04/10/2014

Bioinformatique microbienne

IFB & FG variants

Marine microbial and metagenomic developement : *Data & Tools*

Erwan Corre

ABiMS

Context

IFB-Grand Ouest Platform

Team leader : C. Caron

- E-infrastructure : HPC environment (600 core cluster, 500 TB storage) and Bio analysis tools, Galaxy
- Expertise in the analysis of omics data
- Software engineering (information systems, databases, etc.)
- Training modules

Research team

Marine and environmental Biology

- Tools and methodology developments ,
- Research projects in marine genomics and metagenomics

Research projects and developments

Metagenomics

- Carnivorous sponge microflora (col. Museum)
- Abalone stomach microflora (PIA Idealg)
- Macroalgae epibiont microflora (PIA Idealg)
- Meta-barcoding , -genomic, -transcriptomic analysis "Post TARA" (PIA OCEANOMICS).

- Metagemonic on thermophilic hydrothermal vent flora (LM2E : Lois Maignien)
- Metagenomics of microbial flora associated with the methane cycle in marine sediments on continental margins (IFREMER).

Comparative Microbial genomics

- Marine Cyanobacteria (L. Garczareck)
- Marine vibrio (F. Le Roux)
- Marine Flavobacteria (G. Michel)

Developments

DATA

- Comparative genomic database

TOOLS

- Genome scaffolding tool
- Metabarcoding pipeline

Data : comparative genomique database

Cyanorak v.2, an information system dedicated to the expert curation and annotation of clusters of orthologous sequences from marine picocyanobacteria

Antoine Bisch, Loraine Guéguen, Gregory Farrant, Mark Hoebeke, Gildas Le Corguillé, Erwan Corre, Wilfrid Carré, Christophe Caron, Laurence Garczarek and Frédéric Partensky

Marine Phototrophic Prokaryotes (MaPP) Team, Phytoplankton Group, UMR7144 CNRS-UPMC, Station Biologique, Roscoff
Analysis and Bioinformatics for Marine Science (ABiMS), FR2424 CNRS, Station Biologique, Roscoff

Cyanorak database

Cyanorak

- Storage and visualization of picocyanobacteria sequences
- Clusters of orthologous groups
- Aims: Curation, annotation and Export

History

- ***Cyanorak v.1** (2005) : 14 genomes (11 Syn., 3 Proc.)*
  - Manual annotation of a fixed set of genomes
- ***Cyanorak v.2** (2013): 57 genomes (40 Syn., 3 Cya., 14 Proc.)*
  - Genome sequence origin:
    - Cyanorak v.1: 11 *Syn,* 3 *Proc.*
    - NCBI/RefSeq: 4 *Syn.,* 2 *Cya.*, 11 *Proc.*
    - Newly sequenced genomes : 25 *Syn.*, 1 *Cya.*
  - Import and semi-automatic annotation of new sets of genomes

Database Contents

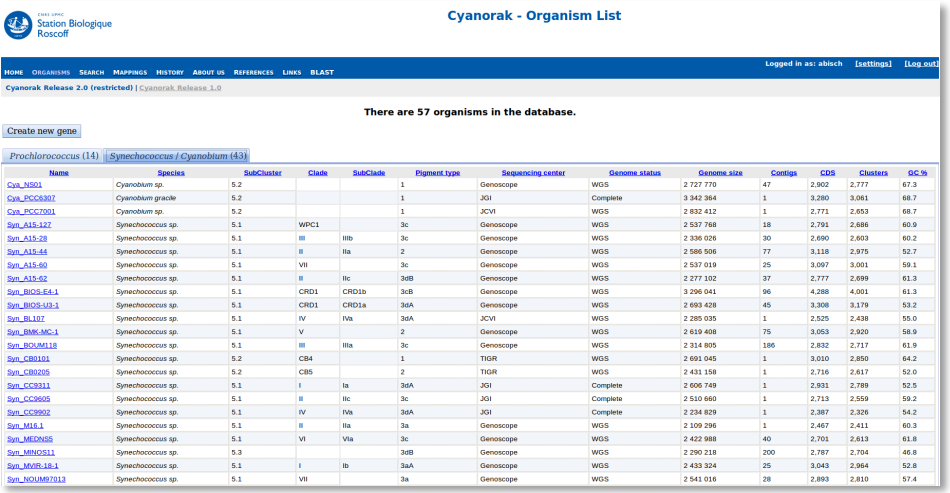
Organisms	14
Genes	35,993
Clusters	7,847

Database Contents

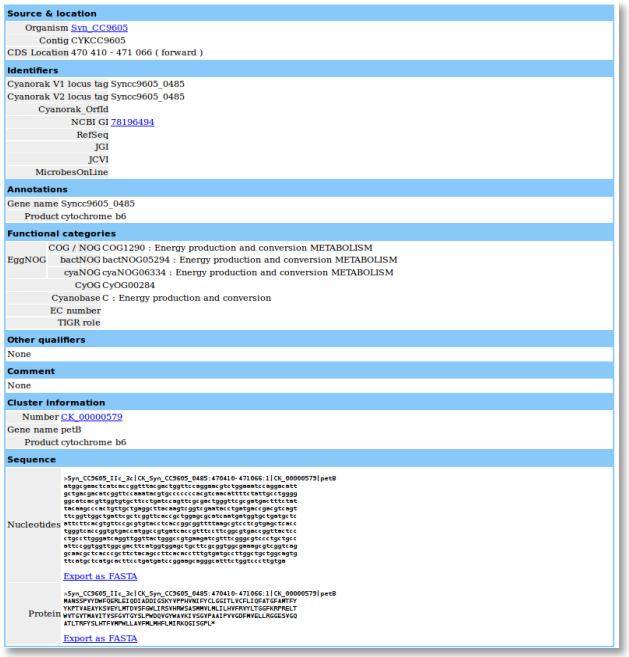
Organisms	57
Genes	154,723
Clusters	24,415

Online interface:

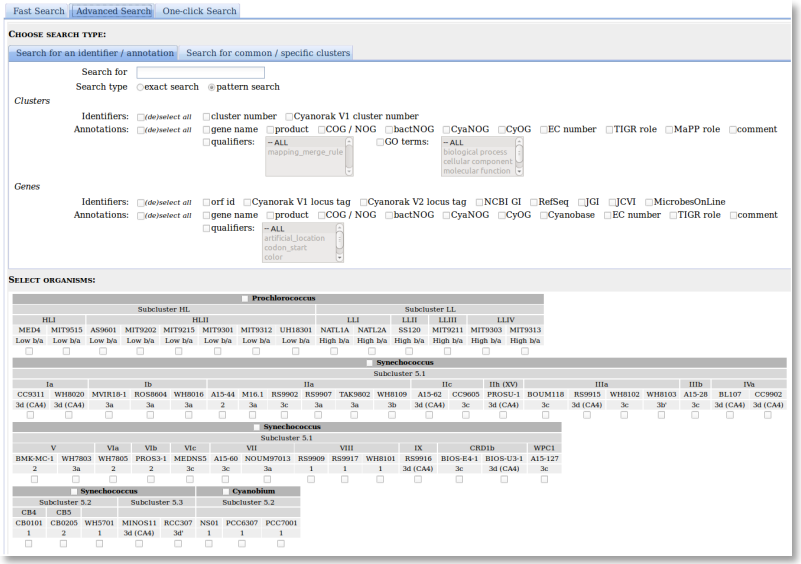
Organisms page



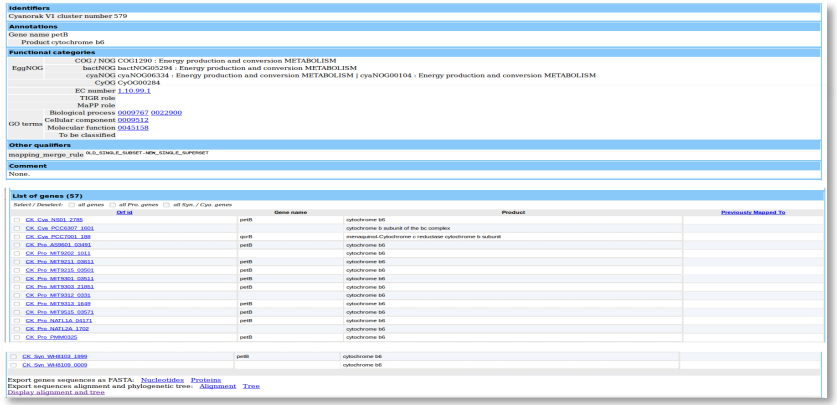
Gene page

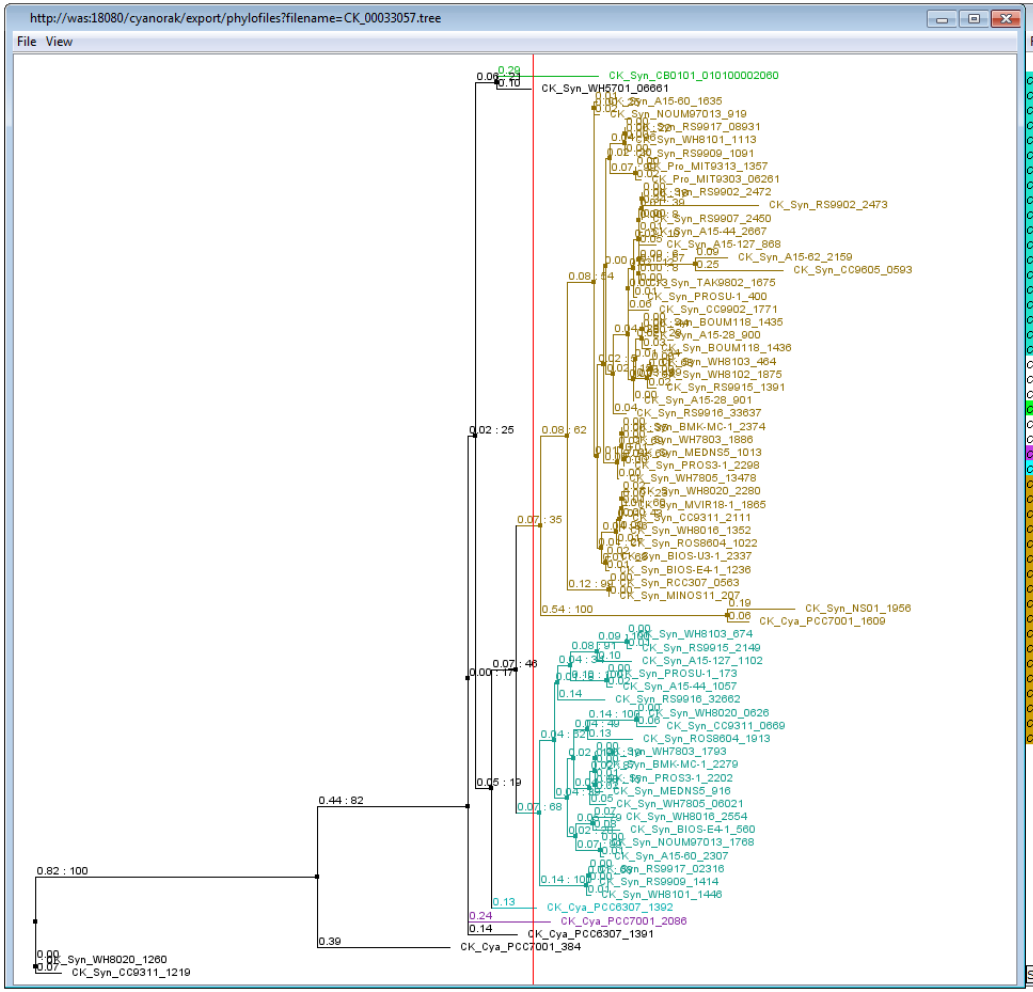


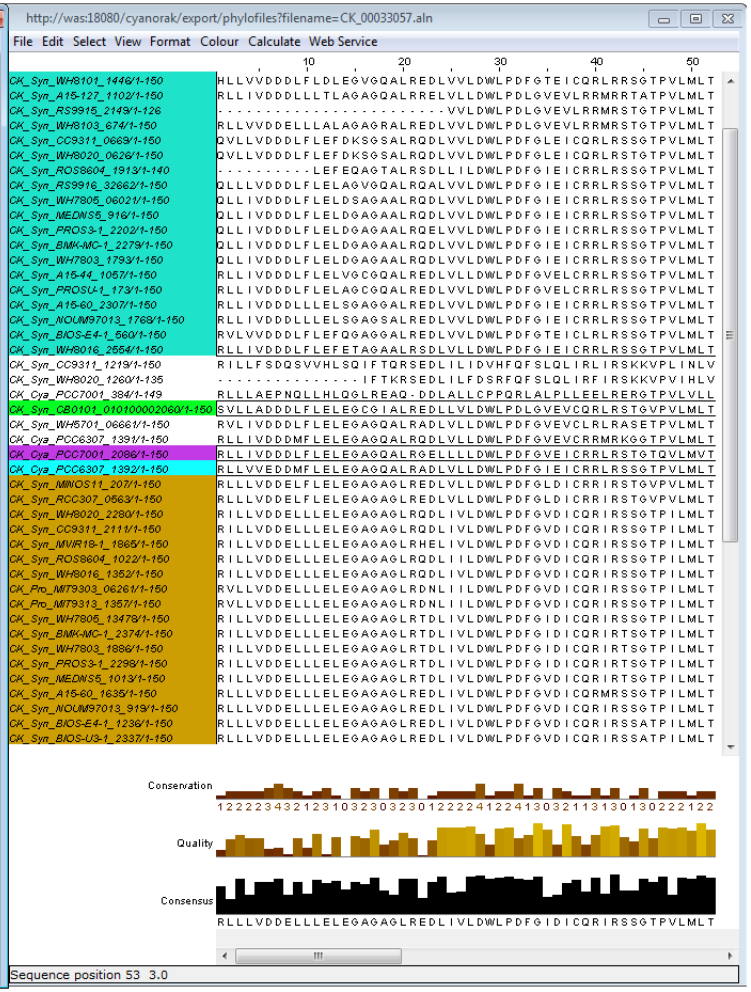
Search tool



Cluster page



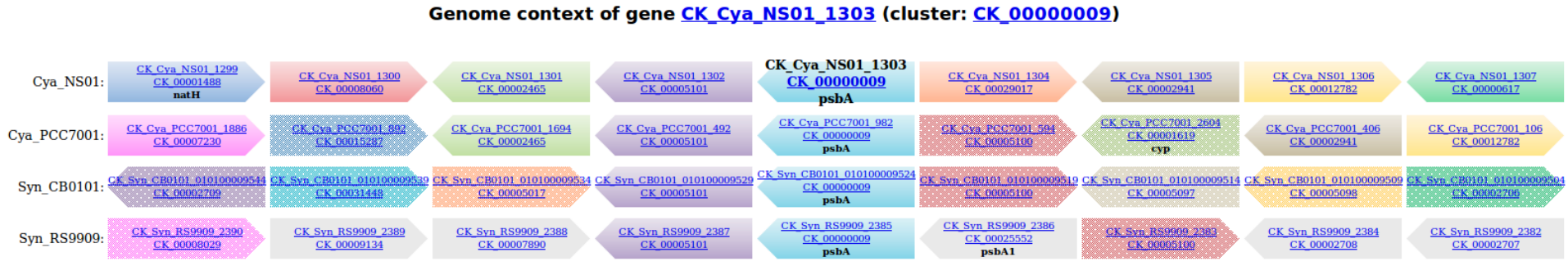




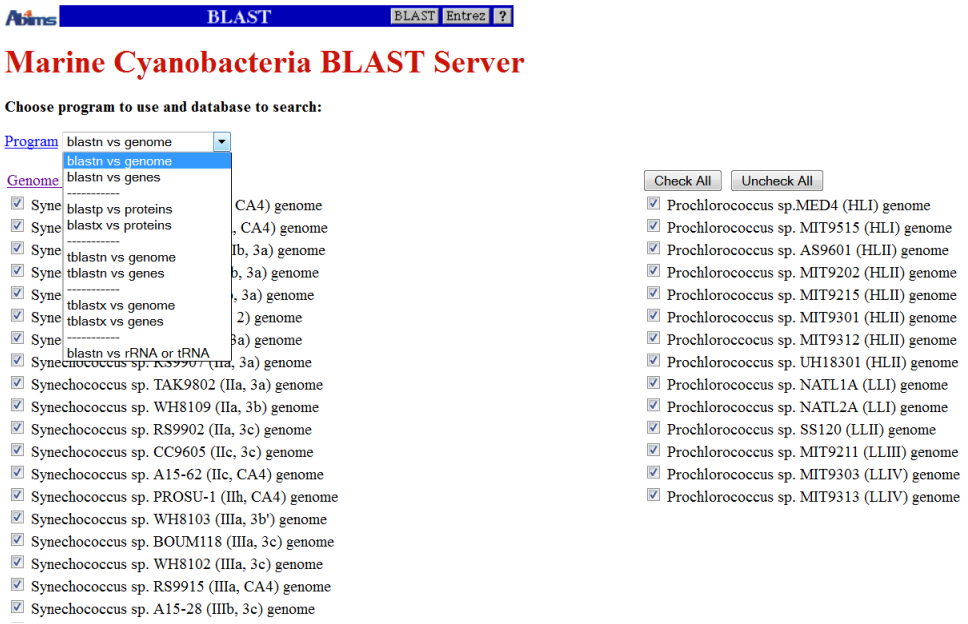
PhyML tree

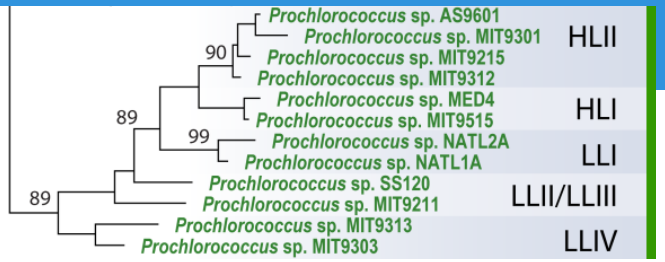
MAFFT Alignment

- # Genome context



- # Local BLAST





Scanlan et al., MMBR, 2009

Phyletic pattern:

- core/accessory/specific genes
- genes involved in adaptation to specific environmental niches

Phyletic pattern

Prochlorococcus													
Subcluster HL								Subcluster LL					
HLI		HLII						LLI		LLII	LLIII	LLIV	
MED4	MIT9515	AS9601	MIT9202	MIT9215	MIT9301	MIT9312	UH18301	NATL1A	NATL2A	SS120	MIT9211	MIT9303	MIT9313
Low b/a	Low b/a	Low b/a	Low b/a	Low b/a	Low b/a	Low b/a	Low b/a	High b/a	High b/a	High b/a	High b/a	High b/a	High b/a
1	1	1	1	1	1	1	1	1	1	0	0	0	0

Synechococcus																				
Subcluster 5.1																				
Ia		Ib			IIa						IIc		IIh	IIIa				IIIb	IVa	
CC9311	WH8020	MVIR-18-1	ROS8604	WH8016	A15-44	M16.1	RS9902	RS9907	TAK9802	WH8109	A15-62	CC9605	PROS-U-1	BOUM118	RS9915	WH8102	WH8103	A15-28	BL107	CC9902
3dA	3dA	3aA	3a	3aA	2	3a	3c	3a	3a	3bB	3dB	3c	3dB	3c	3dB	3c	3bB	3c	3dA	3dA
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

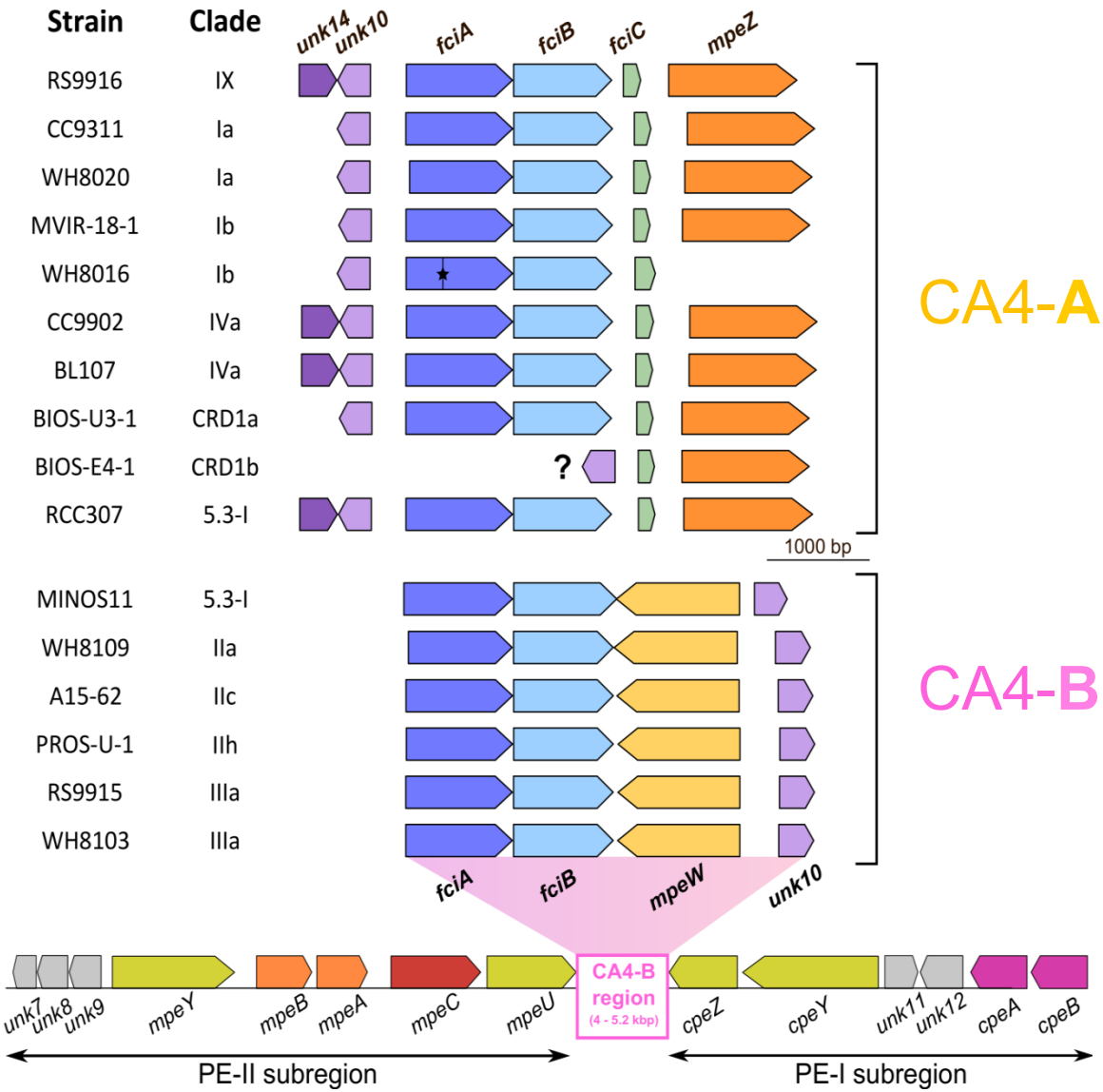
Synechococcus													
Subcluster 5.1													
V		VIa		VIb	VII		VIII			IX	CRD1a	CRD1b	WPC1
BMK-MC-1	WH7803	MEDNS5	WH7805	PROS-7-1	A15-60	NOUM97013	RS9909	RS9917	WH8101	RS9916	BIOS-U3-1	BIOS-E4-1	A15-127
2	3a	3c	2	2	3c	3a	1	1	1	3dA	3dA	3cB	3c
1	1	1	1	1	1	1	1	1	1	1	1	1	1

Synechococcus					Cyanobium		
Subcluster 5.2			Subcluster 5.3		Subcluster 5.2		
CB4	CB5						
CB0101	CB0205	WH5701	MINOS11	RCC307	NS01	PCC6307	PCC7001
1	2	1	3dB	3eA	1	1	1
1	1	1	1	1	1	1	1

DNA photolyase *phrA* gene absent in strickly LL *Prochlorococcus* strains

List of genes (53)

The specific island of chromatic adaptors (CA4)



MpeZ/MpeW
Phycobilin lyases

FciA/B
Transcriptional Activators
AraC/XylS family

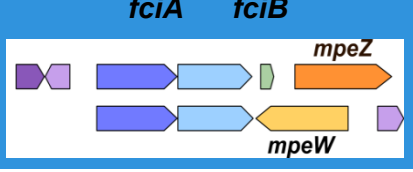
FciC
Transcriptional repressor
found in phage and bacteria

Phycobilisome rod region

Humily et al., 2013

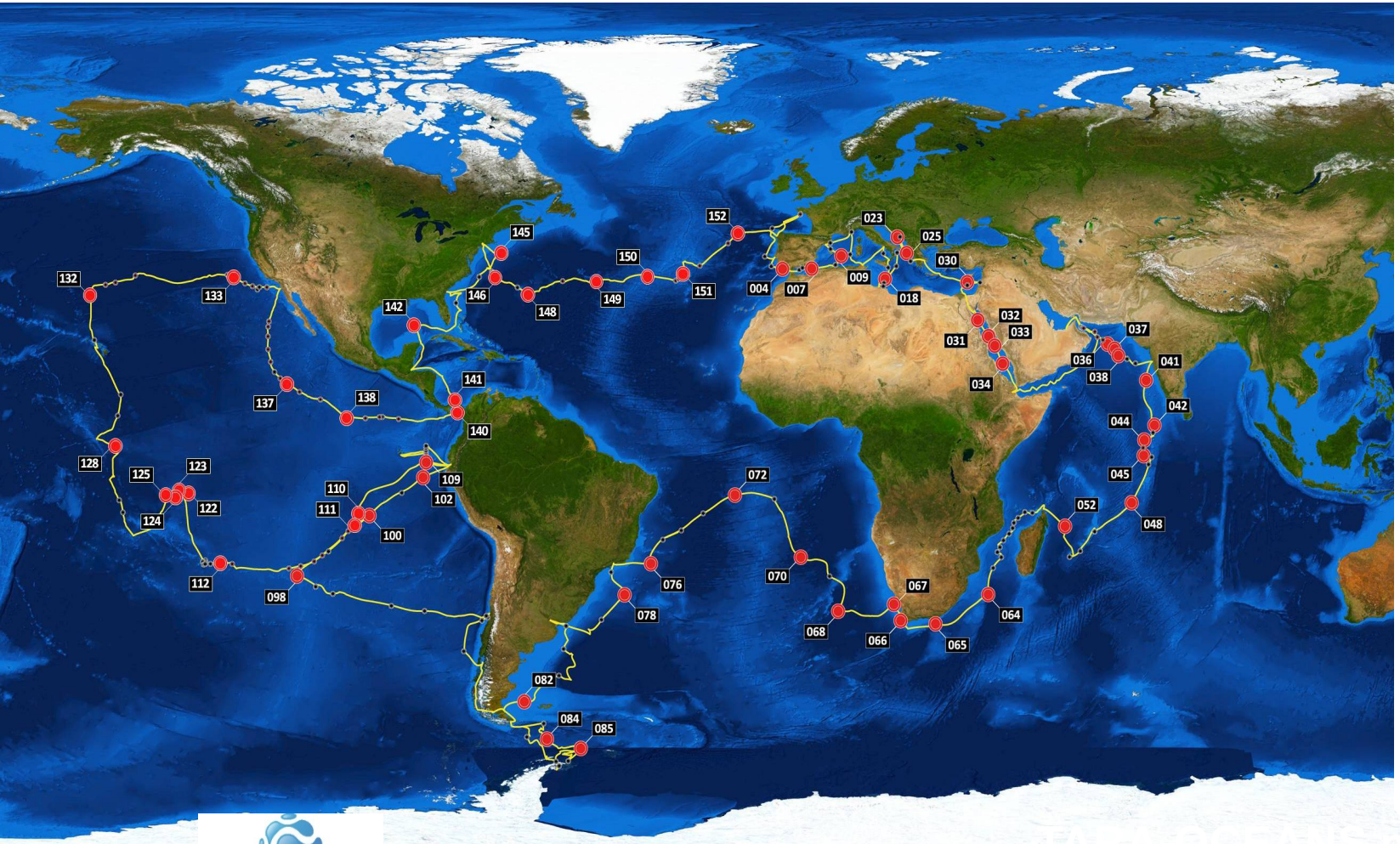
Geographical Distribution of CA4

CA4-A
CA4-B



- *mpeW / mpeZ* recruitment -> Abundance of CA4-A and B
- *petB* recruitment -> Abundance of *Synechococcus*

=> Relative proportion of CA4-A and B



Metagenomes
~5.10^6 reads
(~175 bp)

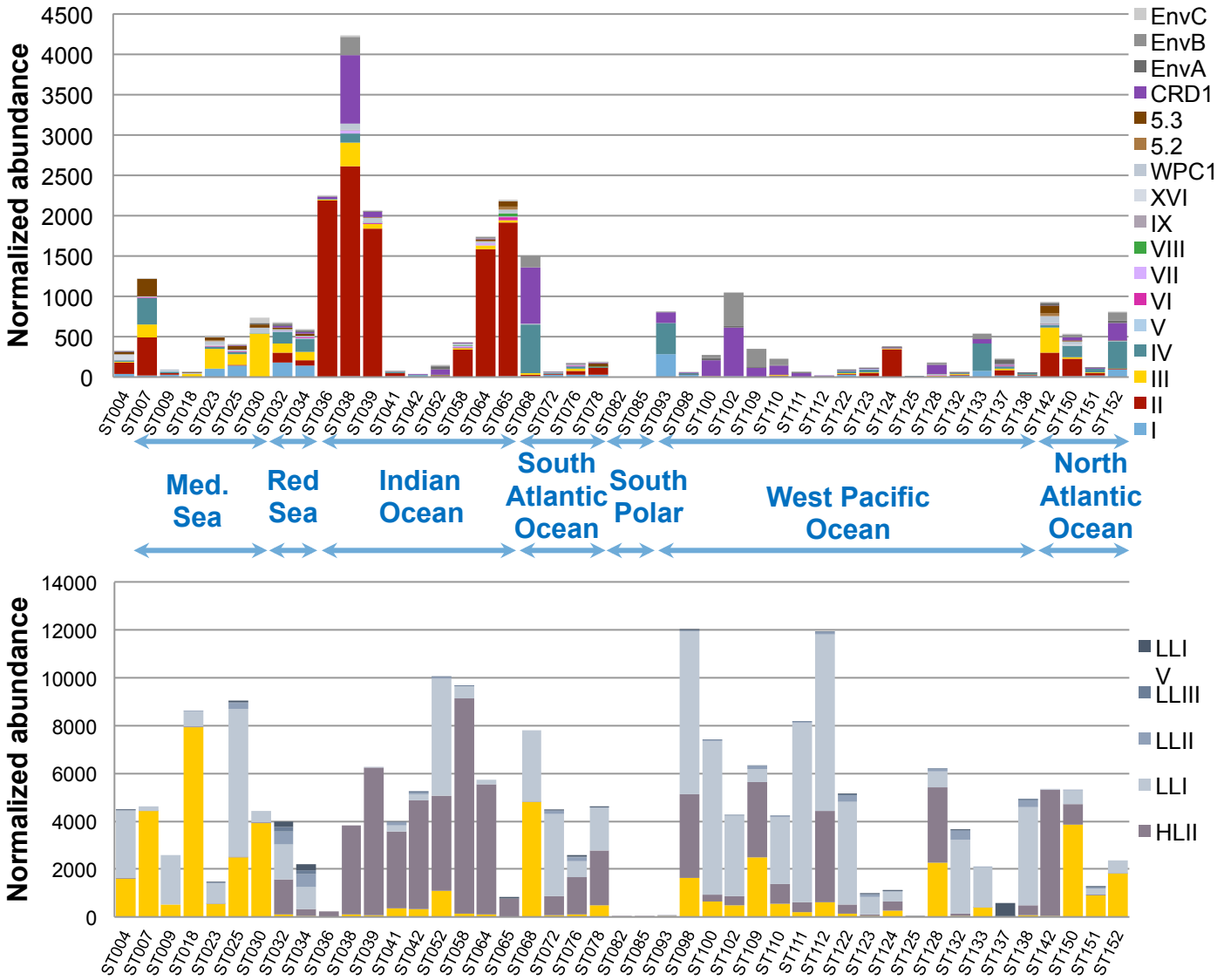
BLASTX alignment
vs 2 *petB*

BLASTN alignment
vs 280 *petB*

Taxonomic Assignement
Abundance profiles

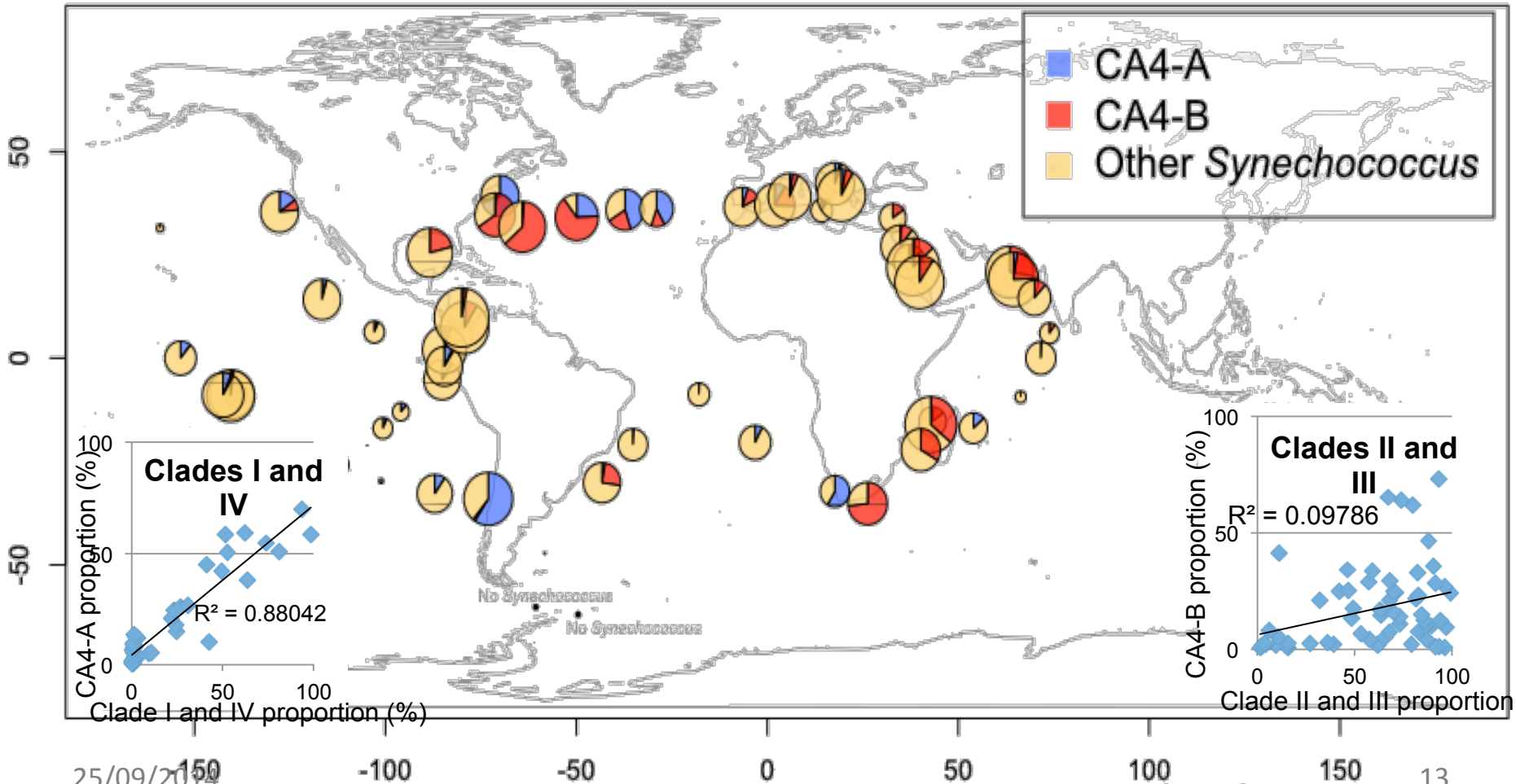
OCEANOMICS

Abundance of *Synechococcus* (A) and *Prochlorococcus* (B) clades in DCM stations of the TARA Oceans metagenomes (small size fraction) based on *petB* miTAG recruitments



Geographical Distribution of CA4

- Abundant in coastal region / Temperate oceanic waters
  - Until 90% of *Synechococcus* populations
- Most clades I and IV are CA4
- Only some clades II and III display this capacity



Farrant, Doré and Garczarek, unpublished

The future of Cyanorak (v.3):

- More data type stored/visualized
  - Additional sequences types (rRNA, ncRNA, tRNA, UTR)
  - Transcriptomic data (Arrays and RNAseq)

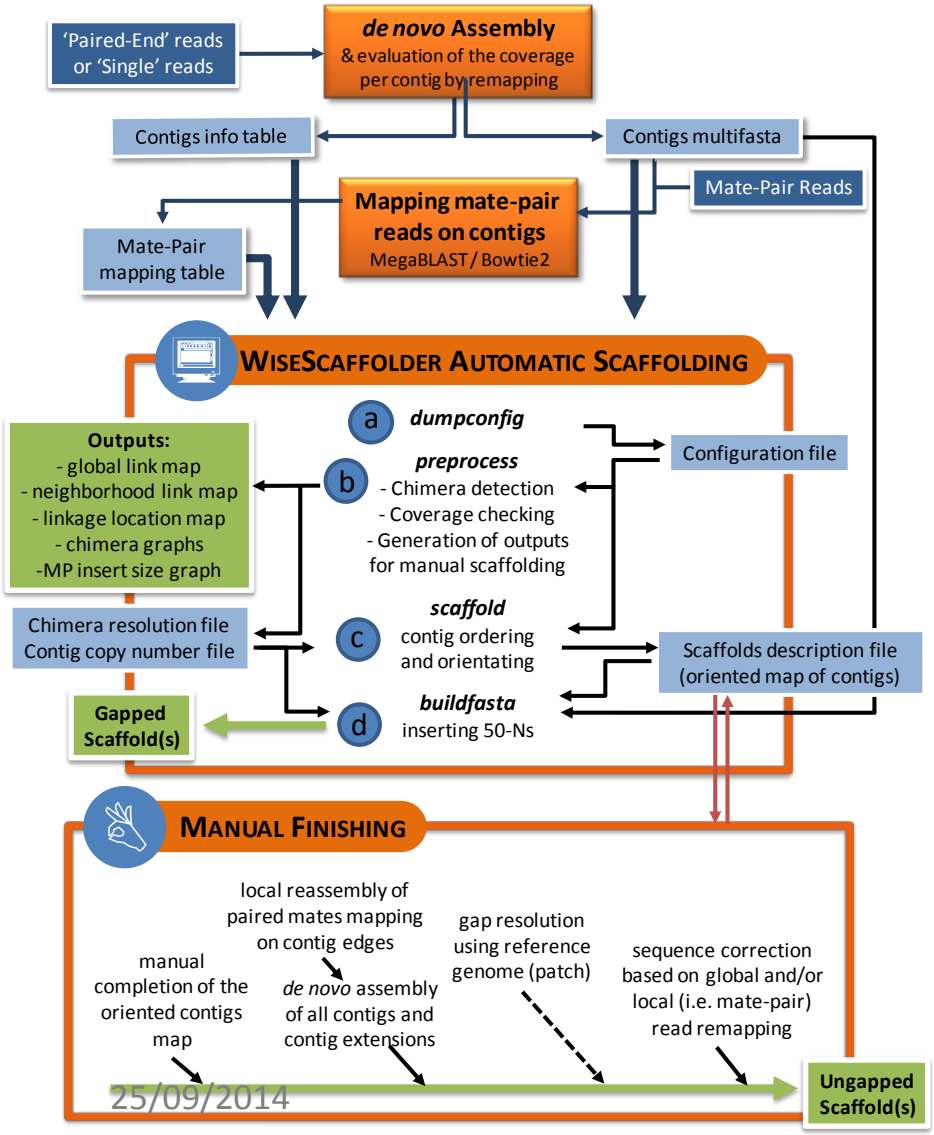
    Transcriptomes from 4 *Proc.* and 7 *Syn.* (Genoscope)

    Transcriptomes from stress experiments (HL/UV, stress ox, T°C, etc.) on 4 *Syn.* strains (~ 334 transcriptomes, ANR SAMOSA project)
  - Metagenomic and metatranscriptomic data

    Recruitment plots of environmental reads
  - Information about mutants

- New modules
  - Genome Browser (JBrowse)
  - Genomic island viewer
  - Comparative analysis tools (core/accessory/unique genes)

TOOLS : WiseScaffolder Workflow



Inputs:
- Contigs produced by CLC
- Contigs description (coverage, length)
- Mapping of mate-pairs

4 subcommands, 4 steps of tuning
- dumpconfig
- preprocess
- scaffold
- buildfasta

Chimera & contig copy number detection

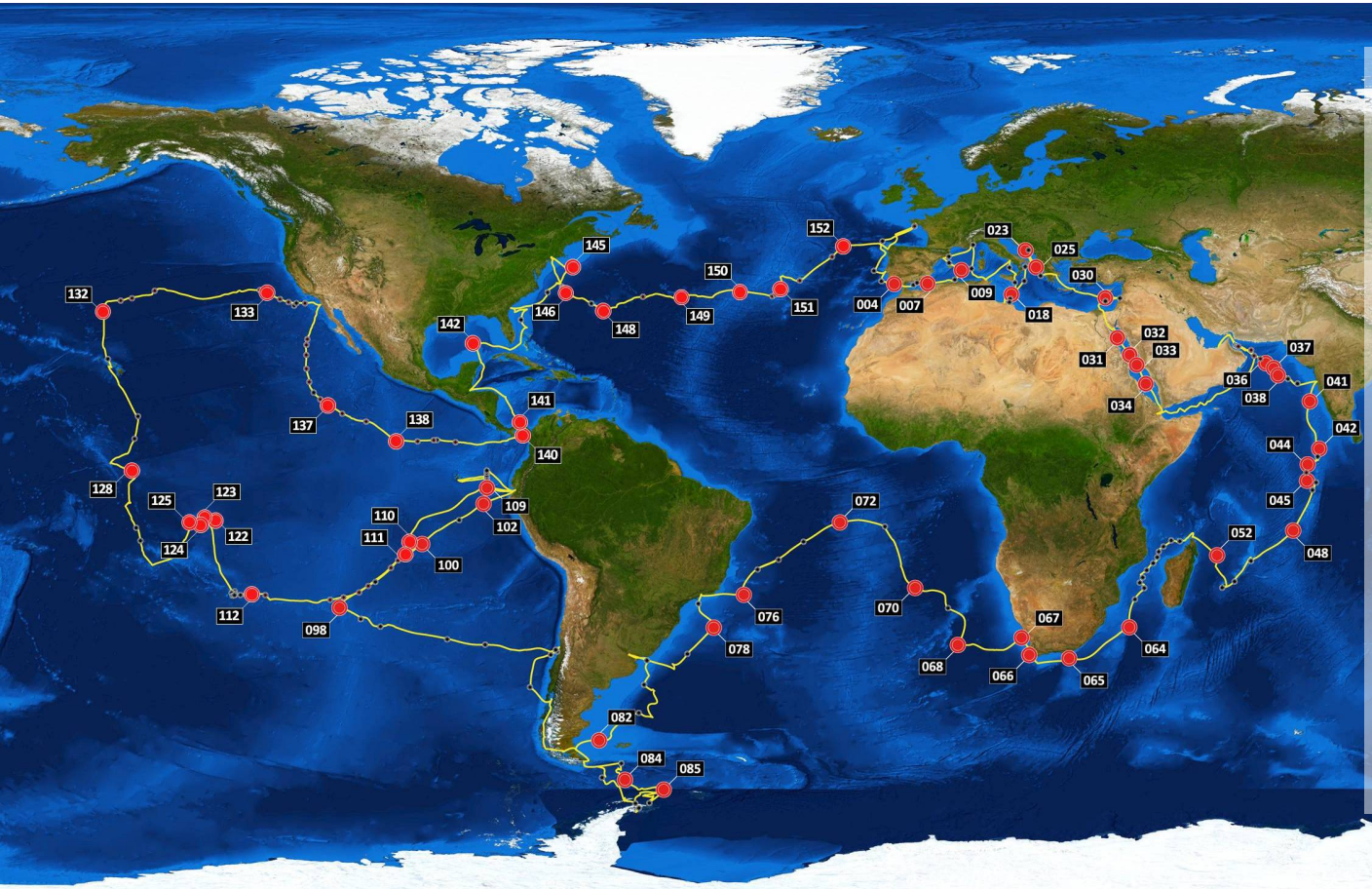
Outputs:
- scaffolds description
- fasta of scaffolds and unscaffolded contigs
- various tables and graphs to help with manual scaffolding

Tools : Workflow for metabarcoding data under Galaxy

Stéphane Audic (Évolution du Plancton et Écosystèmes Pélagiques, Adaptation et Diversité en Milieu Marin, Station biologique de Roscoff)
stephane.audic@sb-roscoff.fr

Collaboration Abims (Analyses and bioinformatics for Marine Science) http://abims.sb-roscoff.fr

METAGENOMICS
TARA Oceans Metagenomic Dataset



TARA Oceans:
2.5 years cruise

66 sampling stations

111 metagenomes for the bacterial size fraction
0.2-1.6 µm – 0.2-3 µm

63 SURface
45 Deep Chlorophyll Max.
3 Oxygen Min. Zone

Completed 2014 – G. Farrant

Metagenic Analysis

Environment

Metagenomic

Sequencing of the full DNA content of an environment. Works for prokaryotic and small eukaryotic genomes

Metatranscriptomic

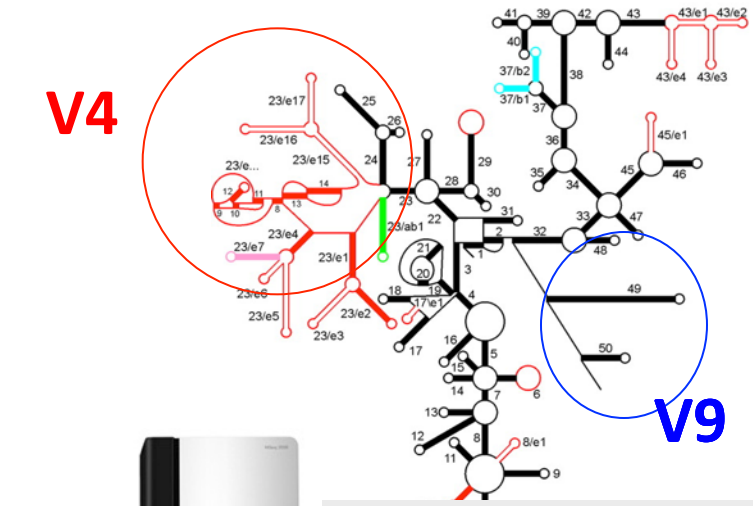
Sequencing of the full RNA content of an environment. Selection of organism categories : ex - eucaryotes poly-A selection

metabarcodingMetagenic

Sequencing of all « instances » of a gene in an environment.

Selected gene by PCR with primers with a great taxonomic coverage but specific enough to amplify only the gene of interest

Qualitative and semi-quantitative approach.



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

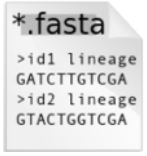
Step 1: raw reads-> table of occurrences

From sequence files and configuration files, get number of occurrences for **each detected sequence** in **each sample**.

sequence files

configuration files

reference sequences (chimera)

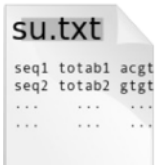



```
Seq1 Sample1 n1
Seq2 Sample1 n2
Seq3 Sample1 n3
..     ..    ..

Seq1  Sample2 n1
Seq2  Sample2 n2
..     ..    ..

Seq1  Sample3 n1
Seq3  Sample3 n2

Seq   Sample  n
  i         j  ij
```




```
Seq1 totab1 acgtgtgcggtgcgtttaag
Seq2 totab2 acgtgtgcgatgcgtttaag
Seq3 totab3 acgtgtgcggtgggtttaag
Seq4 totab4 tcgtgtgcggtgcgtttaag
Seq5 totab5 acttgtgcggtgcgtttaag
Seq6 totab6 acgtgtccggtgcgtttaag
Seq7 totab7 acgtgaccggtgcgtttaag
Seq8 totab8 acgtgtgcgaagcgtttaag
Seq9 totab9 acgtgtgcggtgcgtttaag
   ...       ...
...
Seq   totab       dnaseq
  i        i             i
```

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
Seq1	n1	n2	n3	n4	n5	n6
Seq2	n7	n8	n9	..	..	..
Seq3	..	..	..	..	..	..
Seq4	..	..	..	..	..	..

Input

Output

Creation of "Configuration file"

Separate samples:
In which file?
Wich MID?
Wich primerset? (= Which marker)
Wich primer direction

Index description for sample separation (MID)

sequence files

MID file

```
MID1 ACGCGTG
MID2 GCTAGTG
MID3 CCGTGTA
MID4 GCTGGTC
etc...
```

Sample file

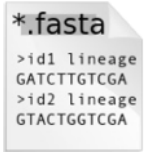
```
SMP1 fichier1 MID1 pset Dir
SMP2 fichier1 MID2 pset BOTH
SMP1 fichier2 MID1 pset CMP
etc...
```

configuration files

Pset file

```
pset1 primerF primerR base_de_ref
\ error-rate size-range
pset2 primerF primerR base_de_ref
\ error-rate size-range
etc...
```

reference sequences (chimera)



Input

Description of Primers, reference database, etc.

Raw data extraction

1- convert sff to fastq (if necessary)
2- 'scaling' of quality values (33;64; …)
3- search in raw data the exact motif
`"MID-PrimerF-([acgt]+)primerR(rev)"`
-> output the expected error raw in the corresponding sequence (base on the worst 50 nuc)

One file per sample

sequence files

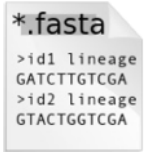
configuration files

np_complete_S*

```
>name1 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGTAGTC
>name2 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGTAGTC
>etc...
```

4- dereplication
Number of occurence

reference sequences (chimera)



5- collect of common identifiers in several samples

md5.several

```
id1
id2
id3
id4
etc.
```

np_uniq_S*

```
>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGTAGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGTAGTC
>etc..
```

Input

Cleaning and compilation of results

np_uniq_S*

```
>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGTAGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGTAGTC
>etc..
```

md5.several

```
id1
id2
id3
id4
etc.
```

Np_filtered_S*

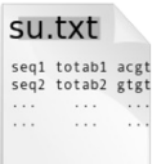
```
>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGTAGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGTAGTC
>etc..
```

Selection of the sequences according seq size and quality

Detection of chimera (uchime):
- Compare to reference sequence database
-Compare to most abondante sequence in the same sample

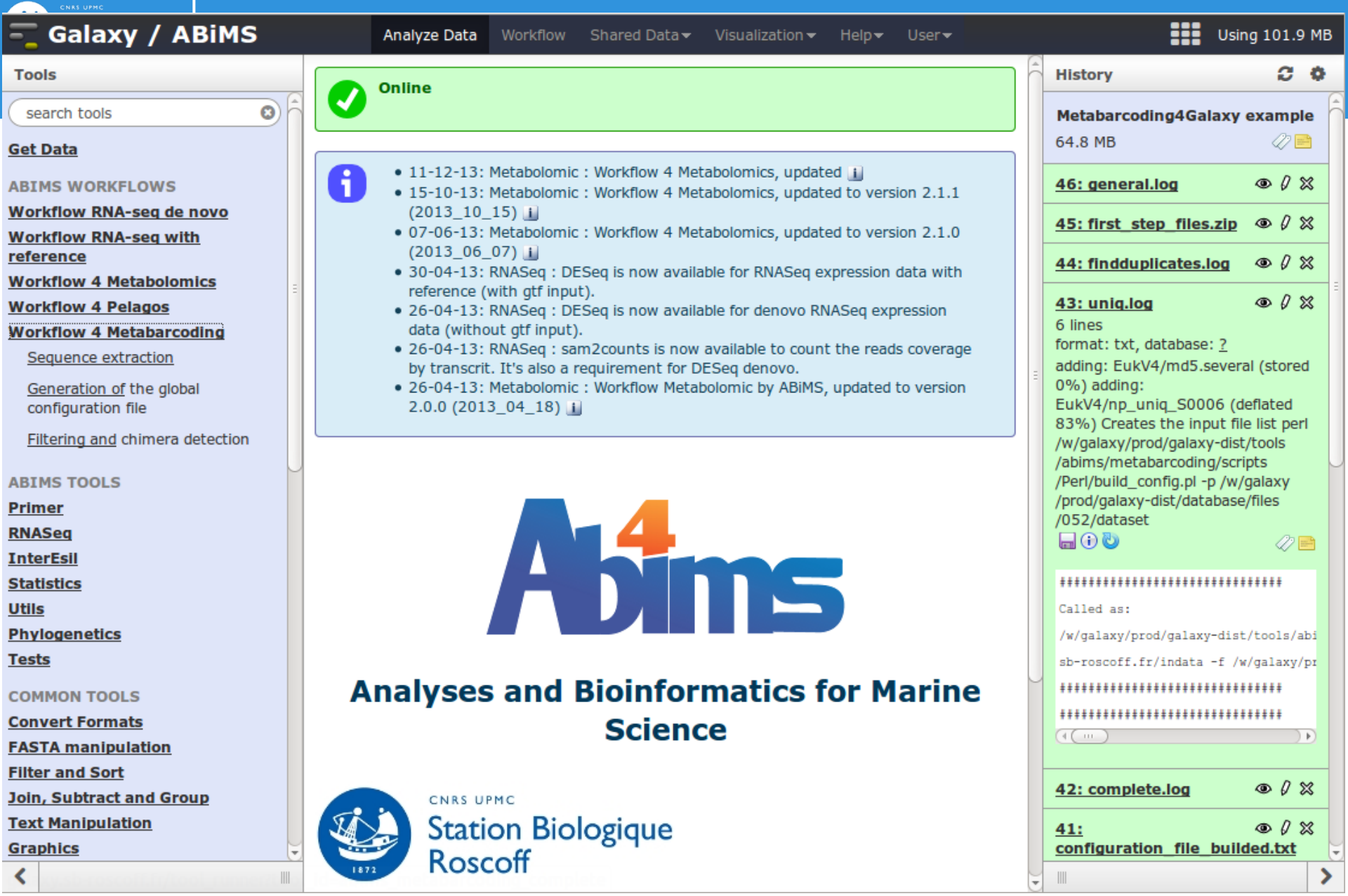
Np_filtered_nc_S**

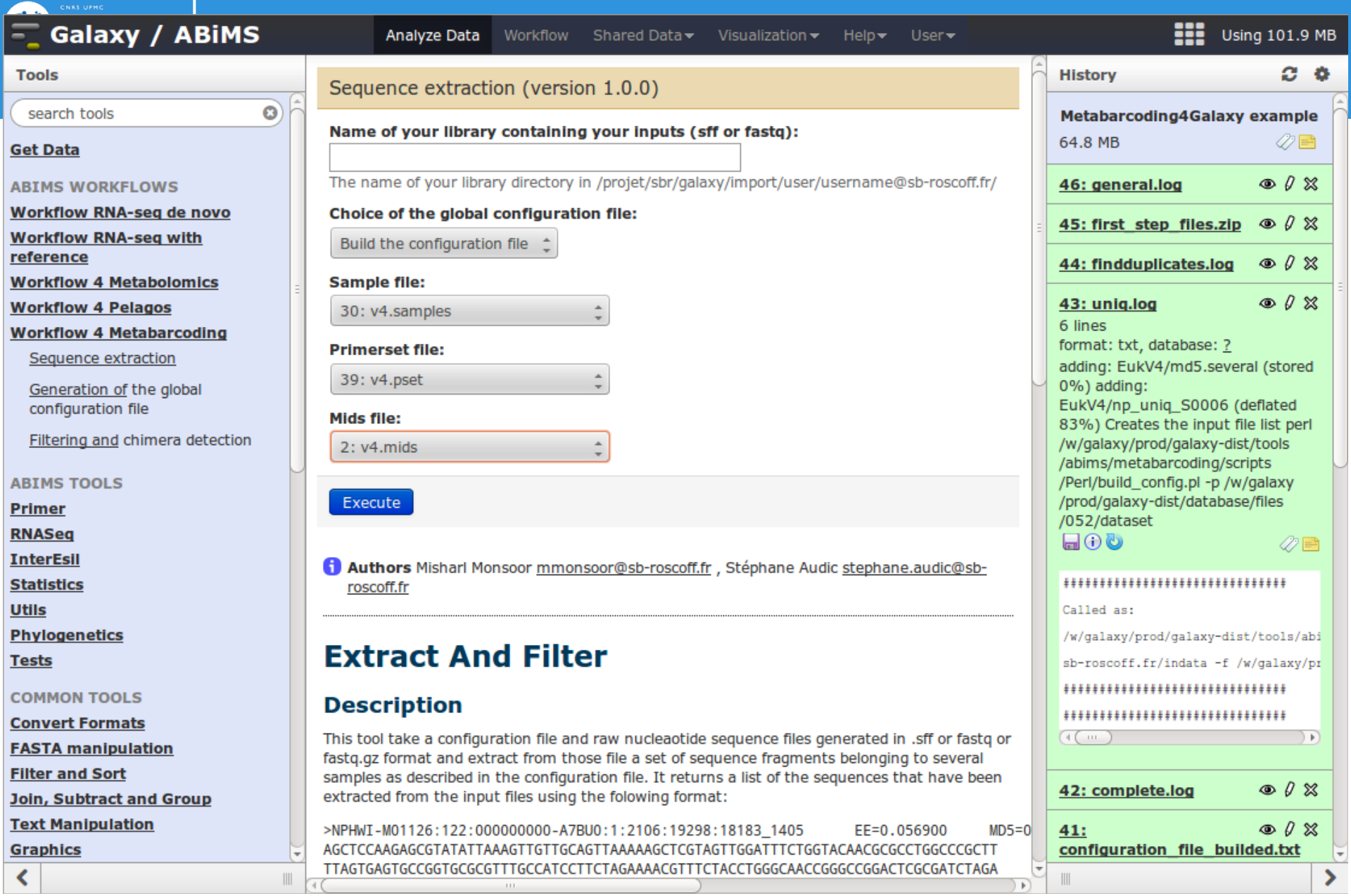
```
>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGTAGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGTAGTC
>etc..
```

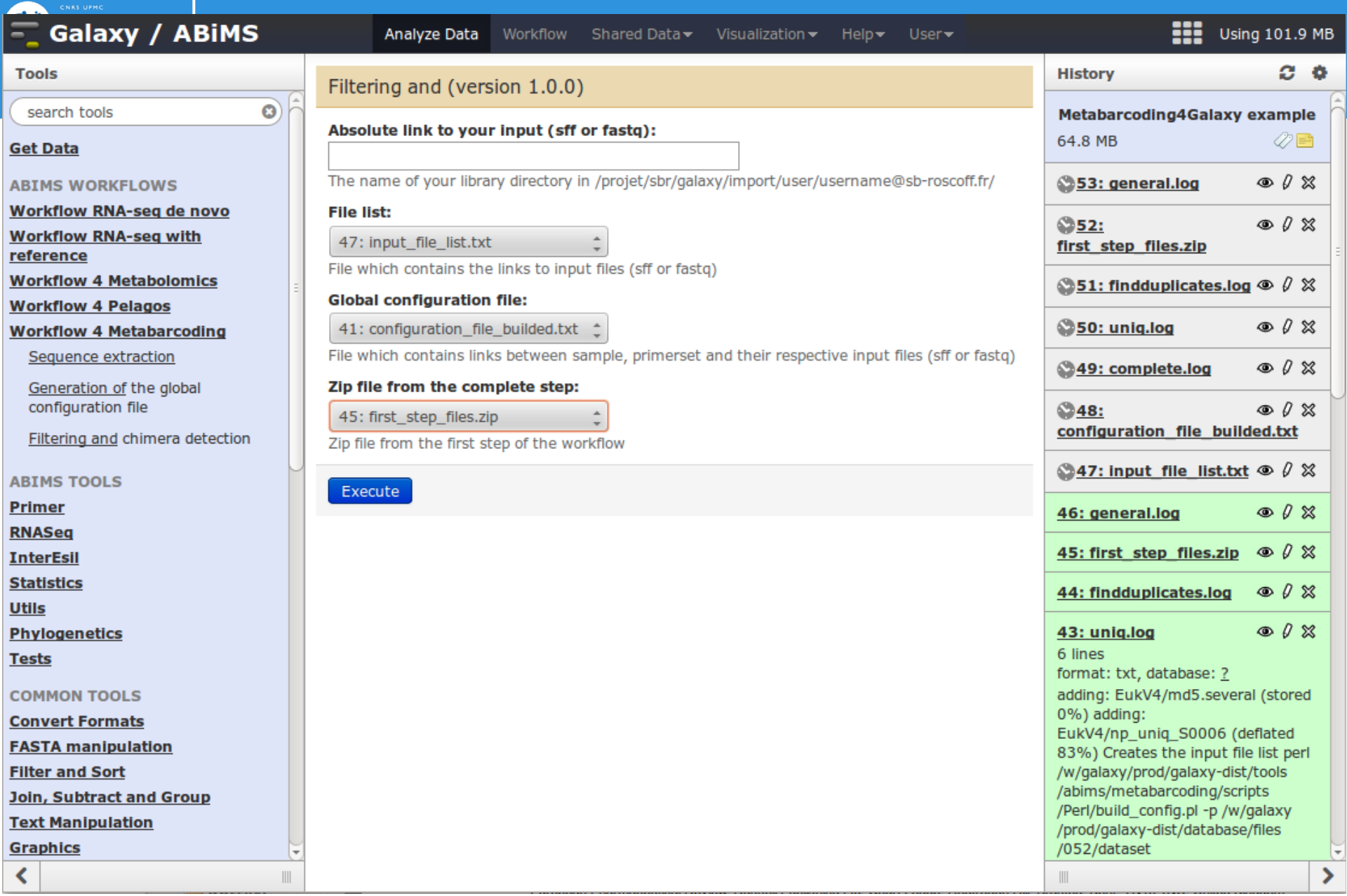


	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
Seq1	n1	n2	n3	n4	n5	n6
Seq2	n7	n8	n9	..	..	..
Seq3	..	..	..	..	..	..
Seq4	..	..	..	..	..	..

```
Seq1 totab1 acgtgtgcggtgcgtttaag
Seq2 totab2 acgtgtgcgatgcgtttaag
Seq3 totab3 acgtgtgcggtgggtttaag
Seq4 totab4 tcgtgtgcggtgcgtttaag
Seq5 totab5 acttgtgcggtgcgtttaag
Seq6 totab6 acgtgtccggtgcgtttaag
Seq7 totab7 acgtgaccggtgcgtttaag
Seq8 totab8 acgtgtgcgaagcgtttaag
Seq9 totab9 acgtgtgcggtgcgtttaag
   ...       ...
...
Seq   totab       dnaseq
```


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 101.9 MB
Tools
search tools
Get Data
ABIMS WORKFLOWS
Workflow RNA-seq de novo
Workflow RNA-seq with reference
Workflow 4 Metabolomics
Workflow 4 Pelagos
Workflow 4 Metabarcoding
Sequence extraction
Generation of the global configuration file
Filtering and chimera detection
ABIMS TOOLS
Primer
RNASeq
InterEsil
Statistics
Utils
Phylogenetics
Tests
COMMON TOOLS
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Text Manipulation
Graphics
Online
11-12-13: Metabolomic : Workflow 4 Metabolomics, updated
15-10-13: Metabolomic : Workflow 4 Metabolomics, updated to version 2.1.1 (2013_10_15)
07-06-13: Metabolomic : Workflow 4 Metabolomics, updated to version 2.1.0 (2013_06_07)
30-04-13: RNASeq : DESeq is now available for RNASeq expression data with reference (with gtf input).
26-04-13: RNASeq : DESeq is now available for denovo RNASeq expression data (without gtf input).
26-04-13: RNASeq : sam2counts is now available to count the reads coverage by transcrit. It's also a requirement for DESeq denovo.
26-04-13: Metabolomic : Workflow Metabolomic by ABiMS, updated to version 2.0.0 (2013_04_18)
Analyses and Bioinformatics for Marine Science
CNRS UPMC
Station Biologique Roscoff
History
Metabarcoding4Galaxy example
64.8 MB
46: general.log
45: first_step_files.zip
44: findduplicates.log
43: uniq.log
6 lines
format: txt, database: ?
adding: EukV4/md5.several (stored 0%) adding: EukV4/np_uniq_S0006 (deflated 83%) Creates the input file list perl /w/galaxy/prod/galaxy-dist/tools /abims/metabarcoding/scripts /Perl/build_config.pl -p /w/galaxy /prod/galaxy-dist/database/files /052/dataset
Called as:
/w/galaxy/prod/galaxy-dist/tools/abi
sb-roscoff.fr/indata -f /w/galaxy/pr
42: complete.log
41: configuration_file_builded.txt


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 101.9 MB
Tools
search tools
Get Data
ABIMS WORKFLOWS
Workflow RNA-seq de novo
Workflow RNA-seq with reference
Workflow 4 Metabolomics
Workflow 4 Pelagos
Workflow 4 Metabarcoding
Sequence extraction
Generation of the global configuration file
Filtering and chimera detection
ABIMS TOOLS
Primer
RNASeq
InterEsil
Statistics
Utils
Phylogenetics
Tests
COMMON TOOLS
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Text Manipulation
Graphics
Sequence extraction (version 1.0.0)
Name of your library containing your inputs (sff or fastq):
The name of your library directory in /projet/sbr/galaxy/import/user/username@sb-roscoff.fr/
Choice of the global configuration file:
Build the configuration file
Sample file:
30: v4.samples
Primerset file:
39: v4.pset
Mids file:
2: v4.mids
Execute
Authors Misharl Monsoor mmonsoor@sb-roscoff.fr , Stéphane Audic stephane.audic@sb-roscoff.fr
Extract And Filter
Description
This tool take a configuration file and raw nucleaotide sequence files generated in .sff or fastq or fastq.gz format and extract from those file a set of sequence fragments belonging to several samples as described in the configuration file. It returns a list of the sequences that have been extracted from the input files using the folowing format:
>NPHWI-M01126:122:000000000-A7BU0:1:2106:19298:18183_1405 EE=0.056900 MD5=0
AGCTCCAAGAGCGTATATTAAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGGATTTCTGGTACAACGCGCCTGGCCCGCTT
TTAGTGAGTGCCGGTGCGCGTTTGCCATCCTTCTAGAAAACGTTTCTACCTGGGCAACCGGGCCGGACTCGCGATCTAGA
History
Metabarcoding4Galaxy example
64.8 MB
46: general.log
45: first_step_files.zip
44: findduplicates.log
43: uniq.log
6 lines
format: txt, database: ?
adding: EukV4/md5.several (stored 0%) adding: EukV4/np_uniq_S0006 (deflated 83%) Creates the input file list perl /w/galaxy/prod/galaxy-dist/tools /abims/metabarcoding/scripts /Perl/build_config.pl -p /w/galaxy /prod/galaxy-dist/database/files /052/dataset
##########
Called as:
/w/galaxy/prod/galaxy-dist/tools/abi
sb-roscoff.fr/indata -f /w/galaxy/pr
##########
##########
42: complete.log
41: configuration_file_builded.txt


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 101.9 MB
Tools
search tools
Get Data
ABIMS WORKFLOWS
Workflow RNA-seq de novo
Workflow RNA-seq with reference
Workflow 4 Metabolomics
Workflow 4 Pelagos
Workflow 4 Metabarcoding
Sequence extraction
Generation of the global configuration file
Filtering and chimera detection
ABIMS TOOLS
Primer
RNASeq
InterEsil
Statistics
Utils
Phylogenetics
Tests
COMMON TOOLS
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Text Manipulation
Graphics
Filtering and (version 1.0.0)
Absolute link to your input (sff or fastq):
The name of your library directory in /projet/sbr/galaxy/import/user/username@sb-roscoff.fr/
File list:
47: input_file_list.txt
File which contains the links to input files (sff or fastq)
Global configuration file:
41: configuration_file_builded.txt
File which contains links between sample, primerset and their respective input files (sff or fastq)
Zip file from the complete step:
45: first_step_files.zip
Zip file from the first step of the workflow
Execute
History
Metabarcoding4Galaxy example
64.8 MB
53: general.log
52: first_step_files.zip
51: findduplicates.log
50: uniq.log
49: complete.log
48: configuration_file_builded.txt
47: input_file_list.txt
46: general.log
45: first_step_files.zip
44: findduplicates.log
43: uniq.log
6 lines
format: txt, database: ?
adding: EukV4/md5.several (stored 0%) adding: EukV4/np_uniq_S0006 (deflated 83%) Creates the input file list perl /w/galaxy/prod/galaxy-dist/tools /abims/metabarcoding/scripts /Perl/build_config.pl -p /w/galaxy /prod/galaxy-dist/database/files /052/dataset

Step 2: occurence table -> assignation

Node	Lineage
Eukaryota|Opisthokonta|Mesomycetozoa|Ichthyosporea|Rhinosporidia|Rhinosporidae	
Eukaryota|Archaeplastida|Chlorophyta|Trebouxiophyceae|Chlorellales|Chlorellales_X|Meyerella|Meyerella+sp.	
Eukaryota|Opisthokonta|Metazoa|Arthropoda|Hexapoda|Insecta|Coccoblus	
Eukaryota|Opisthokonta|Metazoa|Arthropoda|Hexapoda|Insecta|Saldula|Saldula+sp.	
Eukaryota|Alveolata|Ciliophora|Litostomatea|Trichostomatia|Amylovoracidae|Amylovorax|Amylovorax+dogiell	
Eukaryota|Opisthokonta|Metazoa|Arthropoda|Crustacea|Ostracoda|Parasterope|Parasterope+pollex	
Eukaryota|Archaeplastida|Chlorophyta|Chlorophyceae|Chlorophyceae_X|CW-Chlamydomonadales|Pteromonas|Pteromonas+angulosa	
Eukaryota|Opisthokonta|Metazoa|Mollusca|Gastropoda|Caenogastropoda|Crepidula|Crepidula+fornicata	
Eukaryota|Stramenopiles|Stramenopiles_X|Bacillariophyta|Bacillariophyta_X|Raphid-pennate|Cymbella|Cymbella+mlnuta	
Eukaryota|Opisthokonta|Metazoa|Arthropoda|Hexapoda|Insecta|Matronoides|Matronoides+cyanelpennis	

Experimental Data

Seq	totab	dnaseq
Seq1	totab1	acgtgtgcggtgcgtttaag
Seq2	totab2	acgtgtgcgatgcgtttaag
Seq3	totab3	acgtgtgcggtgggtttaag
Seq4	totab4	tcgtgtgcggtgcgtttaag
Seq5	totab5	acttgtgcggtgcgtttaag
Seq6	totab6	acgtgtccggtgcgtttaag
Seq7	totab7	acgtgaccggtgcgtttaag
Seq8	totab8	acgtgtgcgaagcgtttaag
Seq9	totab9	acgtgtgcggtgcgtttaag
...	...	...
i	i	i

Reference sequences

Reference	Node
14.1.1837_U	17059
82.1.1888_U	41847
85.1.1715_U	28767
75.1.1781_U	4702
90.1.1711_U	46762
59.1.1955_U	43646
54.1.1769_U	39467
80.1.2094_U	3238
82.1.2144_U	1923
90.1.1695_U	2505

Node	Parent
58684	58175
57573	57566
58017	57997
44919	44918
8563	8562
28336	1194
28801	28800
5006	2228
55548	55546
29429	2292

ggsearch

alignement global
multi-threaded

Hit List

Seq1	%id	ref1
Seq1	%id	ref2
Seq2	%id	ref3
Seq3	%id	ref4
Seq3	%id	ref5
...	...	...

Assignation

Seq1	%id	ref1,ref2	lineage
Seq2	%id	ref3	lineage2
Seq3	%id	ref4,ref5	lineage3
...	...	.....	.....
...	...	.....	.....

Step 3:co-occurence table -> OTUisation

Experimental data

```
Seq1 totab1 acgtgtgcggtgcgtttaag
Seq2 totab2 acgtgtgcgatgcgtttaag
Seq3 totab3 acgtgtgcggtgggtttaag
Seq4 totab4 tcgtgtgcggtgcgtttaag
Seq5 totab5 acttgtgcggtgcgtttaag
Seq6 totab6 acgtgtccggtgcgtttaag
Seq7 totab7 acgtgaccggtgcgtttaag
Seq8 totab8 acgtgtgcgaagcgtttaag
Seq9 totab9 acgtgtgcggtgcgtttaag
  ...    ...           ...
Seq    totab        dnaseq
  i        i             i
```

Swarms

```
Seq1 totab1 acgtgtgcggtgcgtttaag swarm_id1
Seq2 totab2 acgtgtgcgatgcgtttaag swarm_id2
Seq3 totab3 acgtgtgcggtgggtttaag swarm_id3
Seq4 totab4 tcgtgtgcggtgcgtttaag swarm_id3
Seq5 totab5 acttgtgcggtgcgtttaag .....
Seq6 totab6 acgtgtccggtgcgtttaag .....
Seq7 totab7 acgtgaccggtgcgtttaag .....
Seq8 totab8 acgtgtgcgaagcgtttaag .....
Seq9 totab9 acgtgtgcggtgcgtttaag .....
  ...    ...           ...
Seq    totab        dnaseq
  i        i             i
```

Swarm
fast and accurate
OTU construction

Swarm: robust and fast clustering method for amplicon-based studies

Frédéric Mahé[1,2,3], Torbjørn Rognes[4,5], Christopher Quince[6], Colomban de Vargas[1,2], and Micah Dunthorn[3]

PeerJ PrePrints 2:e386v1 http://dx.doi.org/10.7287/peerj.preprints.386v1

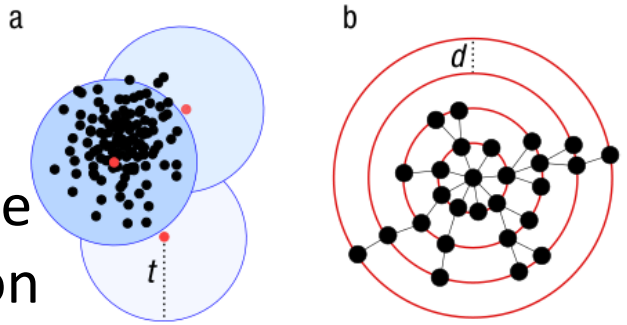


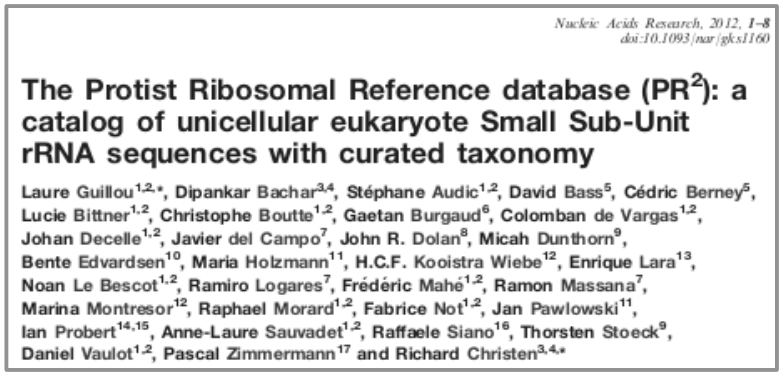
Figure 1. Visualization of the widely used greedy clustering approach based on centroid selection and a global clustering threshold, *t*, where closely related amplicons can be placed into different OTUs. (b) By contrast, Swarm clusters iteratively by using a small local clustering threshold, *d*, allowing OTUs to reach their natural limits.

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
Swarm1	n1	n2	n3	n4	n5	n6
Swarm2	n7	n8	n9	..	..	..
Swarm3	..	..	..	..	..	..
Swarm4	..	..	..	..	..	..

Pipeline as a service

- Metagenetic is being developed, but it is an area that needs to offer STANDARDIZATION and dataset easily COMPARABLE.
- That's why we want to encourage the use of set of primers COMPATIBLE. (Developed as part of BioMarks / Tara Ocean project).

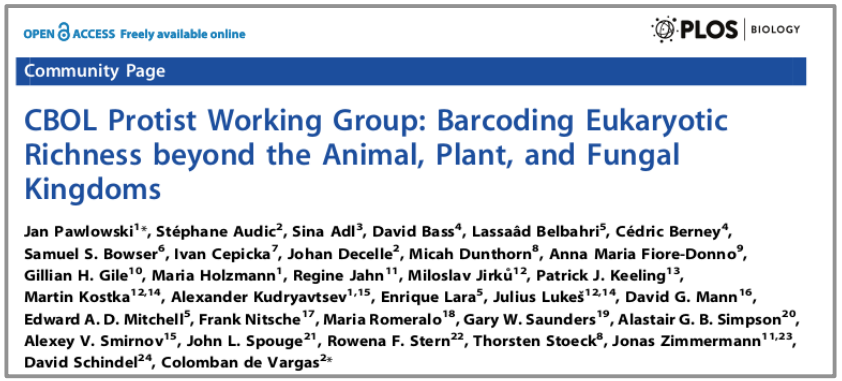
  - Dissemination of protocols / primers
  - Reference Database of specific marker: V4 specific database - V9 specific database - LSUD1D2 specific database (D. Grzebyk, lagunar ecosystemes) - Chloroplastic 16S specific database (J. Decelle, S. Romac)



Nucleic Acids Research, 2012, 1–8
doi:10.1093/nar/gks1160

The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy

Laure Guillou[1,2,*], Dipankar Bachar[3,4], Stéphane Audic[1,2], David Bass[5], Cédric Berney[5], Lucie Bittner[1,2], Christophe Boutte[1,2], Gaetan Burgaud[6], Colomban de Vargas[1,2], Johan Decelle[1,2], Javier del Campo[7], John R. Dolan[8], Micah Dunthorn[9], Bente Edvardsen[10], Maria Holzmann[11], H.C.F. Kooistra Wiebe[12], Enrique Lara[13], Noan Le Bescot[1,2], Ramiro Logares[7], Frédéric Mahé[1,2], Ramon Massana[7], Marina Montresor[12], Raphael Morard[1,2], Fabrice Not[1,2], Jan Pawlowski[11], Ian Probert[14,15], Anne-Laure Sauvadet[1,2], Raffaele Siano[16], Thorsten Stoeck[9], Daniel Vaulot[1,2], Pascal Zimmermann[17] and Richard Christen[3,4,*]



OPEN ACCESS Freely available online

PLOS BIOLOGY

Community Page

CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms

Jan Pawlowski[1*], Stéphane Audic[2], Sina Adl[3], David Bass[4], Lassaâd Belbahri[5], Cédric Berney[4], Samuel S. Bowser[6], Ivan Cepicka[7], Johan Decelle[2], Micah Dunthorn[8], Anna Maria Fiore-Donno[9], Gillian H. Gile[10], Maria Holzmann[1], Regine Jahn[11], Miloslav Jirků[12], Patrick J. Keeling[13], Martin Kostka[12,14], Alexander Kudryavtsev[1,15], Enrique Lara[5], Julius Lukeš[12,14], David G. Mann[16], Edward A. D. Mitchell[5], Frank Nitsche[17], Maria Romeralo[18], Gary W. Saunders[19], Alastair G. B. Simpson[20], Alexey V. Smirnov[15], John L. Spouge[21], Rowena F. Stern[22], Thorsten Stoeck[8], Jonas Zimmermann[11,23], David Schindel[24], Colomban de Vargas[2*]

Why Galaxy?

User point of view

- Complexes pipelines available to a larger community
- No complex installation (f.e. Qiime)
- Allows a transparent access to HPC facility

Developer point of view

- Code reviewing / clearly break the steps of an analysis / identify what can be parallelized and what can not.
- Lighten the distribution step
- Frees up time (to allow others to use its codes without having to run it by yourself)

Platform point of view

- Collaboration research/platform
- Integration of tools (workflow) / data
- Competences and visibility
- Spreading : New training subjects