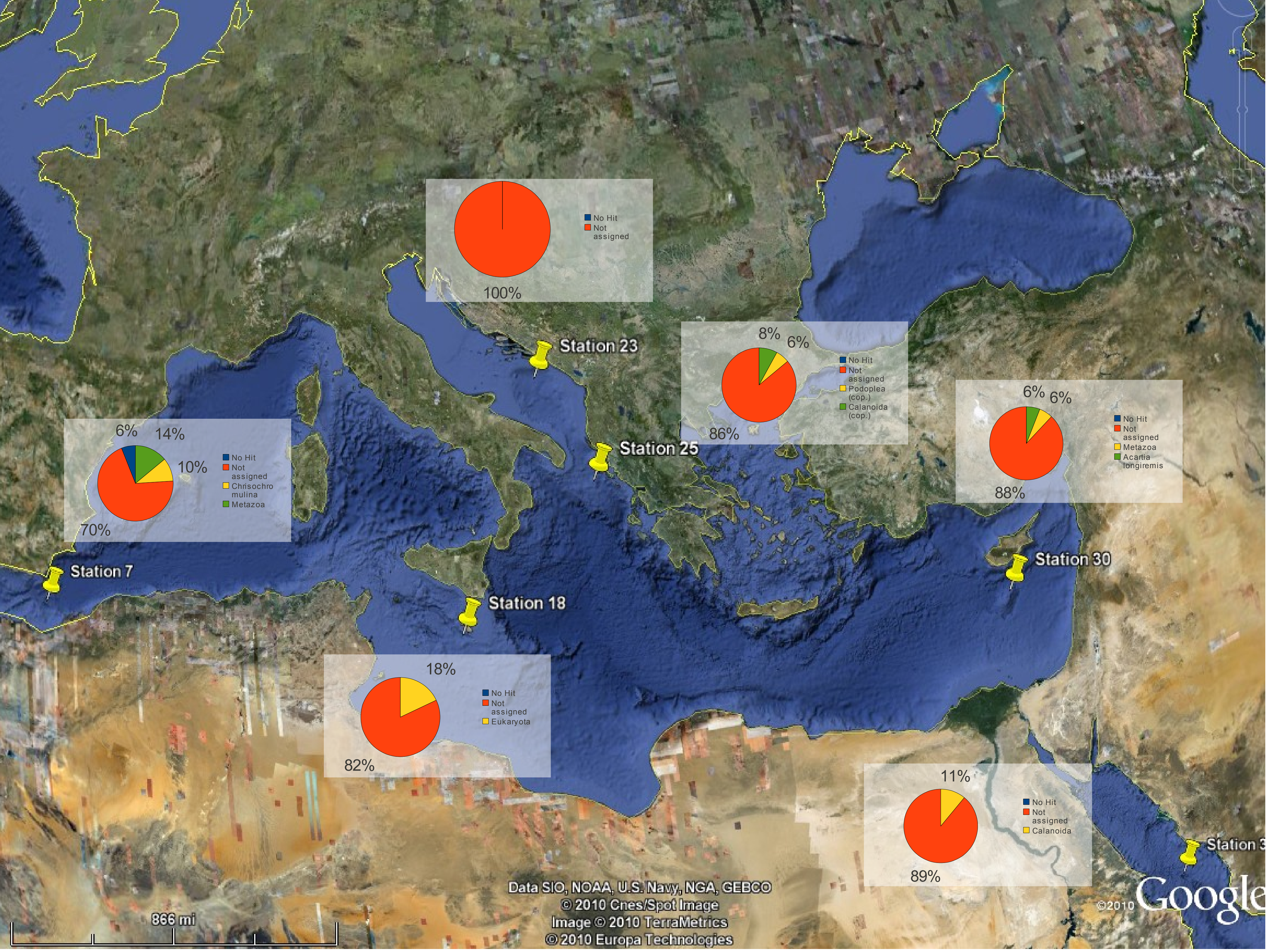


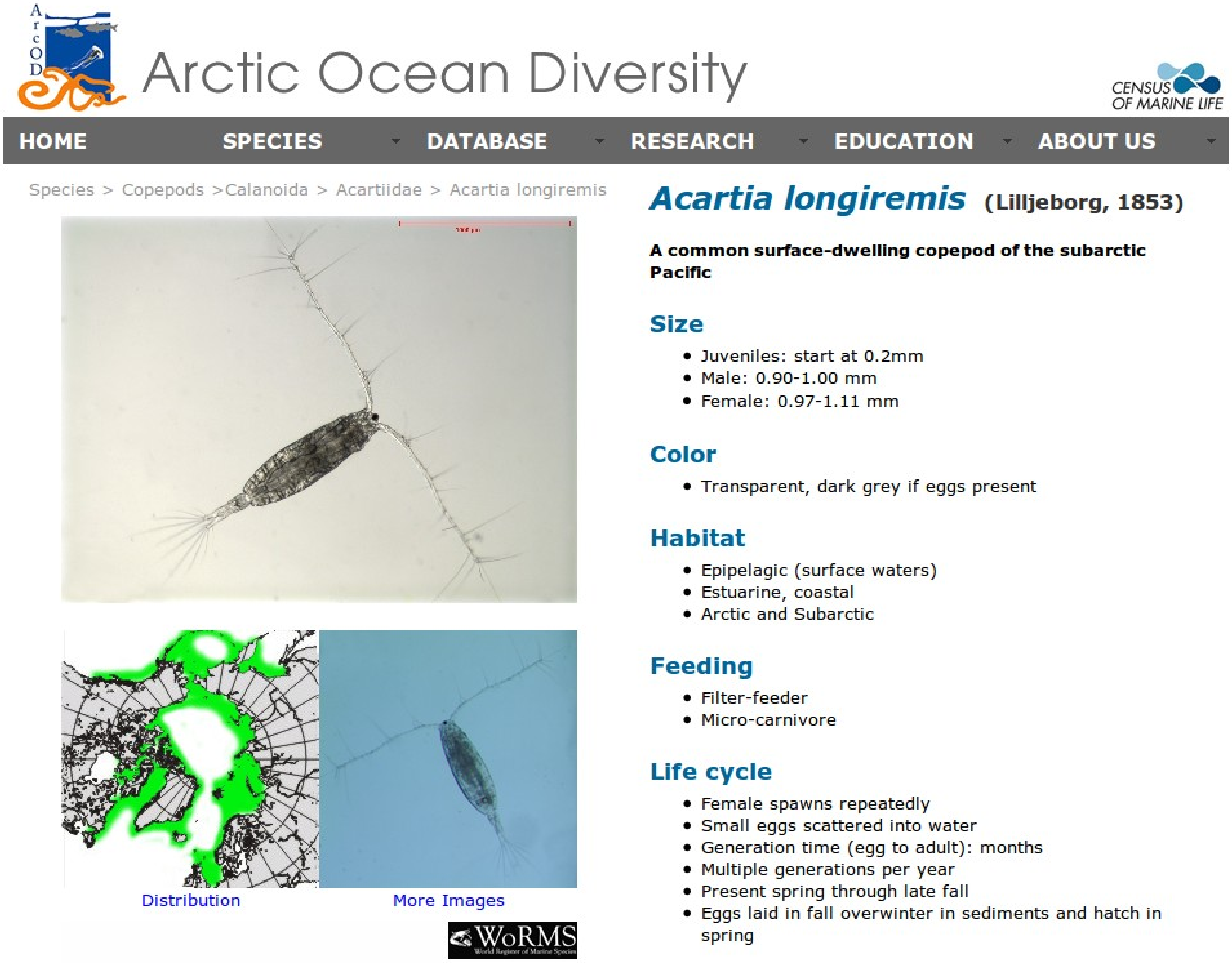
Discovering new species and/or improving reference databases

- Strategy:
  - For each Station, we extract the 100 more abundant sequences not assigned by our clustering based assignation method.

1) Blastn search against nucleotide database (ncbi nt)

2) Assignation with Megan (Metagenome Analysis Software)


100%
No Hit
Not assigned
Station 23
8%
6%
86%
No Hit
Not assigned
Podoplea (cop.)
Calanoida (cop.)
Station 25
6%
6%
88%
No Hit
Not assigned
Metazoa
Acartia longiremis
6%
14%
10%
70%
No Hit
Not assigned
Chrisochromulina
Metazoa
Station 7
Station 18
Station 30
18%
82%
No Hit
Not assigned
Eukaryota
11%
89%
No Hit
Not assigned
Calanoida
Station 3
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2010 Cnes/Spot Image
Image © 2010 TerraMetrics
© 2010 Europa Technologies
©2010 Google
866 mi



Acartia longiremis (Lilljeborg, 1853)

A common surface-dwelling copepod of the subarctic Pacific

Size

- Juveniles: start at 0.2mm
- Male: 0.90-1.00 mm
- Female: 0.97-1.11 mm

Color

- Transparent, dark grey if eggs present

Habitat

- Epipelagic (surface waters)
- Estuarine, coastal
- Arctic and Subarctic

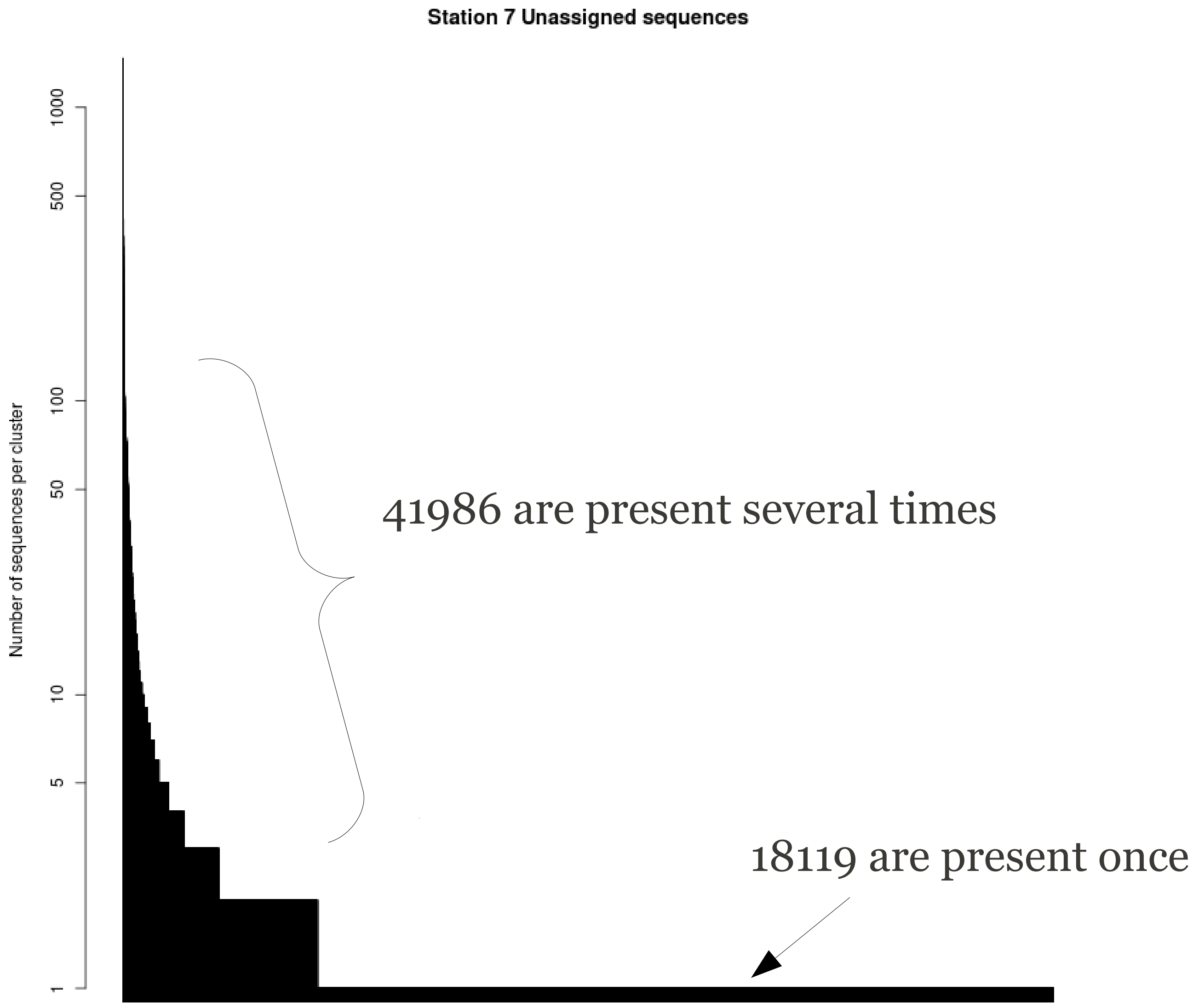
Feeding

- Filter-feeder
- Micro-carnivore

Life cycle

- Female spawns repeatedly
- Small eggs scattered into water
- Generation time (egg to adult): months
- Multiple generations per year
- Present spring through late fall
- Eggs laid in fall overwinter in sediments and hatch in spring

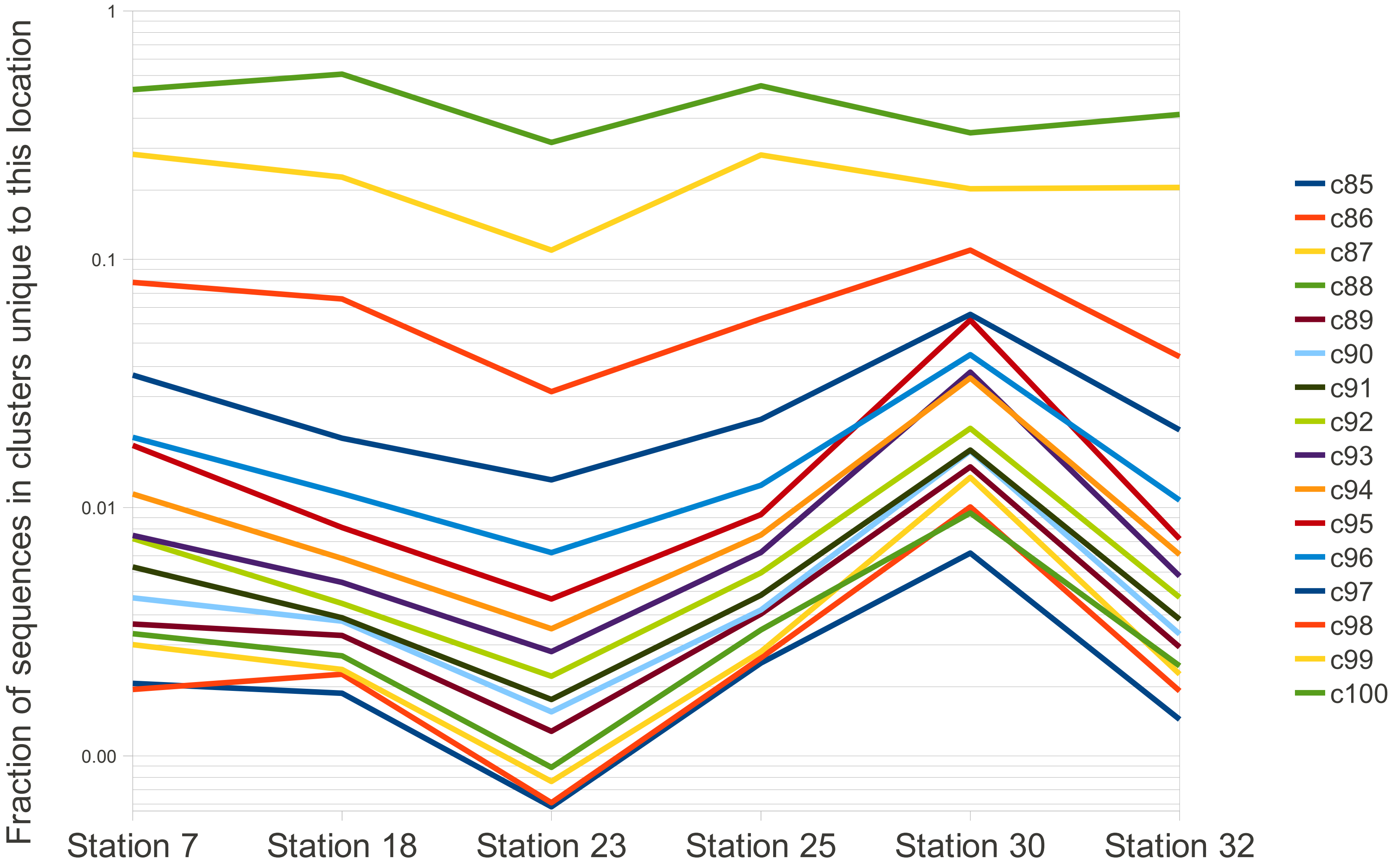
What are unassigned sequences ?



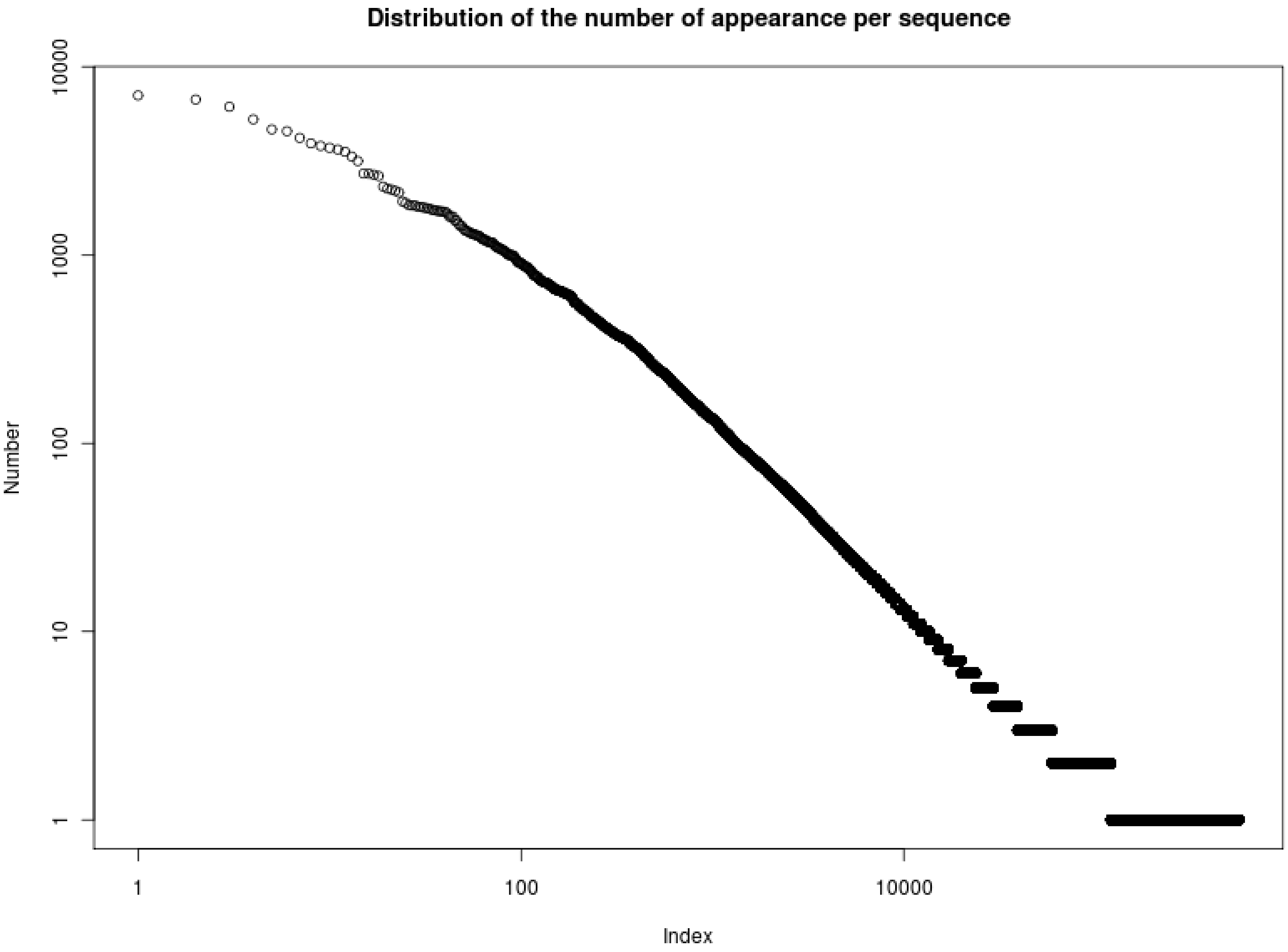
41986 are present several times

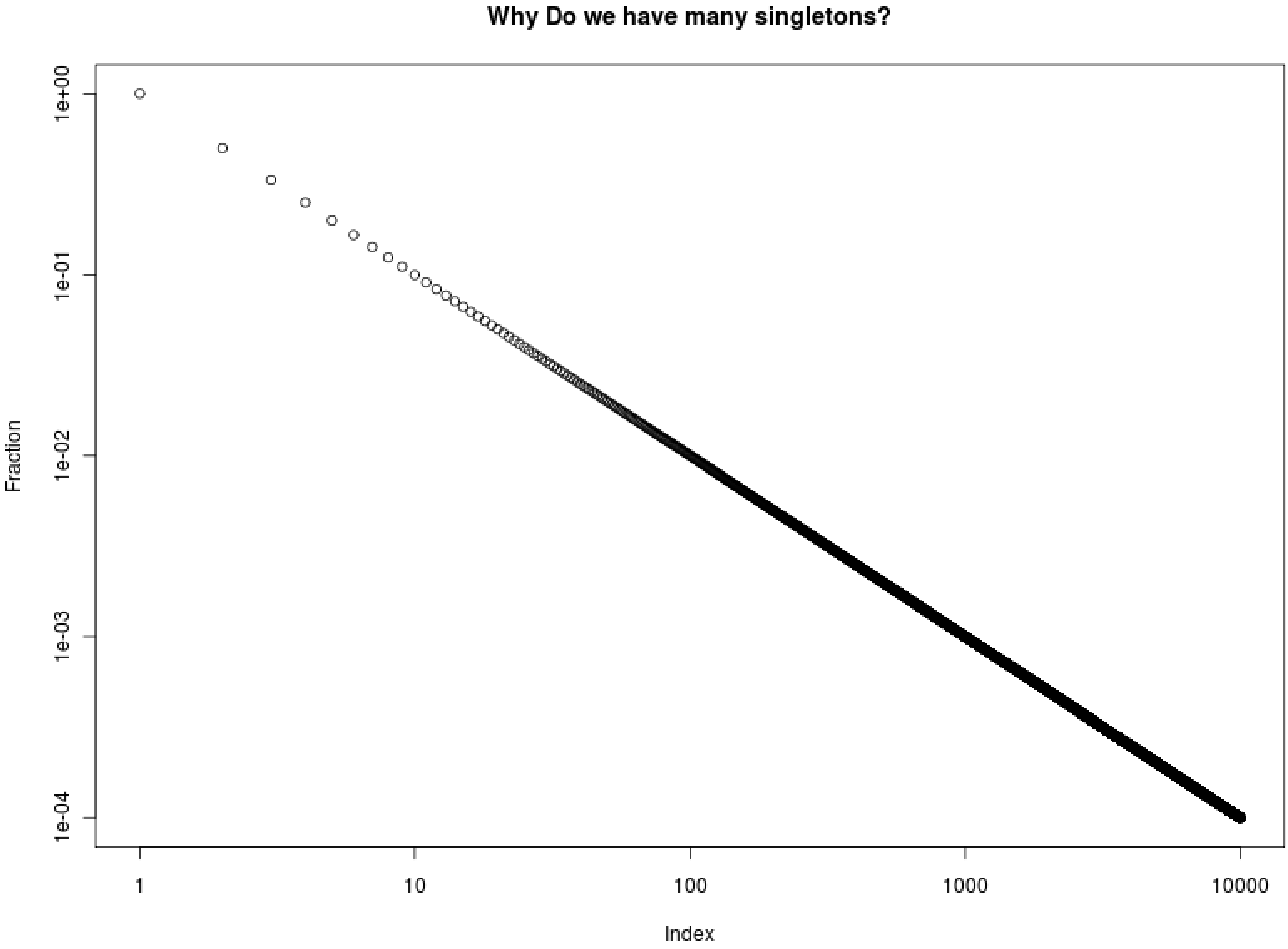
18119 are present once

Large overlap between stations: not so many specific clusters

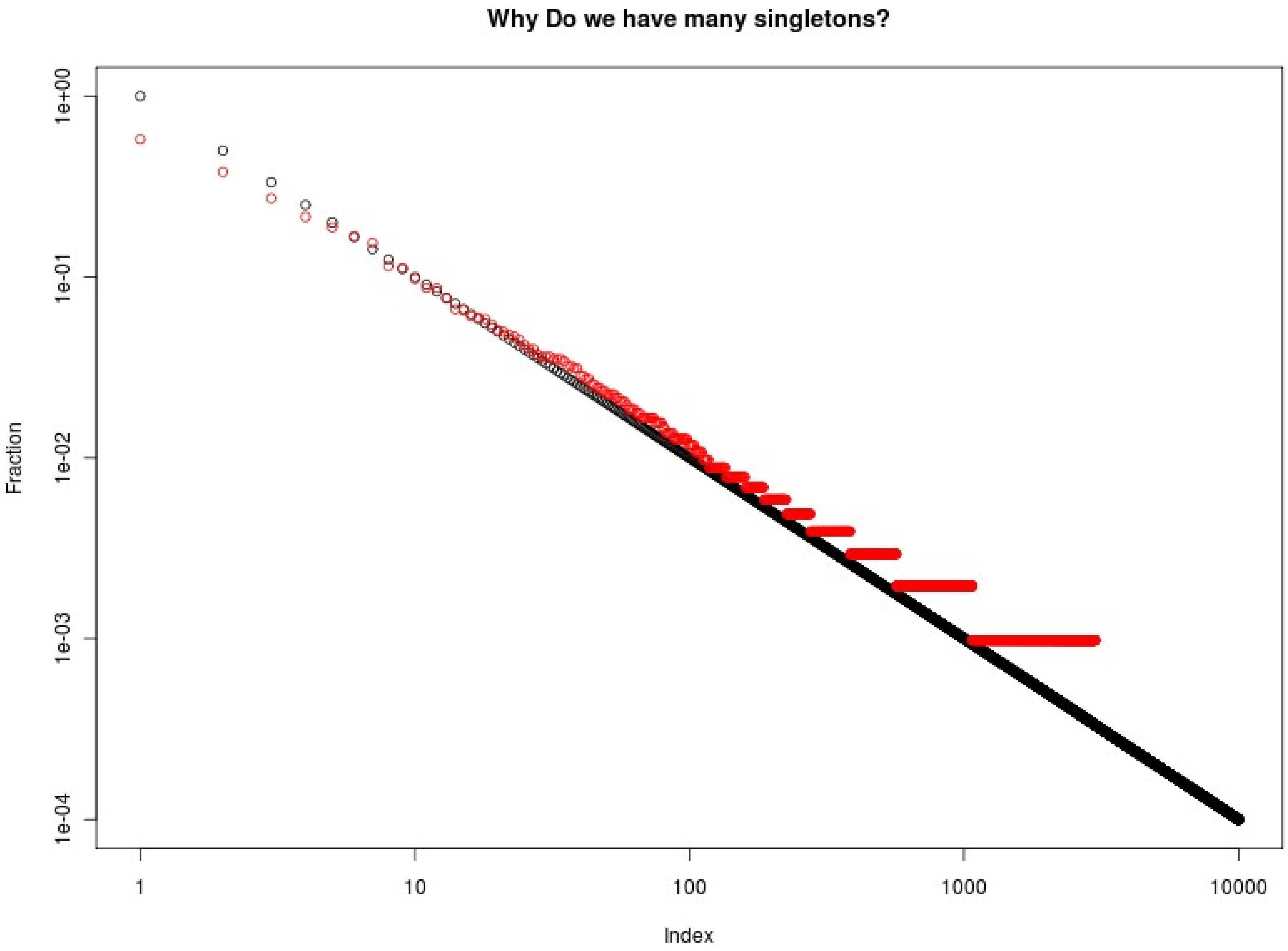


Number of Appearance per sequence in Real Tara V9 Barcode Data





Why Do we have many singletons?


Why Do we have many singletons?
Fraction
1e+00
1e-01
1e-02
1e-03
1e-04
Index
1
10
100
1000
10000

Thanks

- Team Evolution du Plancton et PaléOcéan: (resp. Colomban de Vargas):

Fabrice Not , Lucie Bittner, Frédéric Mahé, Christophe Boutte, Nathalie Babic, Sarah Romac, Sébastien Colin, El Mahdi Bendif, Johan Decelle, Noan Le Bescot, Céline Dimier

- Adaptation et Diversité en Milieu Marin, UMR7144, (resp. Francois Lallier)
- Station Biologique de Roscoff (resp. Bernard Kloareg)