Métagénique marine: Workflows pour l'analyse de données haut-débit sous Galaxy

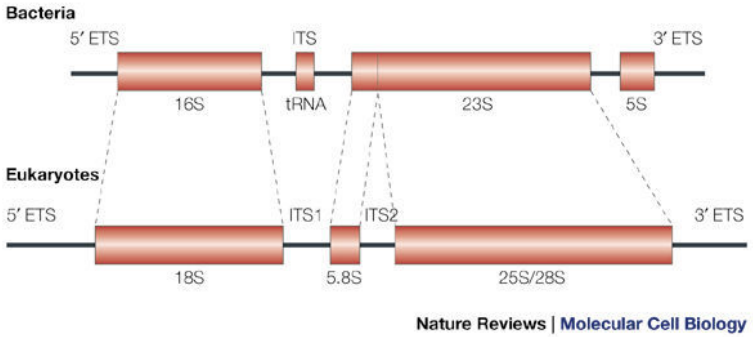
Laure Quintric (Cellule bioinformatique, Centre Ifremer de Brest)

Stéphane Audic (Évolution du Plancton et Écosystèmes Pélagiques, Adaptation et Diversité en
Milieu Marin, Station biologique de Roscoff)

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

Intégration de Qiime dans Galaxy

- Étude diversité des micro-organismes présents dans un environnement par technique NGS
  - Métabarcoding : étude de marqueurs ARN ribosomique : 16s – 18s – ITS...

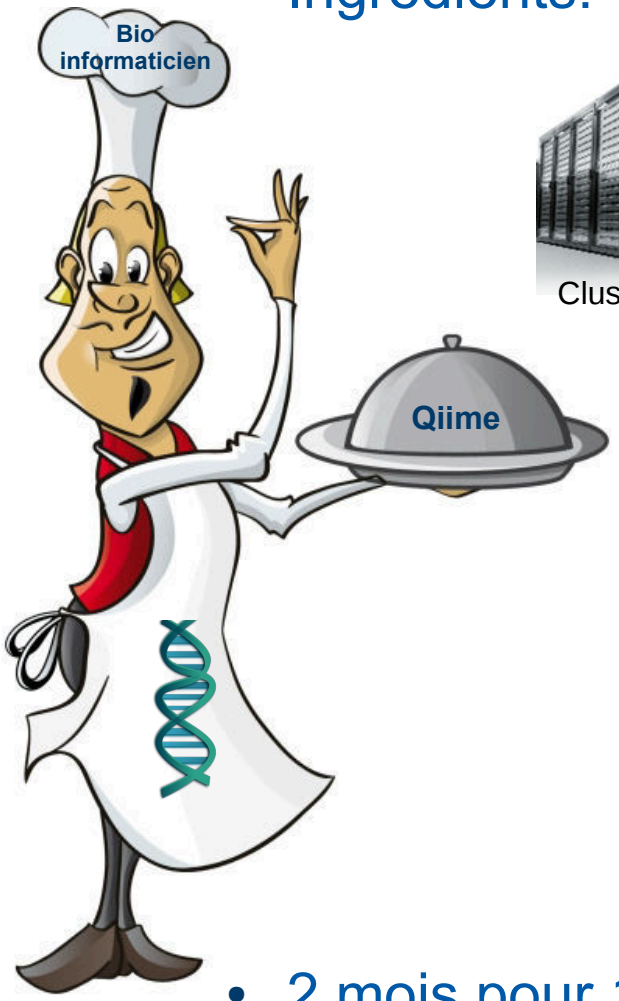


- Contexte :
  - Émergence de projets d'analyse haut-débit d'échantillons environnementaux
  - Sollicitations des laboratoires pour utiliser un outil en interne
- Projet :

Quantitative Insights Into Microbial Ecology

La recette d'intégration de Qiime dans Galaxy

- Ingrédients:



Cluster PCIM

Quantitative Insights Into Microbial Ecology
Qiime
Version 1.8

Python	ChimeraSlayer
Numpy	mothur
MatPlotLib	clearcut
PyCogent	raxml
biom-format	infernal
qcli	cdbtools
PyNAST	muscle
Emperor	pplacer
uclust	ParsInsert
fasttree	usearch
jre1.6	sfffile
rdp_classifier	sffinfo
tax2tree	AmpliconNoise
blast	R
cd-hit	...

Galaxy

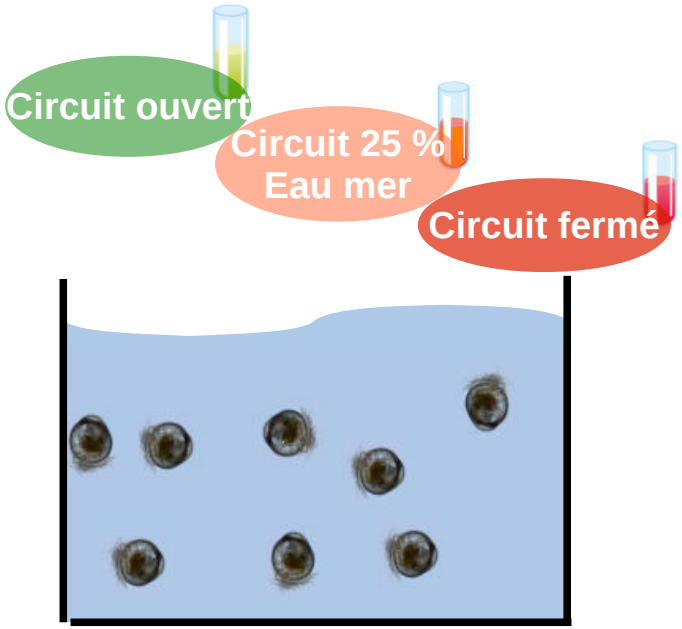
Qiime 1.7
Wrappers

Reconfiguration de 17 wrappers :
- Adaptation à la version 1.8
- Modification de la gestion des sorties (fini les .gz)
- Affichage des sorties graphiques directement dans Galaxy

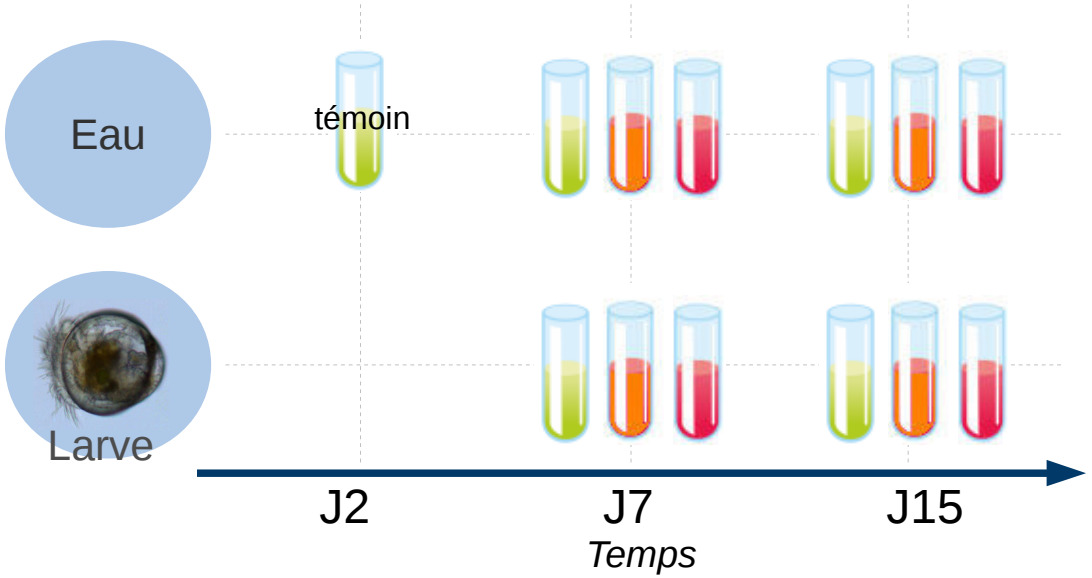
- 2 mois pour 1 ETP

Démo du workflow

- Travaux de thèse de Katia Asmani (sous la direction de Jean-Louis Nicolas), Laboratoire LEMAR (*UMR CNRS/UBO/IRD/Ifremer*)
  - *« Étude du microbiome associé aux élevages larvaires et post-larvaires de l'huître creuse Crassostrea Gigas »*
  - *Librairie : 13 échantillons ARN16S Pyroséquençage Roche 454*



Bac d'élevage larvaire



Données d'entrée du workflow

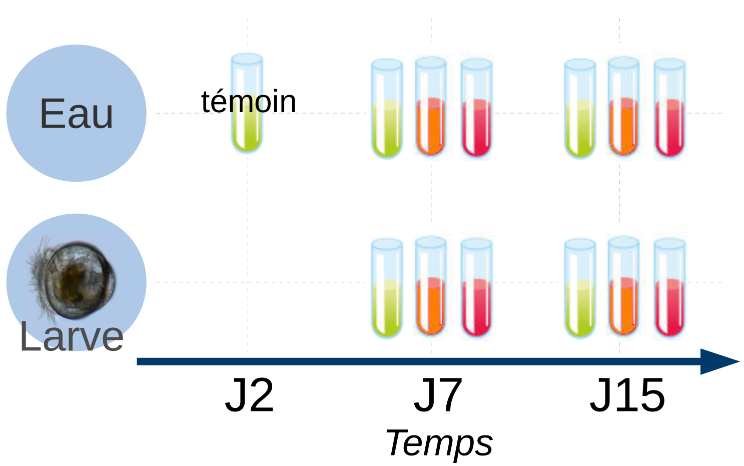
- reads.fasta :

```
>H33RDWS01CN1LE length=474 xy=0976_3456 region=1 run=R_2013_02_06_13_21_22_
ACGAGTGCGTAGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGTATGCTTAACACATGCAAGTCGAACGAGAGTTTTTAACTCTAGTGGCGAACGGGTGCGTAATACGTGAGAATCTGCC
>H33RDWS01EXULU length=286 xy=1908_2720 region=1 run=R_2013_02_06_13_21_22_
ACGAGTGCGTAGAGTTTGATCCTGGCTCAGGGTGAACGCTAGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAATTAGCTTGCTAATTTGCTGACGAGTGGCGCACGGGTGCGTA
>H33RDWS01ELU0D length=468 xy=1772_0427 region=1 run=R_2013_02_06_13_21_22_
ACGAGTGCGTAGAGTTTGATCCTGGCTCAGAATGAACGCTGGCGGCATGCCTAACACATGCAAGTCGAACGAGATCTTCGGATCTAGTGGCGCACGGGTGAGTAACGCGTGGGAACCTACCT
```

- reads.qual :

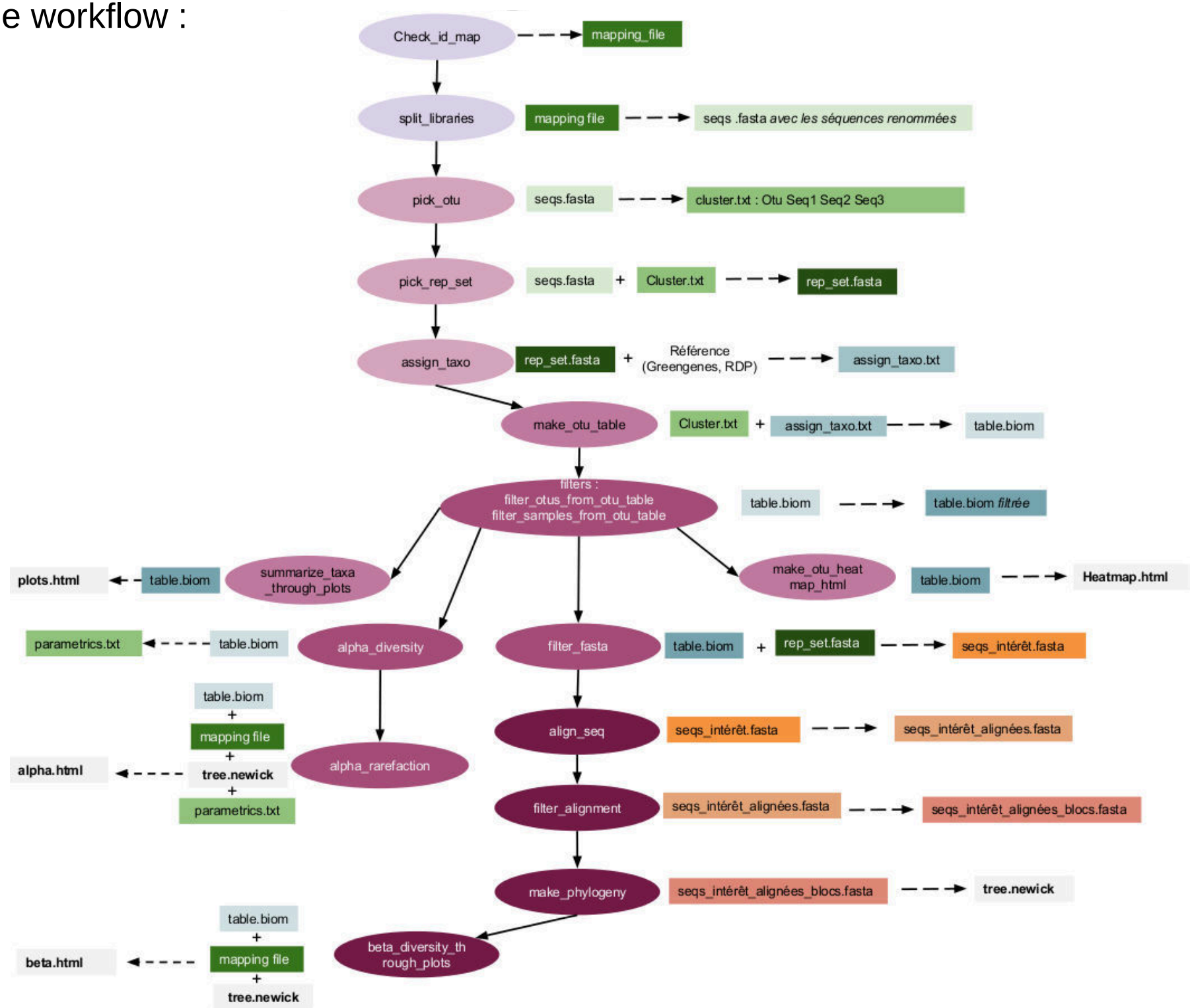
```
>H33RDWS01CN1LE length=474 xy=0976_3456 region=1 run=R_2013_02_06_13_21_22_
40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
>H33RDWS01EXULU length=286 xy=1908_2720 region=1 run=R_2013_02_06_13_21_22_
40 40 40 40 40 40 40 40 40 40 40 40 40 39 21 21 21 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
>H33RDWS01ELU0D length=468 xy=1772_0427 region=1 run=R_2013_02_06_13_21_22_
40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
```

- Description des échantillons :



SampleID	BarcodeSequence	Sample	Circuit	Description
Ech1	ACGAGTGCGT	Eau	Ouvert	Temoin
Ech2	AGACGCACTC	Eau	Ouvert	E C0 j7
Ech3	ATATCGCGAG	Eau	Ferme25	E CF25 j7
Ech4	TCTCTATGCG	Eau	Ferme	E CF0 j7
Ech5	AGCACTGTAG	Eau	Ouvert	E C0 j15
Ech6	CGTGTCTCTA	Eau	Ferme25	E CF25 j15
Ech7	TGATACGTCT	Eau	Ferme	E CF0 j15
Ech8	TACGAGTATG	Larve	Ouvert	L C0 j7
Ech9	TAGAGACGAG	Larve	Ferme25	L CF25 j7
Ech10	ACATACGCGT	Larve	Ferme	L CF0 j7
Ech11	TACTCTCGTG	Larve	Ouvert	L C0 j15
Ech12	TCGTCGCTCG	Larve	Ferme25	L CF25 j15
Ech13	ACGCGAGTAT	Larve	Ferme	L CF0 j15

Qiime workflow :



Metabarcoding4Galaxy

- Souplesse / polyvalence / rapidité
- De nombreux projets de metagénique (=metabarcoding)
- Essaie de standardiser une suite d'analyse :
  - Choix d'environnements
  - Choix de marqueurs: 16S , 18S, boucle V4, V9, etc.
  - Choix de techniques de séquençage
  - Multiplexage
  - Bases de séquences de références marqueur spécifiques
  - Assignations (utilisation de programme d'alignement global)
  - Clustering ; Analyses.

De l'échantillon aux fichiers de séquence (en bref)

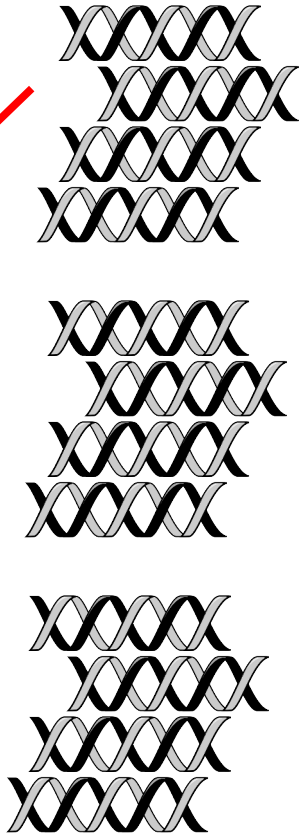
Collecte des échantillons

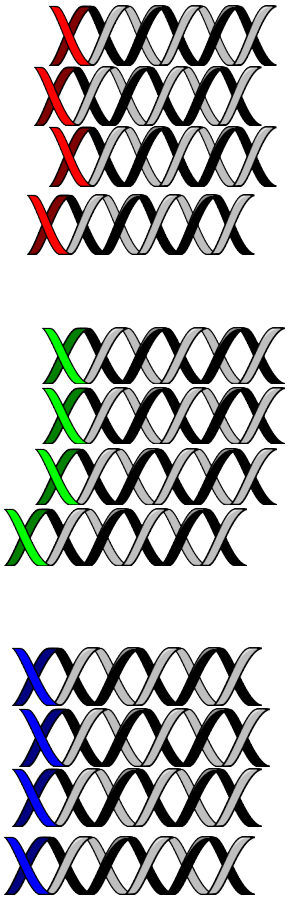
Sélection de l'ADN correspondant à une région particulière par PCR

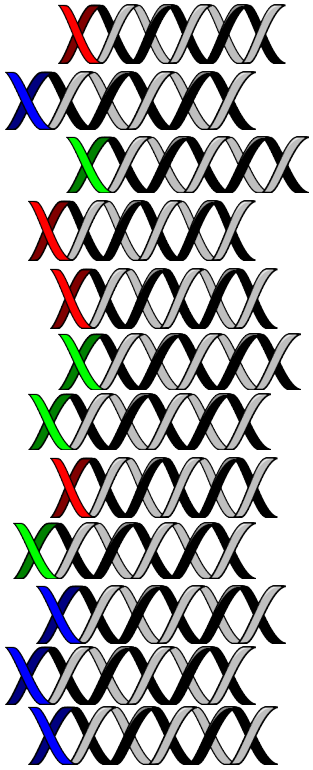
Ajout d'une étiquette pour identifier les échantillons (MID)

Mélange du tout pour séquençage

Plusieurs types de fichiers possibles







454:

.sff

Illumina:

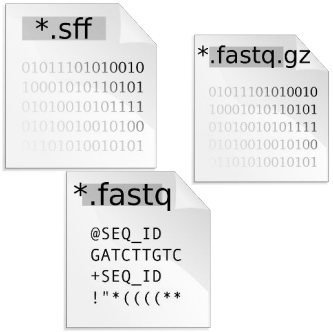
.fastq
.fastq.gz

Le but du Workflow

A partir des fichiers de séquences et de fichiers de configuration, obtenir pour **chaque séquence détectée**, son **nombre d'occurrences** dans **chaque échantillon**.

Entrée

sequence files



configuration files

reference sequences (chimera)

Sortie

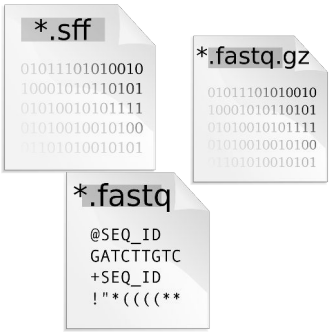

```
Seq1 Sample1 n1
Seq2 Sample1 n2
Seq3 Sample1 n3
   ..          ..
     ..

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

Seq1  Sample3 n1
Seq3  Sample3 n2

Seq_i  Sample_j  n_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_i  totab_i      dnaseq_i
```

Génération fichiers configuration

Entrée

sequence files



configuration files

reference sequences (chimera)

Comment séparer les échantillons: Quel MID, dans quel fichier, quel primerset (= quel marqueur)

Description des indexes pour séparer les échantillons (MID)

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...

Pset file

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

Description primers, sequences de réference, etc.

Extraction des données brutes

Entrée

*.sff

*.fastq.gz

*.fastq

```
@SEQ_ID
GATCTTGTC
+SEQ_ID
!"*((((**
```

*.conf

```
smp1 file1 mid1
smp2 file1 mid2
smp3 file2 mid1
"" "" ""
"" "" ""
```

*.fasta

```
>id1 lineage
GATCTTGTCGA
>id2 lineage
GTACTGGTCGA
```

1- transforme sff en fastq si nécessaire
2- détermine le 'scaling' des valeurs de qualité (pour les spécialistes)
3- Recherche dans les séquences brutes des structures de la forme MID-PrimerF-([acgt]+)primerR(rev) (recherche exacte) et sort le nombre attendue d'erreurs dans la séquence correspondante

Un fichier par échantillon.

np_complete_S*

```
>name1 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGT
AGTC
>name2 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGT
AGTC
>etc...
```

4- déréplication

np_uniq_S*

```
>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGT
AGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGT
AGTC
>etc..
```

5- collecte des identifiants communs à plusieurs échantillons

md5.several

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

Nettoyage et compilation des résultats

Entrée

*.conf

```
smp1 file1 mid1
smp2 file1 mid2
smp3 file2 mid1
""   ""    ""
""   ""    ""
```

*.fasta

```
>id1 lineage
GATCTTGTCGA
>id2 lineage
GTACTGGTCGA
```

np_uniq_S*

>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGT
AGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGT
AGTC
>etc..

md5.several

id1
id2
id3
id4
etc.

Sélection des séquences dans la bonne gamme de taille et qualité convenable

np_filtered_S*

>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGT
AGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGT
AGTC
>etc..

Détection des chimères (avec uchime):

- Par rapport à la base de référence
- Par rapport aux séquences plus abondantes du même échantillon

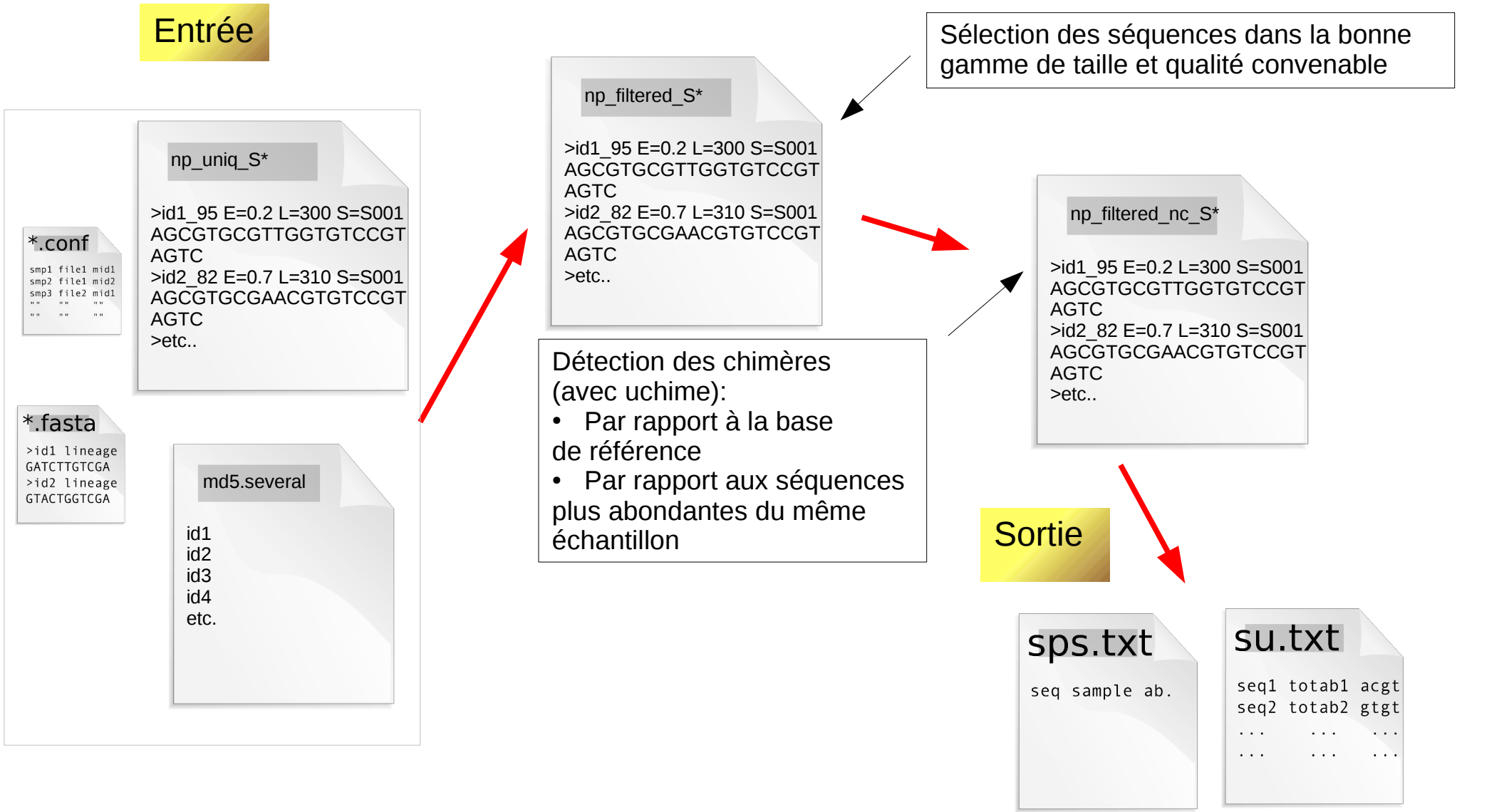
np_filtered_nc_S*

>id1_95 E=0.2 L=300 S=S001
AGCGTGCGTTGGTGTCCGT
AGTC
>id2_82 E=0.7 L=310 S=S001
AGCGTGCGAACGTGTCCGT
AGTC
>etc..

sps.txt

```
seq sample ab.
```

Nettoyage et compilation des résultats



Perspectives

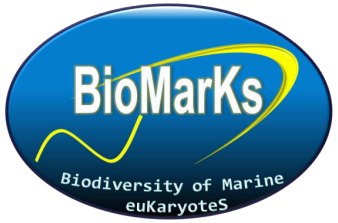
- Immédiate:
  - Parfaire la documentation
  - Optimisation
- Plus long terme:
  - Ajout d'un module de construction d'OTUs
  - Ajout d'un module d'assignation taxonomique

Quel intérêt de passer par Galaxy?

- Mettre à disposition de la communauté des outils difficilement installables par ailleurs / Allège l'aspect distribution
- Oblige à revoir son code
- Libère du temps (permettre à d'autres d'utiliser ses codes sans être obligé de la faire soi-même)
- Permet l'accès à de grandes puissances de calcul de manière transparente.

Remerciements

- Évolution du Plancton et Écosystèmes Pélagiques: C. de Vargas - J. Decelle - S. Romac - N. Henry - ...
- ABIMS: abims.sb-roscoff.fr - Poursuite de la suite de Metabarcoding4Galaxy dans le cadre d'un nouveau mini-projet?
- Données utilisées pour les tests



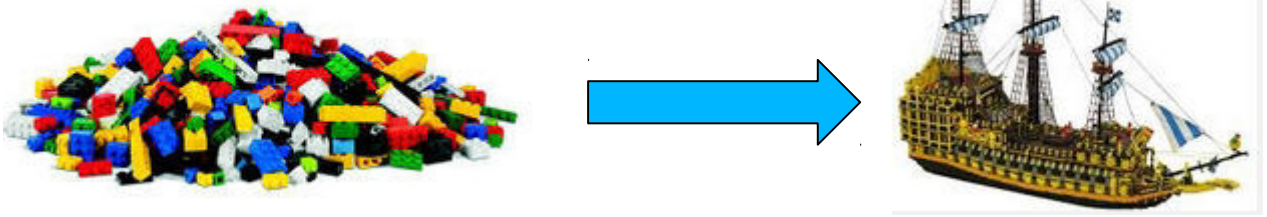
APEGE 2012
Divtrophie_Lagune
D. Grzebyk

Le projet OCEANOMICS bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme "Investissment d'Avenir" portant la référence ANR-11-BTBR-0008.

BioMarKs is a ERA net program funded by BiodivERsA's national partners.

Conclusion

- Galaxy couteau suisse
  - RNA-seq (Gen2Bio 2013), Métabolomique, RAD-seq, etc.
  - MétaGénique : 2 pipelines
- Modularité



- Dynamique
  - Régionale (GUGGO, etc.) & Nationale (IFB)
- Nouvelle communauté et adhésion
  - Besoin de soutien
- Un développement/intégration
  - Contactez-nous ?!