

Métagénomique

Bérénice Batut,

berenice.batut@udamail.fr

DUT Génie Biologique

Option Bioinformatique

Année 2014-2015

Métagénomique

Introduction

Principes de la métagénomique

Importance des métadonnées

Limites et freins

Métatranscriptomique

Métagénomique

Introduction

Principes de la métagénomique

Importance des métadonnées

Limites et freins

Métatranscriptomique

Diversité microbienne

Présence importante

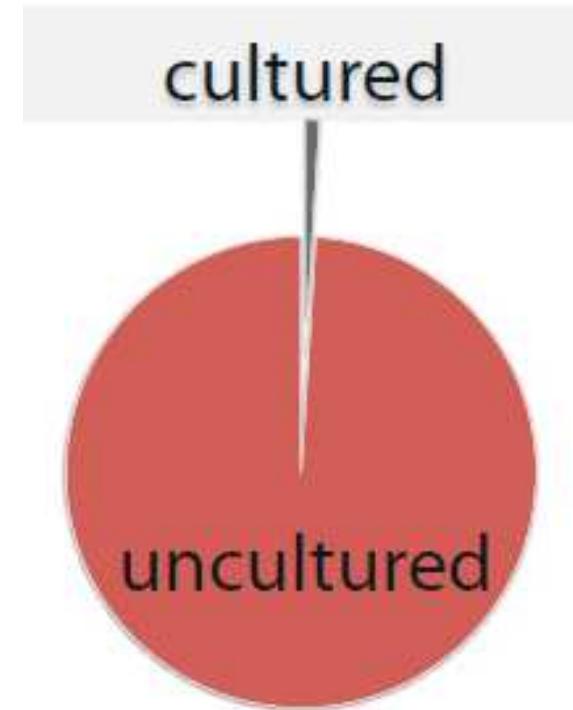
Majeure partie de la biomasse sur Terre

90% des cellules d'un être humain

Omniprésence

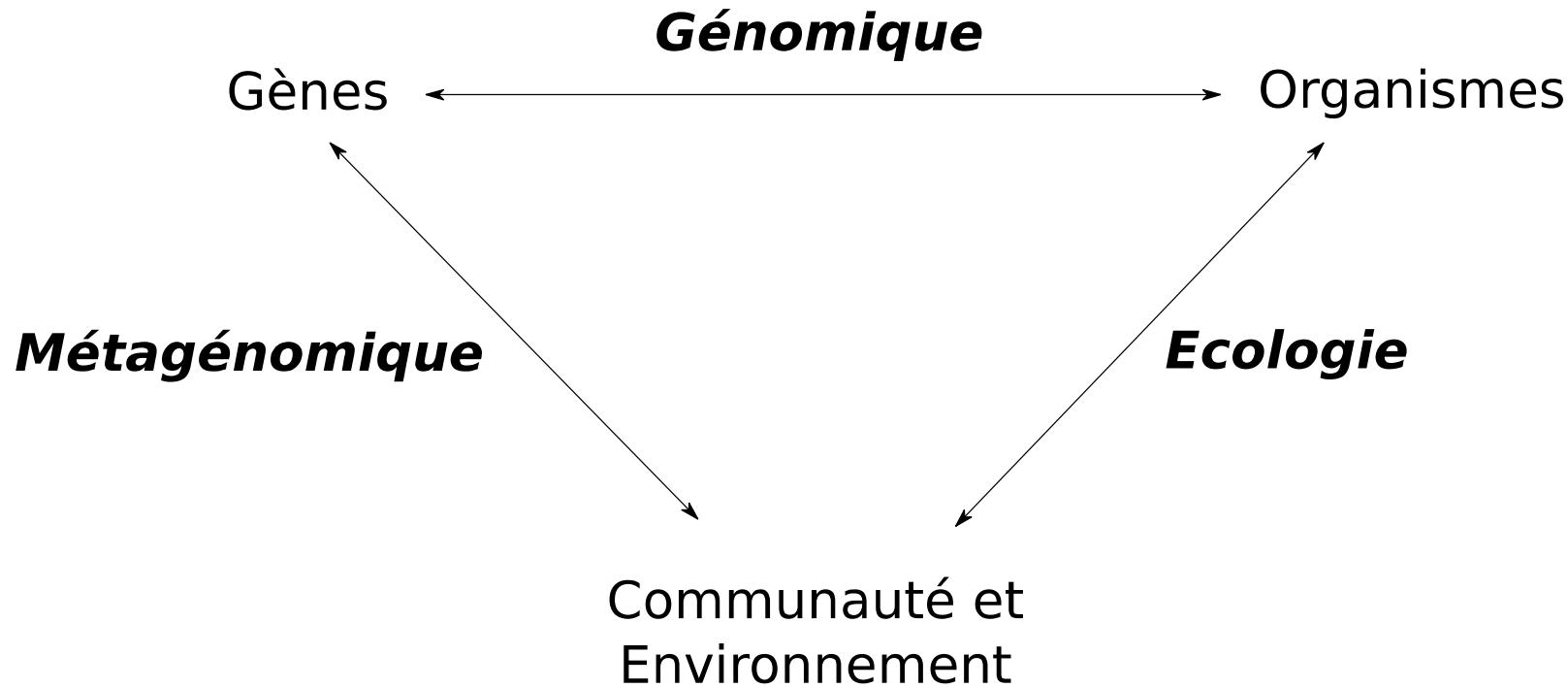
Plupart des niches écologiques

Importance métabolique



Rinke & Woyke

Métagénomique



Métagénomique

Etude de l'ADN des organismes non cultivés (*Handelsman et al, Chem Biol, 1998*)

Information génétique totale d'un ensemble d'organismes, au delà du génome (*Gilbert & Dupont, Annu Rev Marine Sci, 2011*)

Métagénomique

Pas besoin de culture

Identification

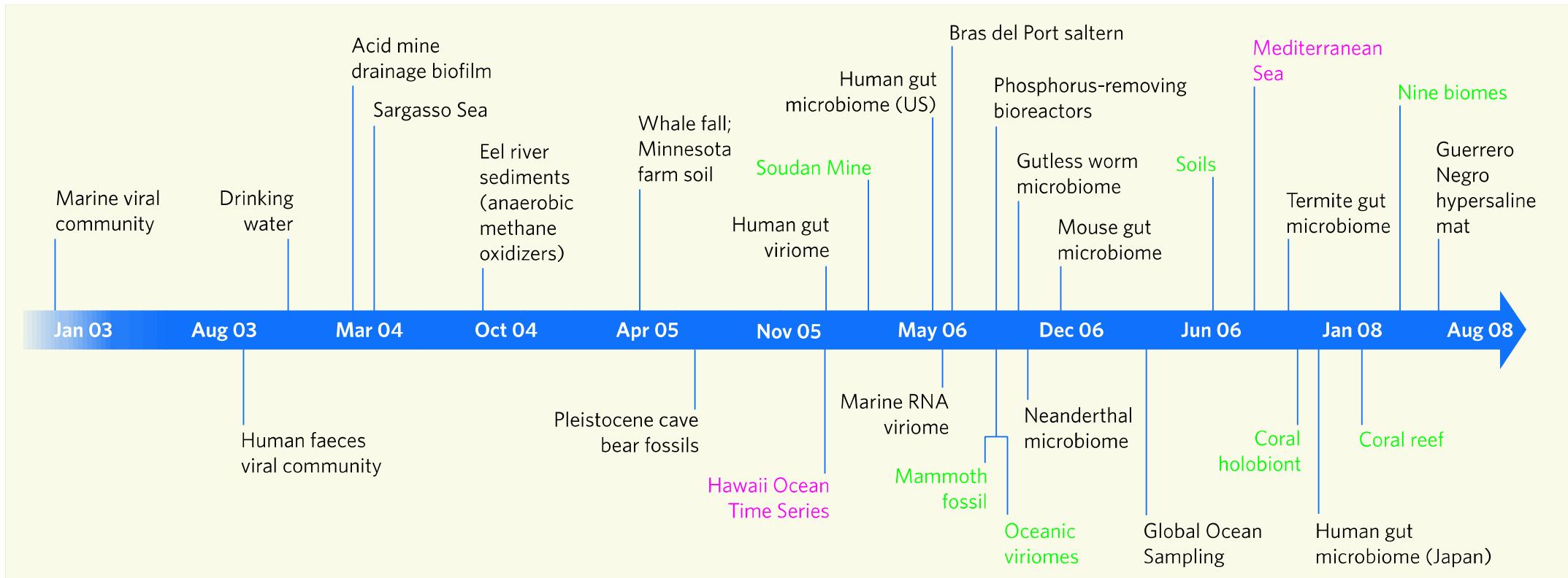
Membres de communautés

Gènes dans un environnement

Fonctions métaboliques caractérisant une communauté bactérienne

Compréhension du rôle écologique, du métabolisme et de l'histoire évolutive d'un écosystème

Projets métagénomiques



Projets métagénomiques

2128 projets métagénomiques sur NCBI

Sable de plage

Moustique

Corail

Glace

Fermenteur de biogaz

Air de la ville de Singapour

Fromages

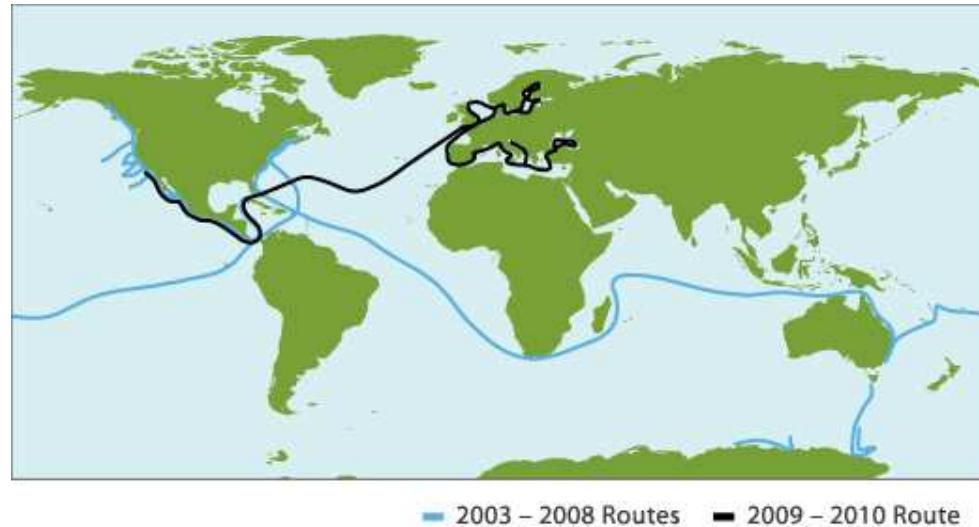
Surface de la cuvette des toilettes

...

Global Ocean Sampling (GOS)



Exploration des génomes des océans



Objectifs

Estimer la diversité génétique des communautés microbiennes marines

Comprendre leur rôle dans les processus fondamentaux de la nature

Global Ocean Sampling (GOS)



Production du plus grand catalogue de gènes de milliers de nouvelles espèces

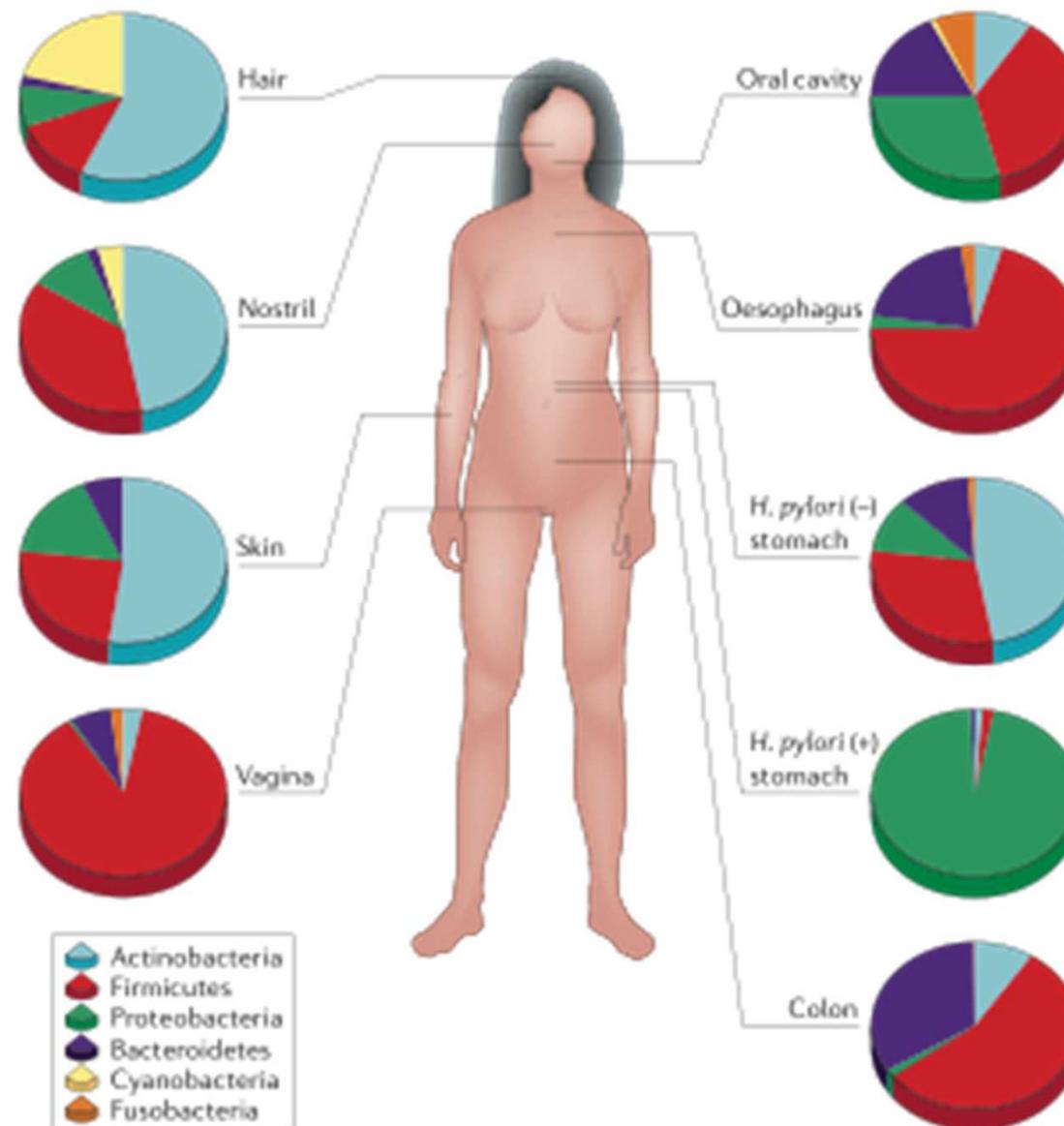
Données générées

3 087 469 séquences nucléotides

6 123 395 séquences protéiques identifiées

8 publications et 1 édition spéciale de PLoS Biology en 2007

Santé humaine



Santé humaine

Human Microbiome Project (HMP)

2,3 Tb de données métagénomiques

35 milliards de reads issus de 690 échantillons de 300 individus prélevés à 15 sites du corps



Metagenomics of the Human Intestinal Tract (MetaHIT)

540 Gb de séquences ADN

124 individus

American Gut Projet participatif



Métagénomique

Introduction

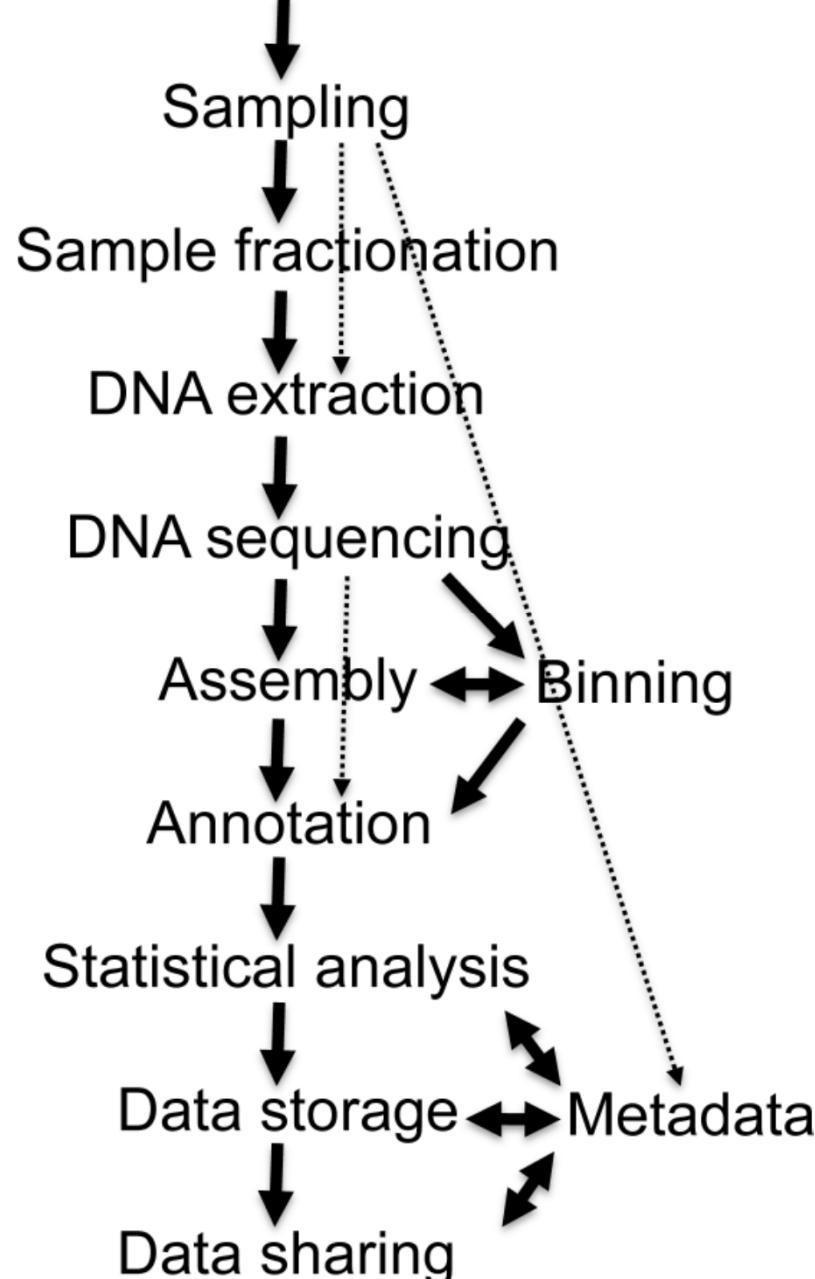
Principes de la métagénomique

Importance des métadonnées

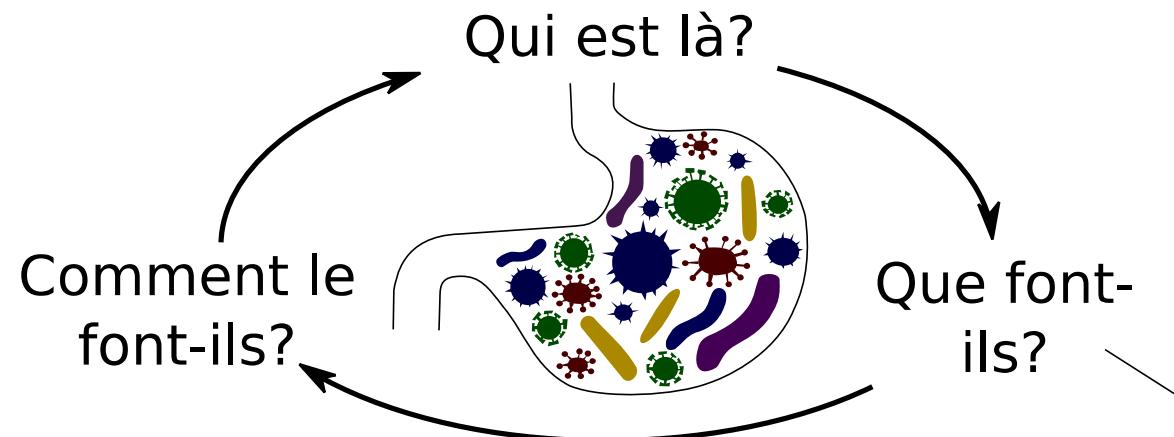
Limites et freins

Métatranscriptomique

Experimental design



Métagénomique



Objectifs

Identifier les organismes présents dans un échantillon

Identifier les fonctions principales effectuées dans un échantillon

3 types d'expériences métagénomiques

« Environmental clone libraries »

Séquençage Sanger après clonage

Etudes amplicon

Séquençage NGS des gènes ribosomaux amplifiés par PCR

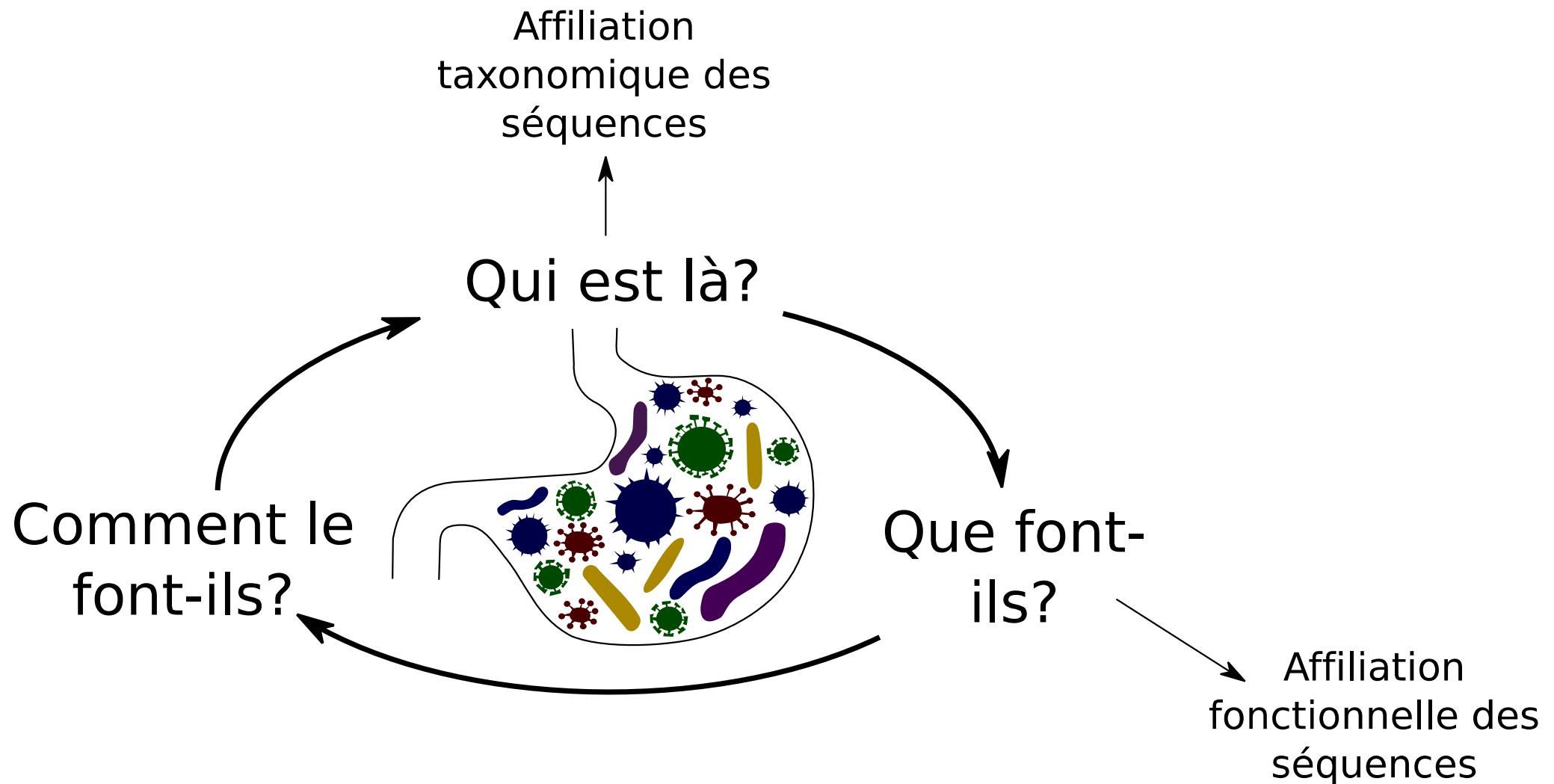
Identification des organismes présents

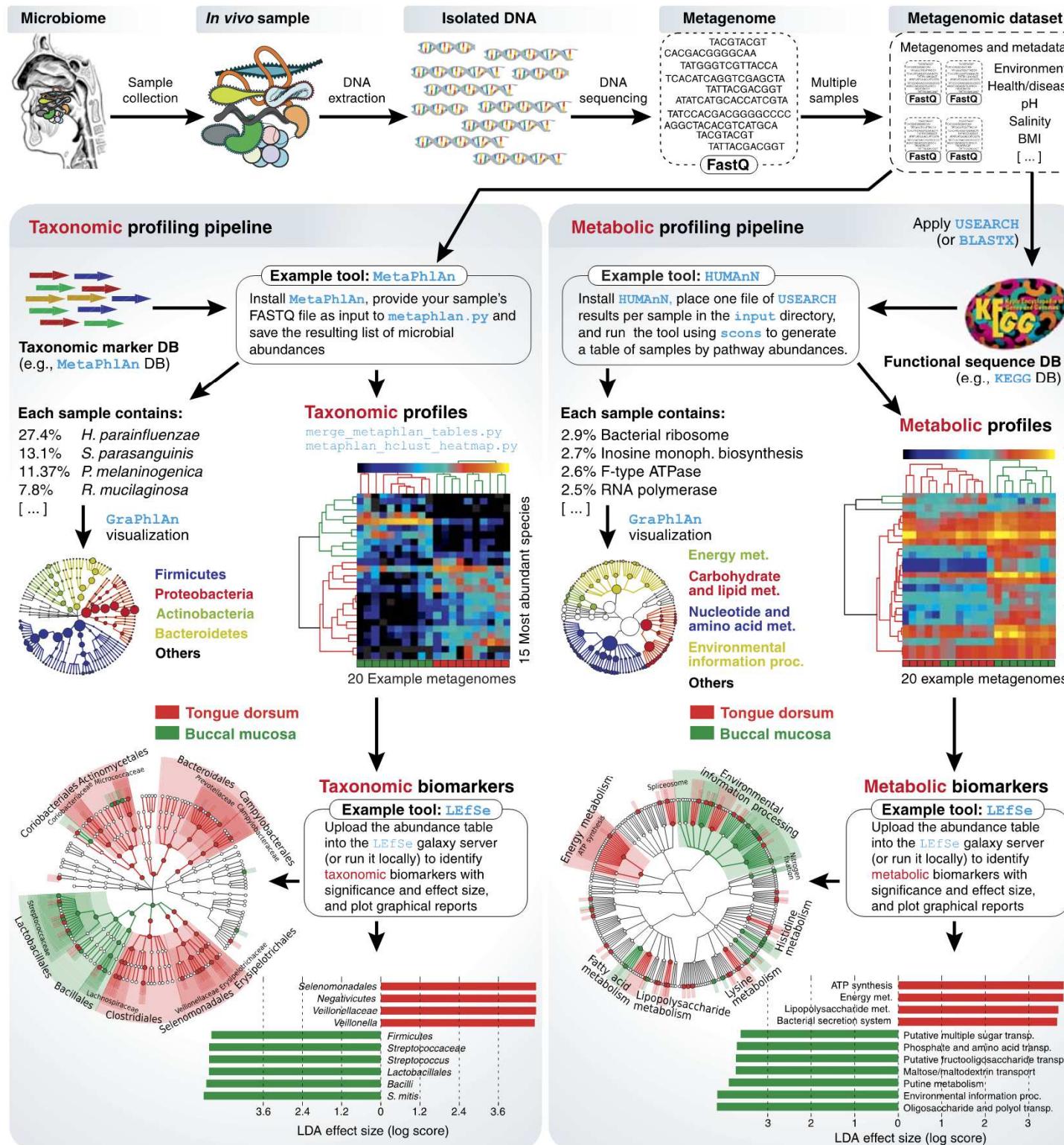
« Environmental shotgun metagenomics » ou WGS

Séquençage NGS de l'ensemble des séquences d'un échantillon

Identification des organismes présents et de leurs fonctions possibles

Métagénomique





Segata et al,
Mol Sys Biol, 2013
17

Métagénomique

Introduction

Principes de la métagénomique

Prétraitement

Analyse taxonomique

Analyse fonctionnelle

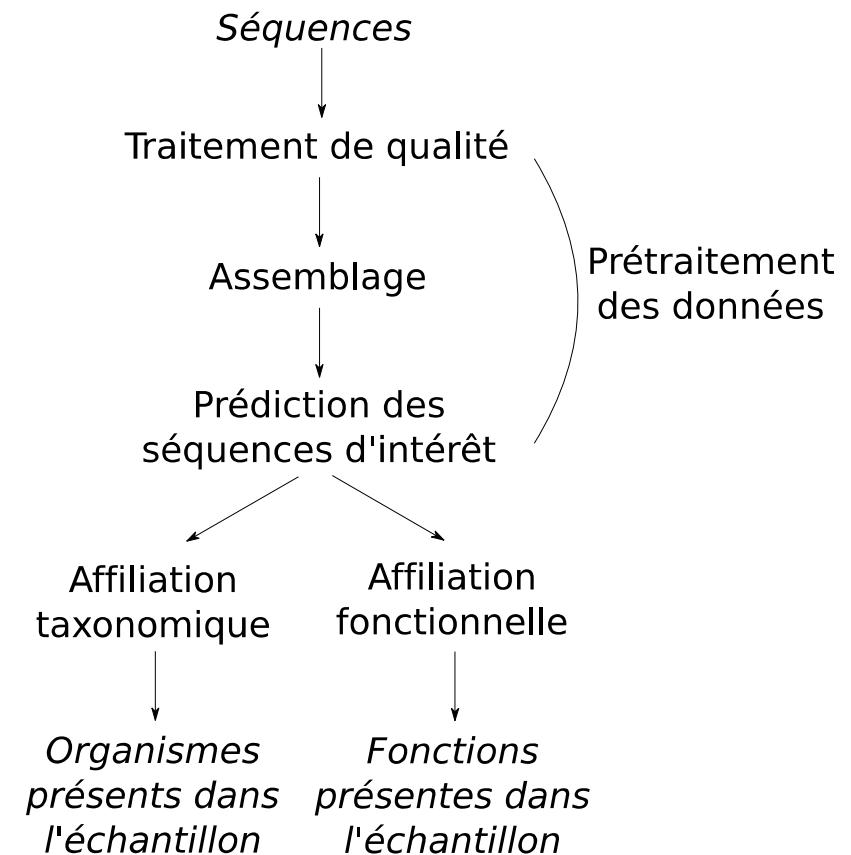
Comparaison des données

Pipelines déjà implémentés

Importance des métadonnées

Limites et freins

Métatranscriptomique



Métagénomique

Introduction

Principes de la métagénomique

Prétraitement

Analyse taxonomique

Analyse fonctionnelle

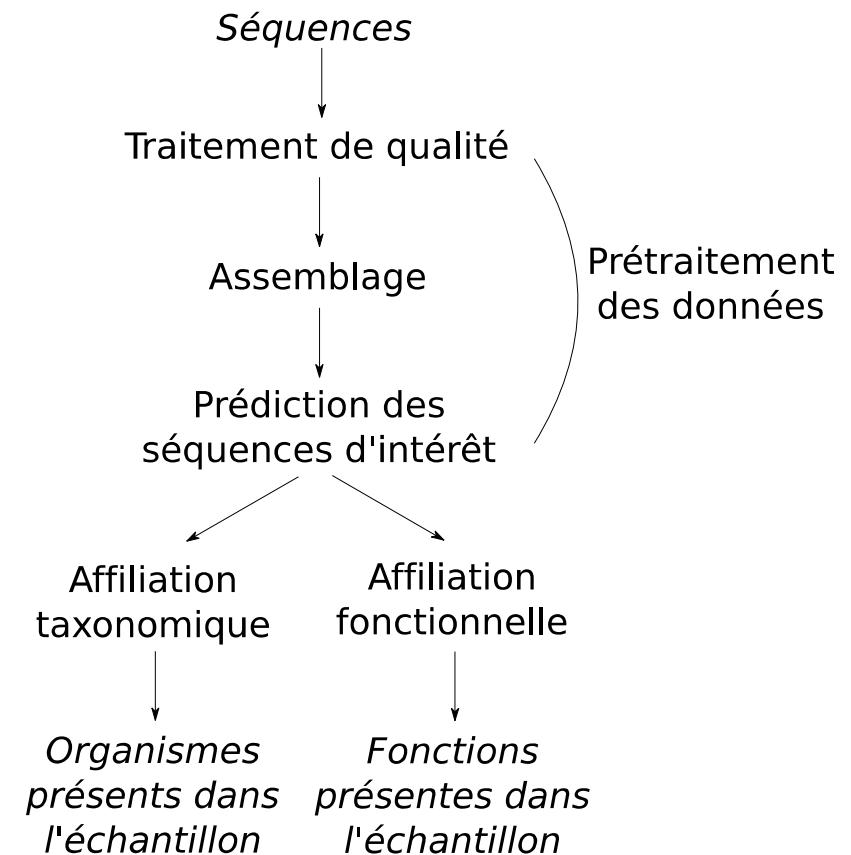
Comparaison des données

Pipelines déjà implémentés

Importance des métadonnées

Limites et freins

Métatranscriptomique



Assemblage

Séquences courtes

Assemblage nécessaire pour caractériser les séquences



“Metagenomics is like a disaster in jigsaw shop”

Assemblage

Attention aux chimères

Assemblage de séquences issues d'espèces différentes

Outils spécifiques pour l'assemblage des données métagénomiques

MetaVelvet, MetaIDBA, khmer

Genovo

Euler

SOAPdenovo, MetAMOS, Newbler

Identification des caractéristiques d'intérêt

Caractéristiques d'intérêt

ORF

Unités taxonomiques opérationnelles (OTU)

Gènes ARNt, ARNr, ...

Utiles pour la recherche de similarité dans les bases de données de séquence

Outils spécifiques

ORF : MetaGeneAnnotator, Prodigal, Orphelia...

OTU : usearch, mothur, uclust, ...

ARNt : tRNAscanSE

ARNr : rRNA selector, Metaxa, ...

Métagénomique

Introduction

Principes de la métagénomique

Prétraitement

Analyse taxonomique

Analyse fonctionnelle

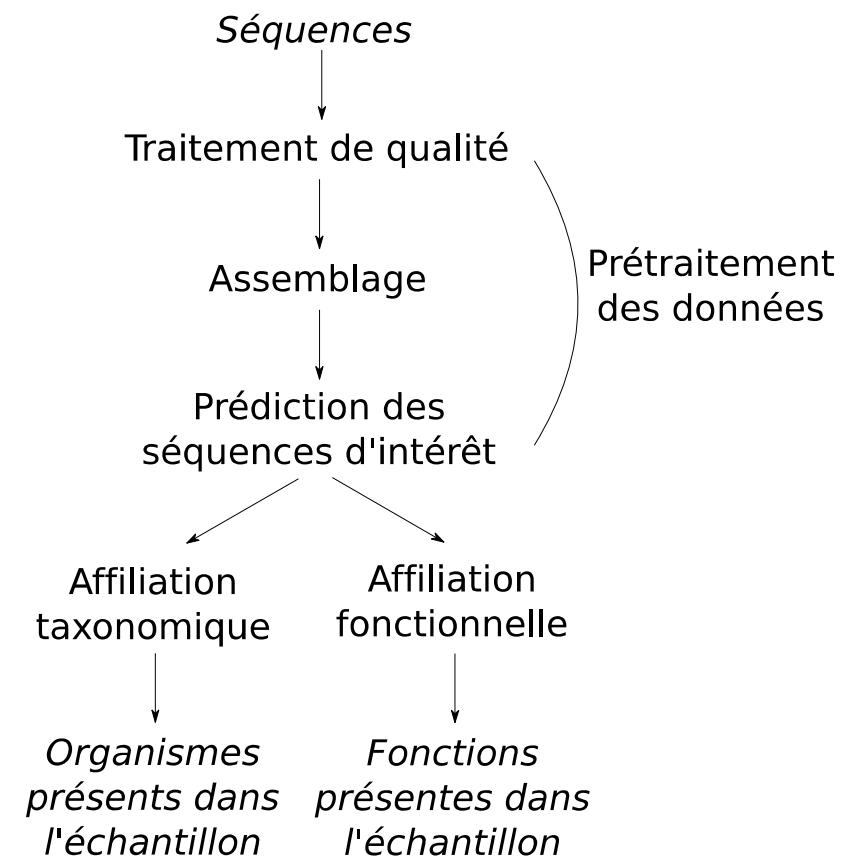
Comparaison des données

Pipelines déjà implémentés

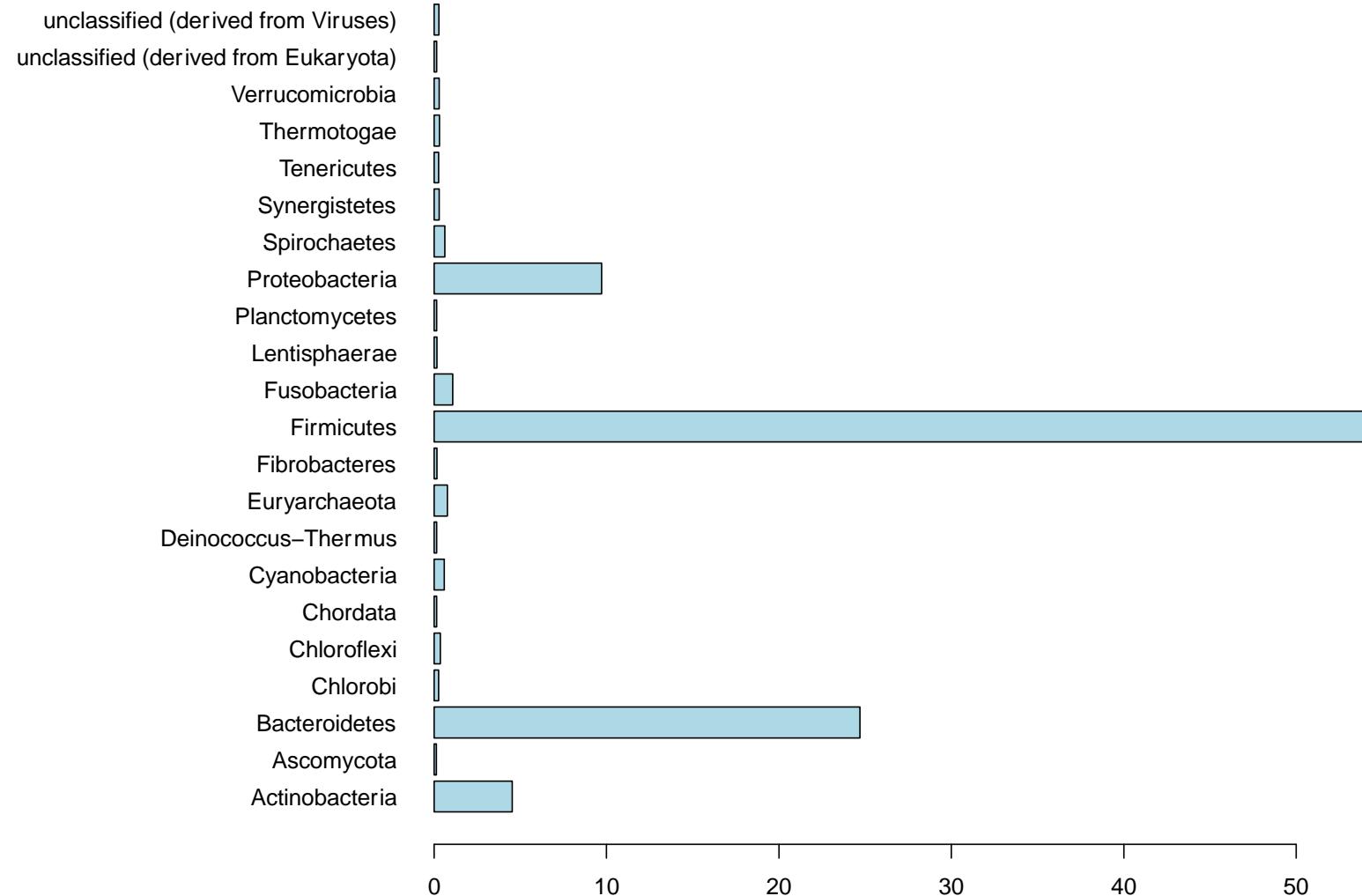
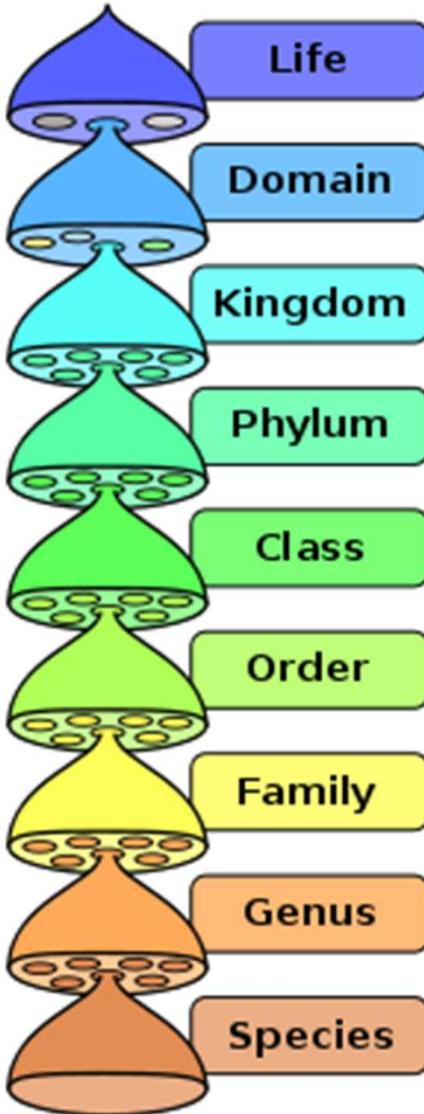
Importance des métadonnées

Limites et freins

Métatranscriptomique

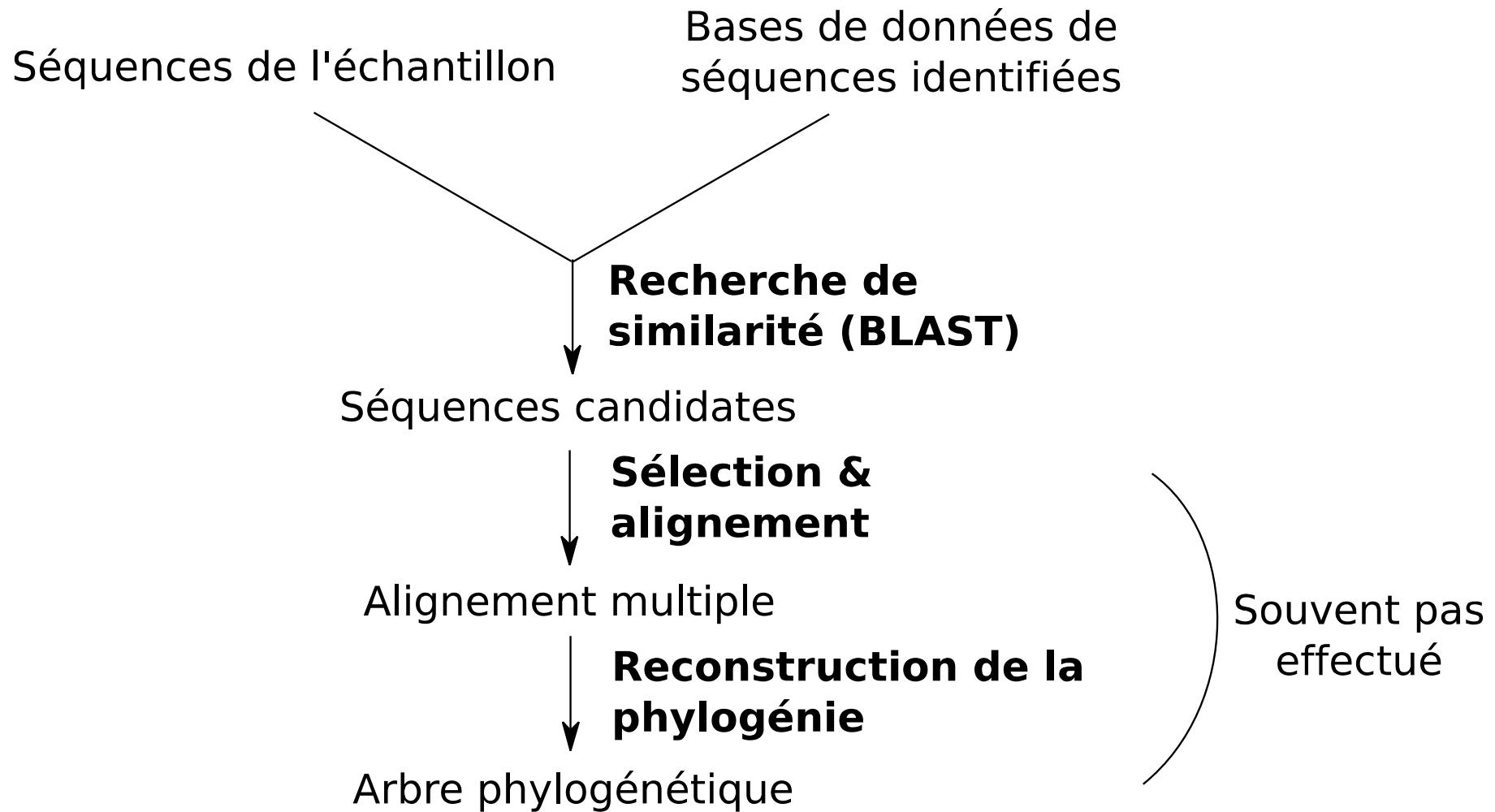


Pourquoi l'affiliation taxonomique?



<https://metagenomics.anl.gov/?page=MetagenomeOverview&metagenome=4448044.3>

Procédure standard d'affiliation taxonomique



Limites de la procédure standard d'affiliation taxonomique

Longueur des séquences obtenues

Absences d'homologues dans les banques

Nouvelles espèces

Faible abondance des marqueurs phylogénétiques classiques (ARNr 16S ou 18S)

Boues d'épuration : 0.17%

Mer des Sargasses : 0.06%

Sol acide, mine du Minnesota : 0.017%

Autres approches pour l'affiliation taxonomique

Clustering

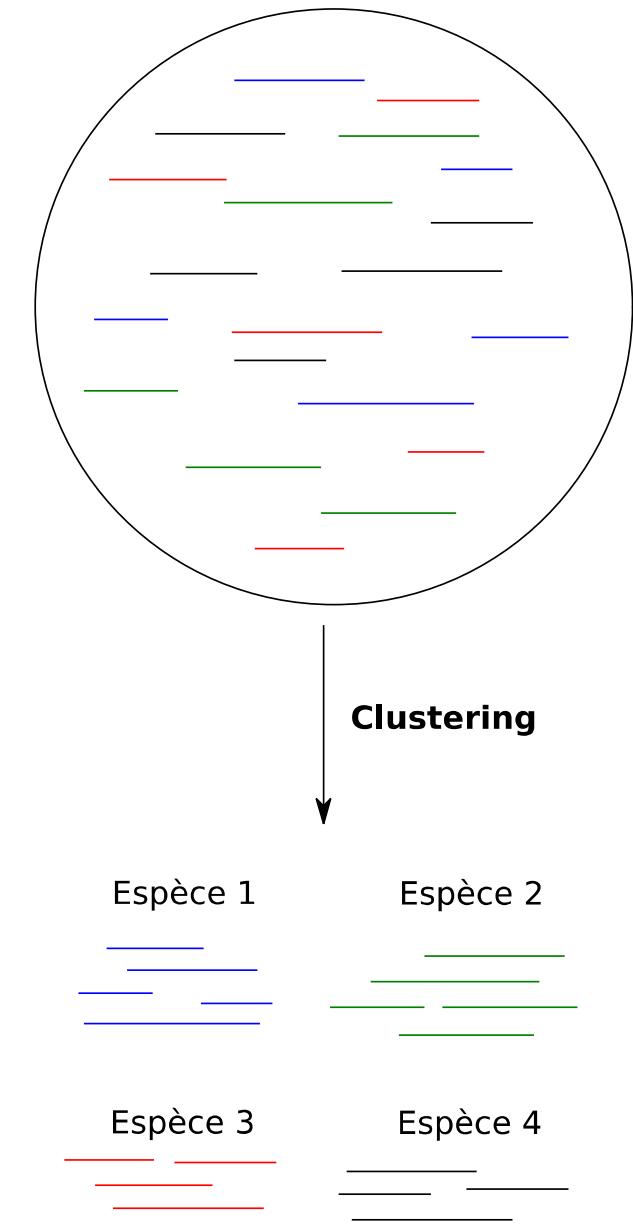
Approches comparatives

Recherche de similarités contre des séquences de références issues des bases de données

Recherche de similarités intrinsèques

Contenu en GC

Usage des codons



Diversité alpha

Mesure de la biodiversité d'un échantillon

Notion définie par Whittaker (Evolution and Measurement of Species Diversity, 1972)

Mesure, pour un échantillon, de
Richesse

Nombre de groupes d'individus génétiquement liés, nombre d'espèces

Egalité

Proportion d'espèces présentes

Plus les espèces sont présentes en proportion équivalentes, plus grande est l'égalité



Grand nombre d'espèces avec
des abondances similaires

Diversité
alpha

Métagénomique

Introduction

Principes de la métagénomique

Prétraitement

Analyse taxonomique

Analyse fonctionnelle

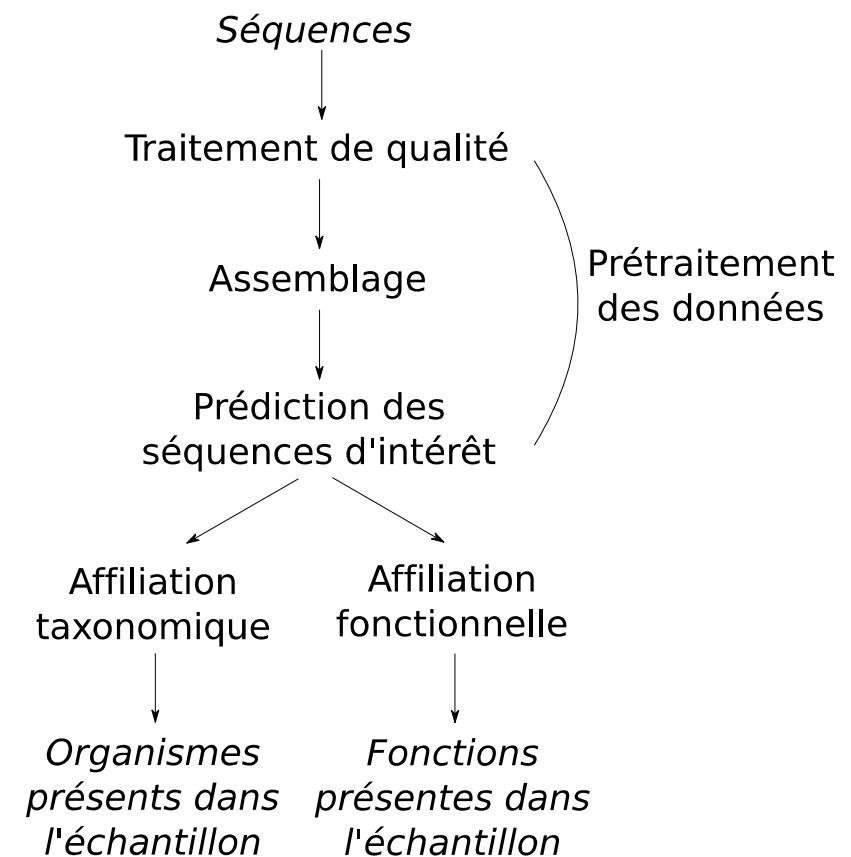
Comparaison des données

Pipelines déjà implémentés

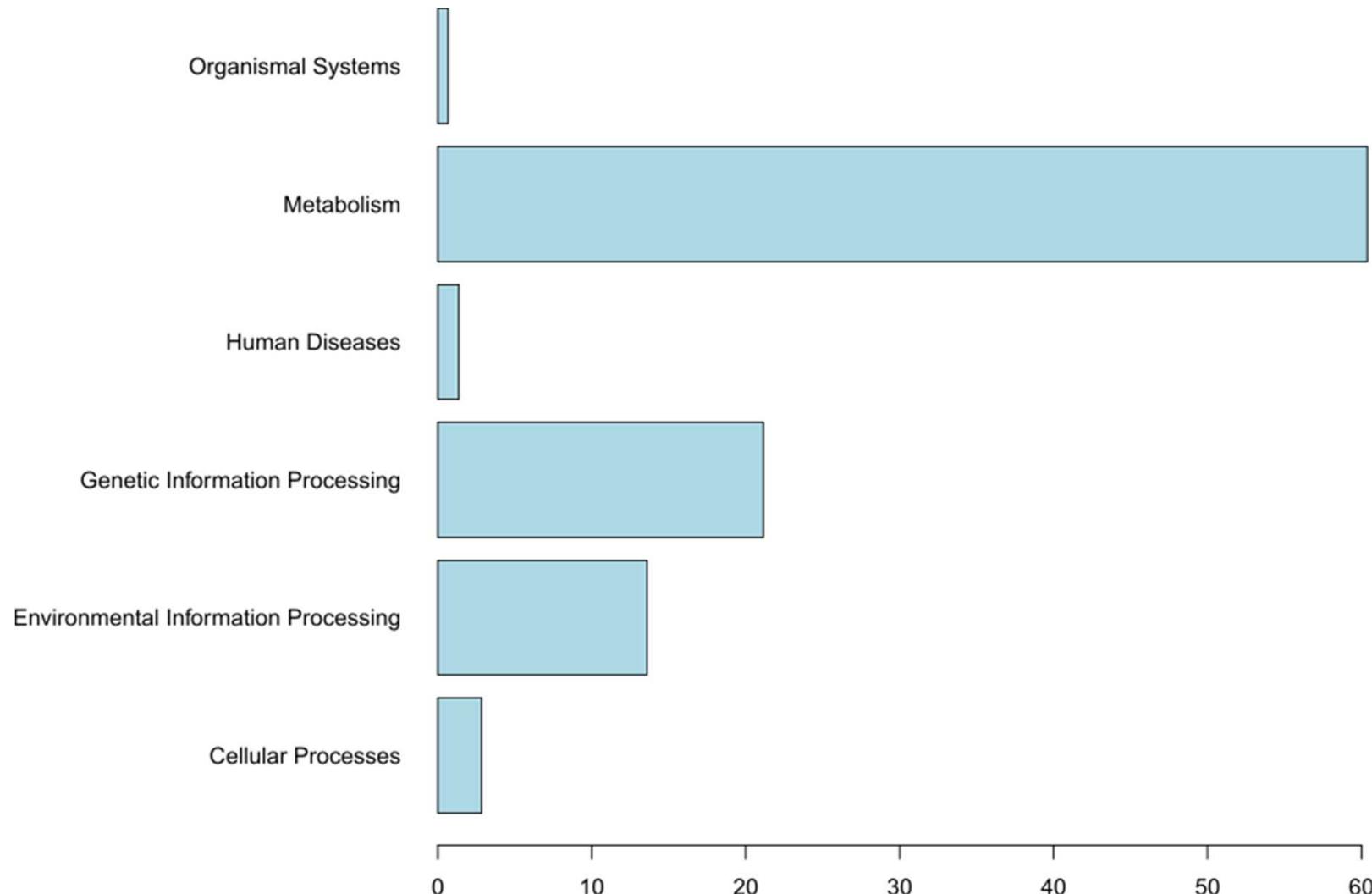
Importance des métadonnées

Limites et freins

Métatranscriptomique

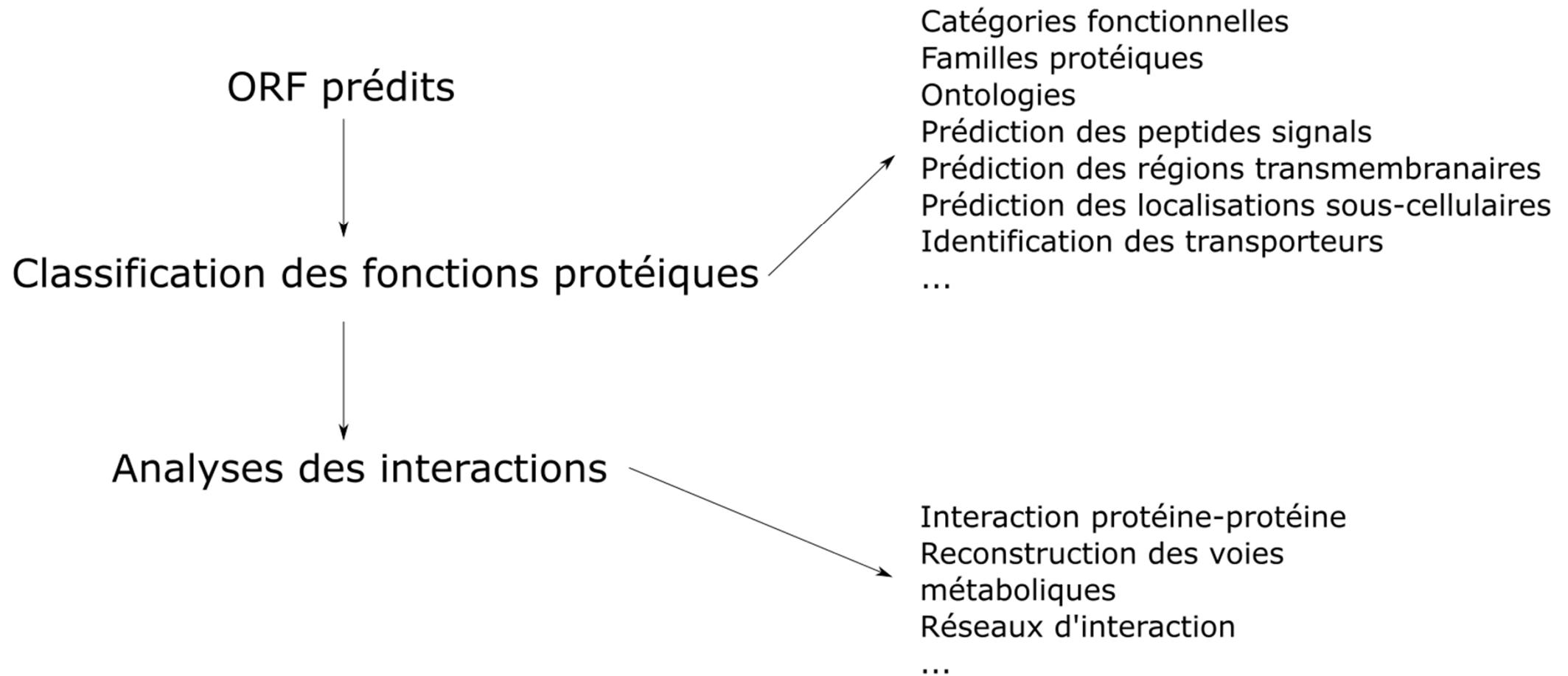


Pourquoi l'affiliation fonctionnelle?



[https://metagenomics.anl.gov/?page=Metagenome
Overview&metagenome=4448044.3](https://metagenomics.anl.gov/?page=MetagenomeOverview&metagenome=4448044.3)

Affiliation fonctionnelle



Classification des fonctions protéiques

Recherche contre différentes bases de données de familles de séquences

Cluster of orthologous groups of proteins (COGs)

SEED

Kyoto Encyclopedia of Genes and Genomes (KEGG)

Pfam

Evolutionary genealogy of genes: non supervised orthologous groups (eggNOG)

...

Recherche de motifs fonctionnels

Interpro

Prosite

...

Métagénomique

Introduction

Principes de la métagénomique

Prétraitement

Analyse taxonomique

Analyse fonctionnelle

Comparaison des données

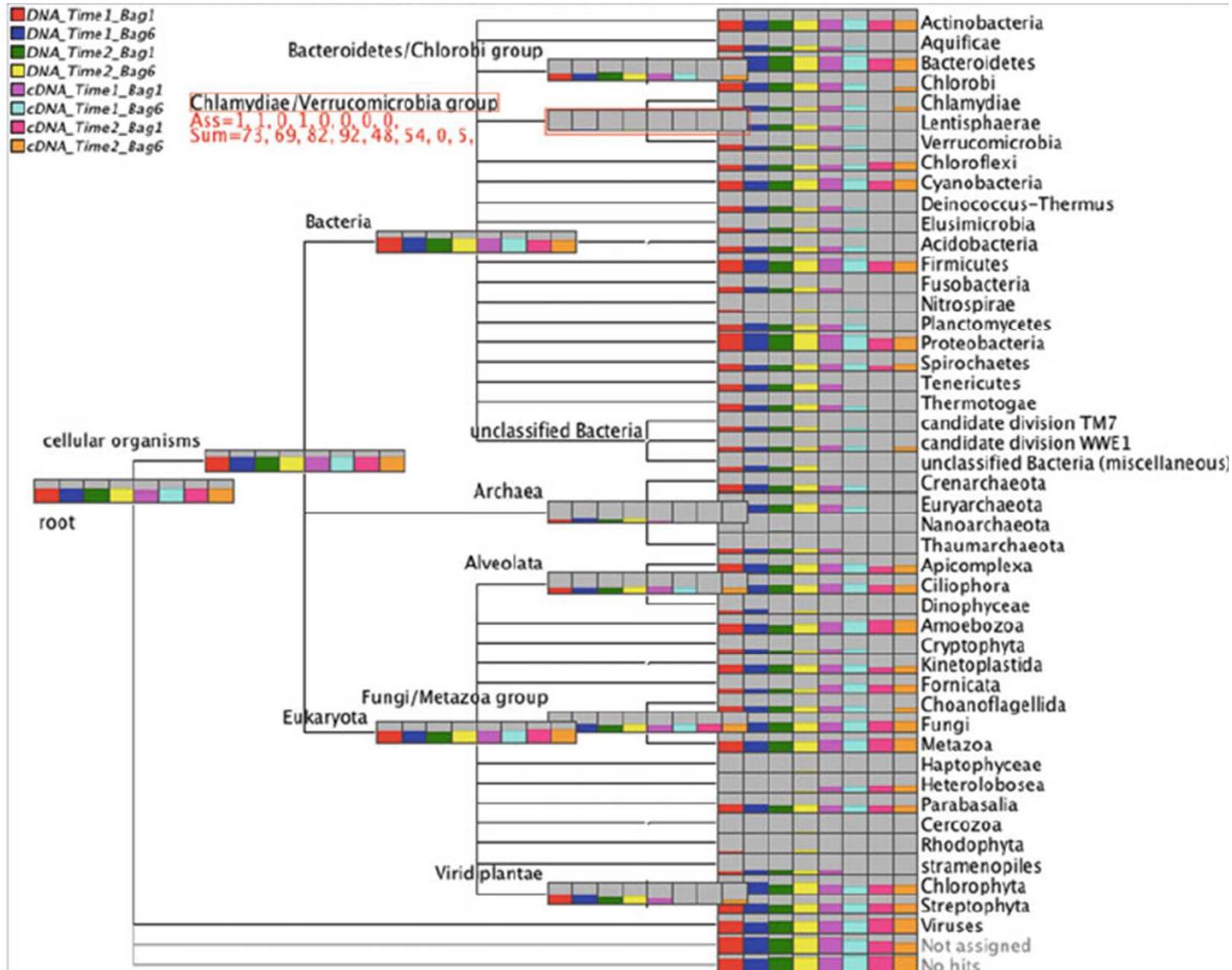
Pipelines déjà implémentés

Importance des métadonnées

Limites et freins

Métagénomique

Comparaison des organismes présents dans plusieurs échantillons



Diversité beta

Comparaison de la diversité entre des échantillons

Notion définie par Whittaker (Evolution and Measurement of Species Diversity, 1972)

$$\beta = \frac{S}{\bar{\alpha}}$$

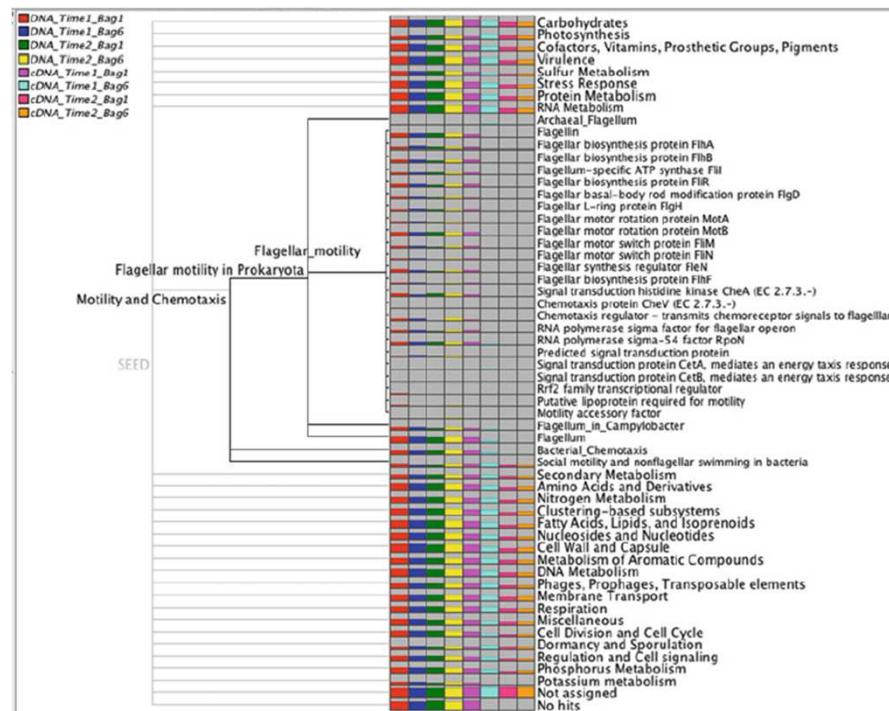
Avec S le nombre total d'espèces enregistrées dans les deux communautés, $\bar{\alpha}$ la moyenne du nombre d'espèces trouvées au sein des communautés

Compare entre deux échantillons la différence de communautés

Différences des espèces ou OTU

Comparaison fonctionnelle d'échantillons

Différences de gènes/fonctions/voies métaboliques



Indicateur des variables environnementales affectant le potentiel fonctionnel des échantillons

Métagénomique

Introduction

Principes de la métagénomique

Prétraitement

Analyse taxonomique

Analyse fonctionnelle

Comparaison des données

Pipelines déjà implémentés

Importance des métadonnées

Limites et freins

Métagénomique

Principaux pipelines d'analyse des données métagénomiques

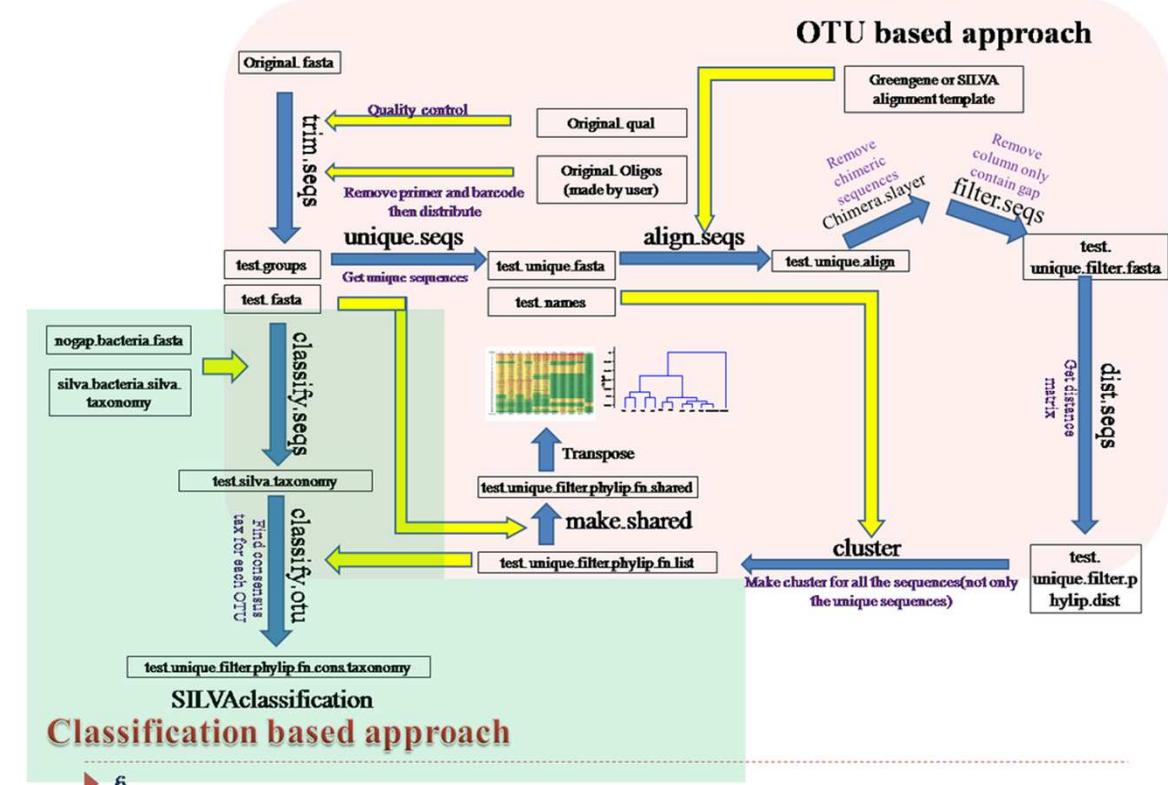
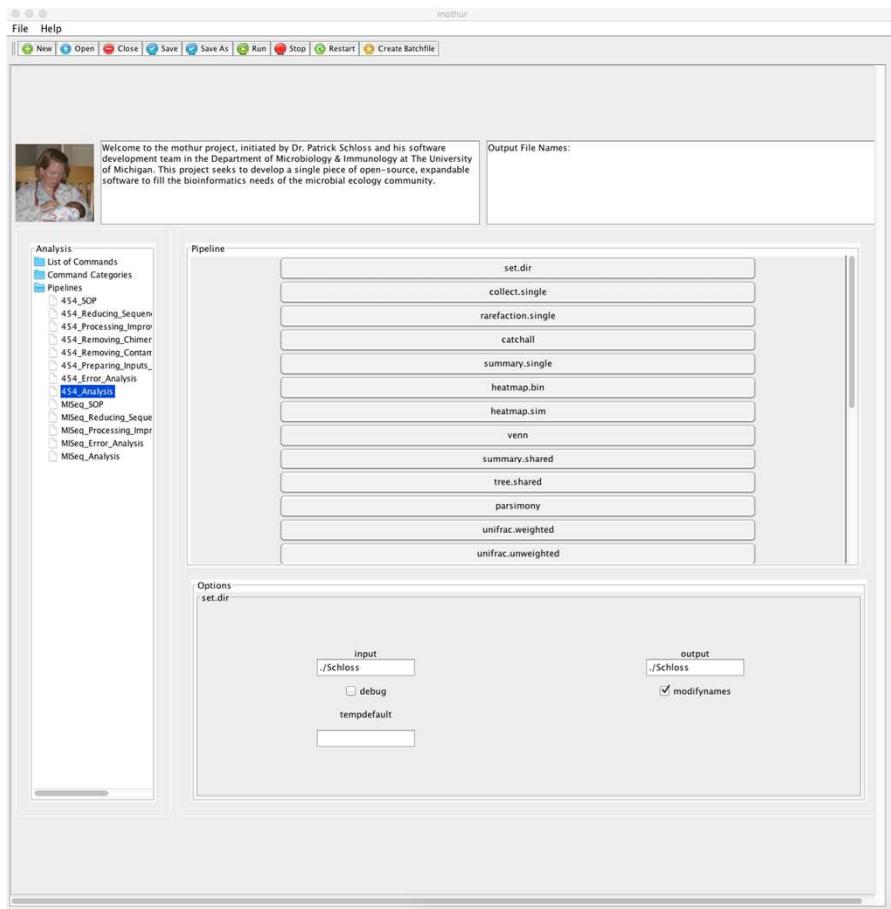


mothur

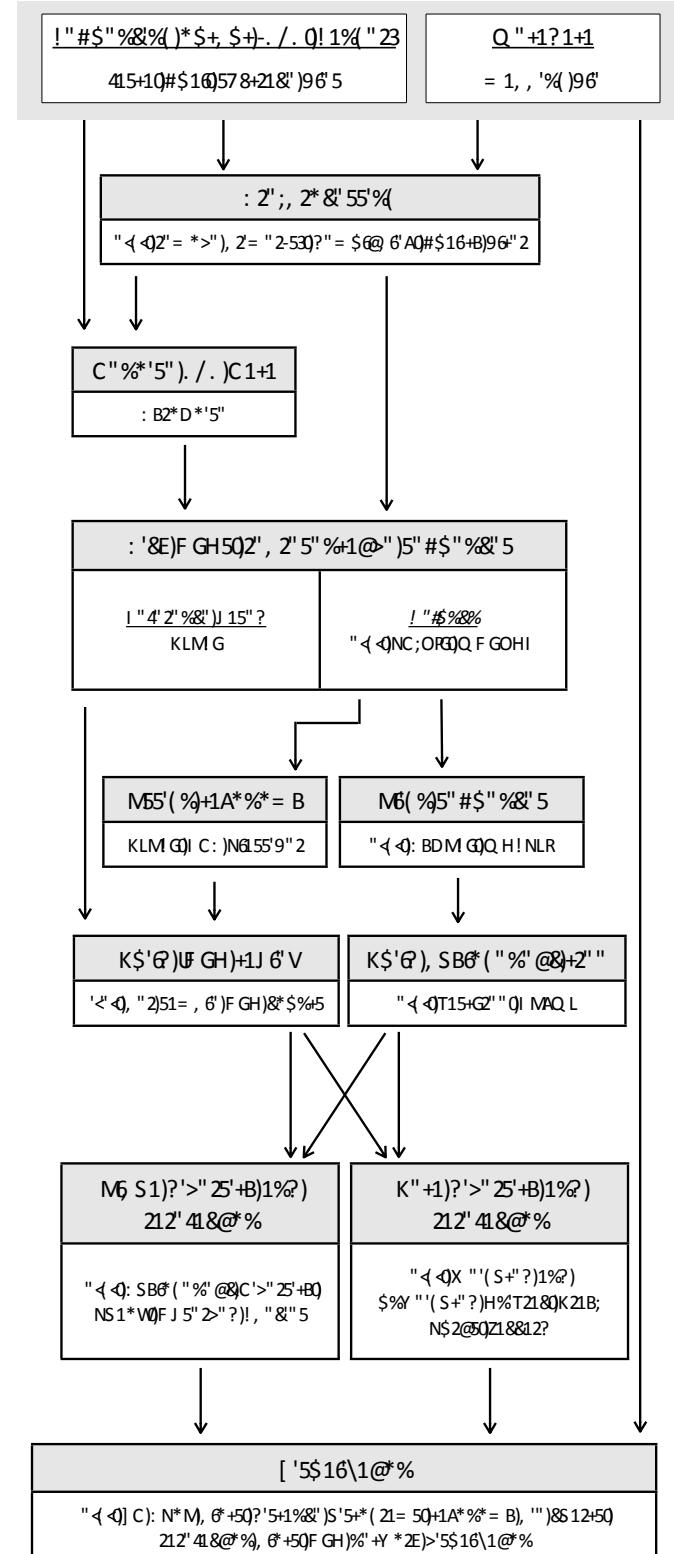
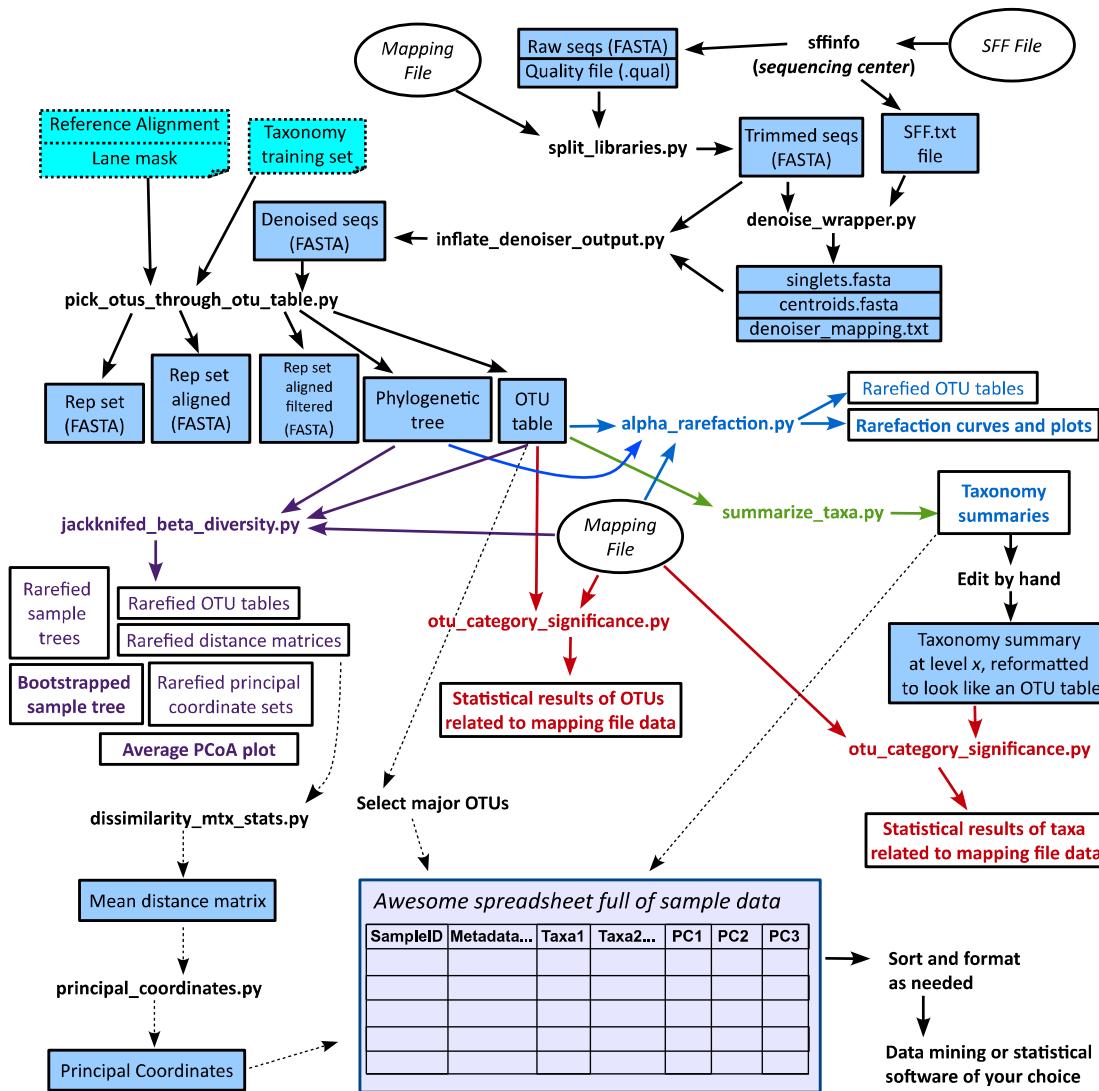


mothur

Schloss et al, Applied Env Microbiol, 2009



Carporaso et al, Nature Methods, 2010



AXIOME

AXIOME – Automation, Extension, and Integration Of Microbial Ecology: Helper tools to make managing numerical ecology pipelines easier



Quantitative Insights Into Microbial Ecology

mothur

Lynch et al, GigaScience,
2013

1 Describe the data
and analyses

```
<?xml version="1.0"?>
<axiome>
<def name="" type="" />
<panda forward="" reverse="" version="" />
<sample tag="" />
</panda>
<alpha/>
<beta/>
<mrrp/>
<duleg/>
</axiome>
```

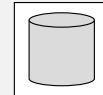
2 Run AXIOME on the
file

```
$ axiome exp.ax
```

3 Invoke make to run
analyses

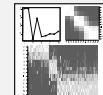
```
$ make
```

AXIOME Plugins (version 1.6)



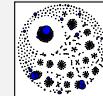
PANDAseq

PAired-eNDA Assembly of Illumina sequences
<panda />



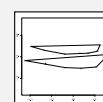
Non-negative Matrix Factorization

Alternative higher-dimesional decomposition
<nmf />



SSUnique

Unique lineage finder
<ssunique />



MRPP Analysis

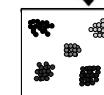
Multiple Response Permutation Procedure
<mrrp />



User-Defined Analyses



Quantitative Insights Into Microbial Ecology



Clustering

ulclust or CD-HIT sequence clustering
<otu-method="" />



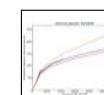
BIOM

OTU Table



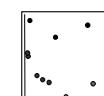
Tree Building

Building phylogenetic trees
<phylogeny-method="" />



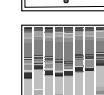
α-diversity

Chao1, phylogenetic diversity/richness
<alpha />



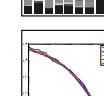
β-diversity

Rarified/summarized by taxonomic level
<beta />



Taxonomic Composition

<taxaplot />



Rank-Abundance

<rankabundance />

AXIOME: Define Metadata

Name	Type
Description	String
Colour	String
Patient	Integer
SampleLocation	String

AXIOME: Define Source Files

Type	File Path
FASTA	/media/HDD/patient1.fa
PANDA	/media/HDD/patient2_forward.fq

AXIOME: Define Samples (File 2 of 2)

Tag	Description	Colour	Patient	SampleLocation
ATTAC	P2-Mouth	red	2	mouth
CGAGTT	P2-Stomach	blue	2	stomach
CGGACT	P2-Colon	green	2	colon

AXIOME: Define Methods and Analyses

Pipeline: qiime

Multi-Core System Number of Cores: 16

Alignment Method: infernal

OTU Method: cdhit

OTU Flags:

OTU Blast DB:

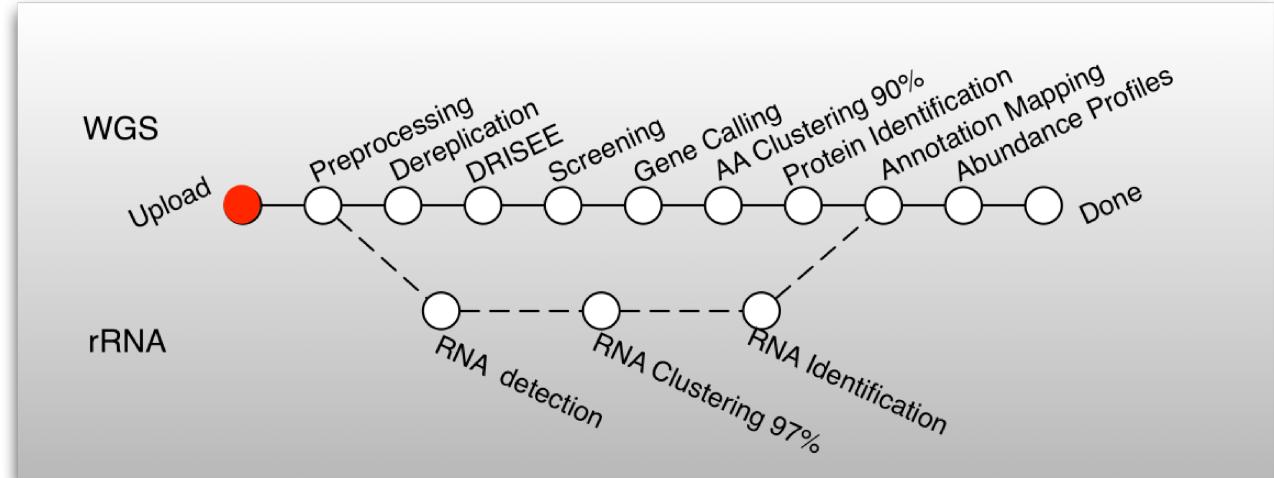
Verbose

Analysis	Parameters
pcoa	method="bray"
mrrp	method="bray"
alpha	
taxaplot	
blast-classifier	seqfile="/home/michael/ref_seqs.fa" taxfile="/home/michael/ref_taxa.txt"

MG-RAST

metagenomics analysis server

Meyer et al, BMC Bioinformatics,
2008



Metagenome Analysis

① Data Type

- ORGANISM ABUNDANCE
- Representative Hit Classification
- » Best Hit Classification**
- Lowest Common Ancestor

FUNCTIONAL ABUNDANCE

- Hierarchical Classification
- All Annotations

OTHER

- Recruitment Plot

② Data Selection

Metagenomes: 4440452.7

Annotation Sources: M5NR, 1e-5, 60 %, 15

Max. e-Value Cutoff: 1e-5

Min. % Identity Cutoff: 60 %

Min. Alignment Length Cutoff: 15

use features from workbench

③ Data Visualization

Analysis Views: Organism Classification

Visualizations:

- barchart
- tree
- table
- heatmap
- PCoA
- rarefaction

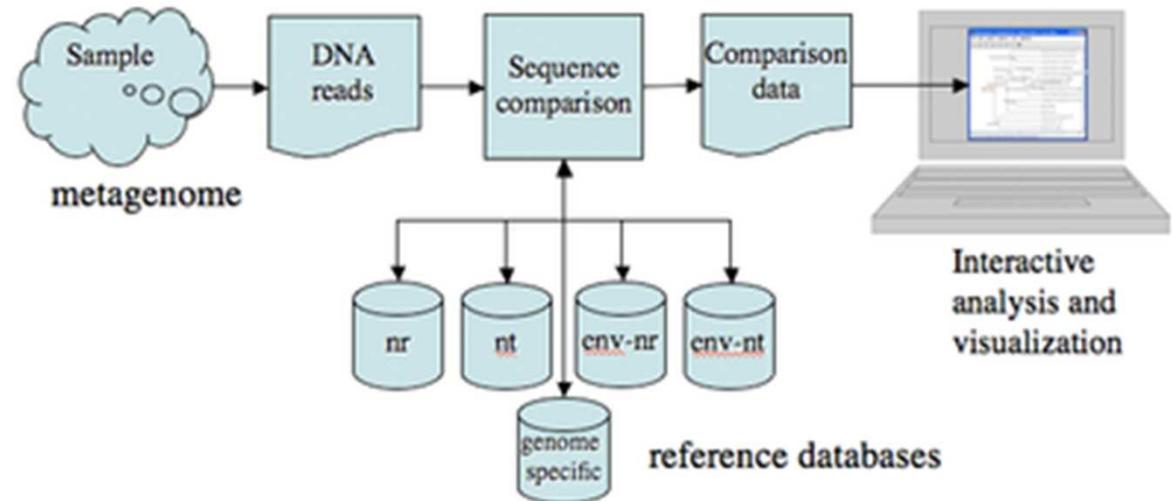
generate

Workbench (0 Features) Getting Started x

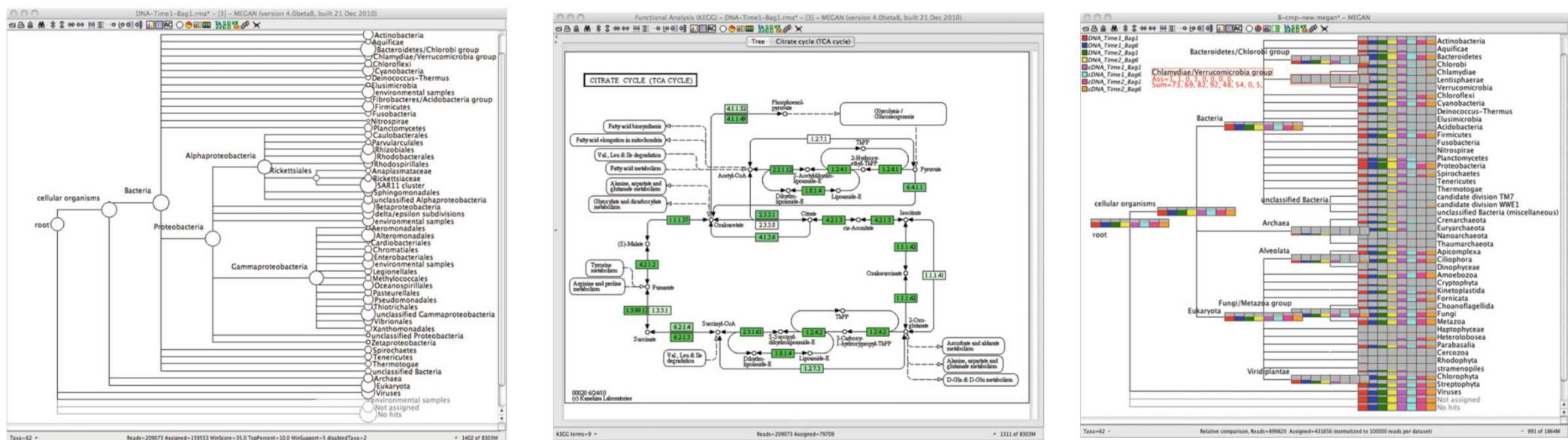
To create a visualization, first select an analysis view from the **Analysis Views** box. The default is 'Organism Classification'. Then choose the data and cutoffs you wish to use in the **Data Selection** box. Depending on the type of data, you might have a set of possible visualizations. Pick one of them and click the **generate** button.

You will see the generated visualizations created in separate tabs in this tab-view. In addition to the visualization, the created tabs will display the settings used to create them. Generating a new visualization will preserve the previous visualizations you created.

You can rename the tabs by double clicking the tab-header and entering a new title. You can remove a tab by clicking the 'x' symbol in the top right corner of the tab-header. You can switch between tabs by single clicking the tab-header.



Huson et al, Genome Res, 2007; Huson et al, Genome Res, 2011; Huson & Mitra, Methods Mol Biol, 2012





CloVR

Anguioli et al, BMC Bioinformatics, 2011

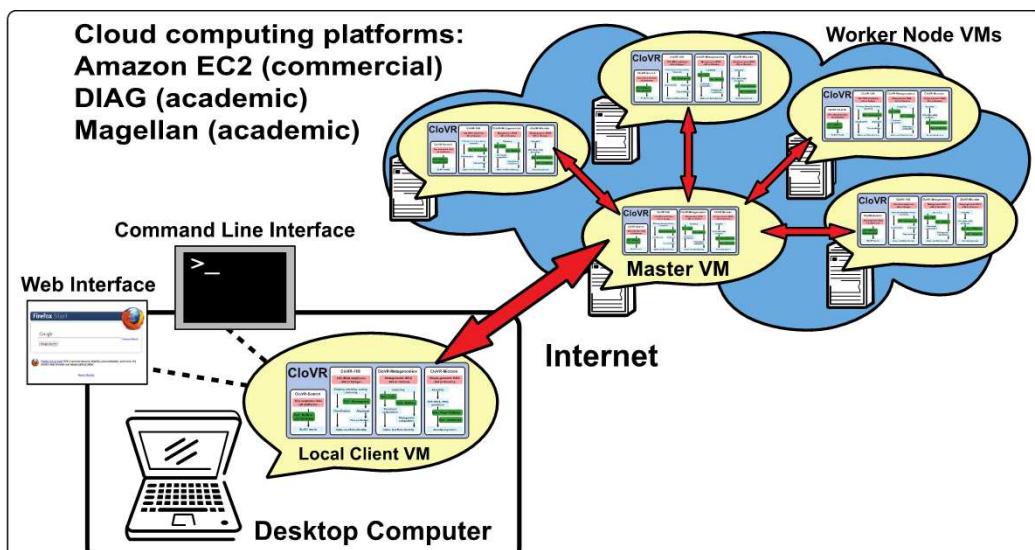
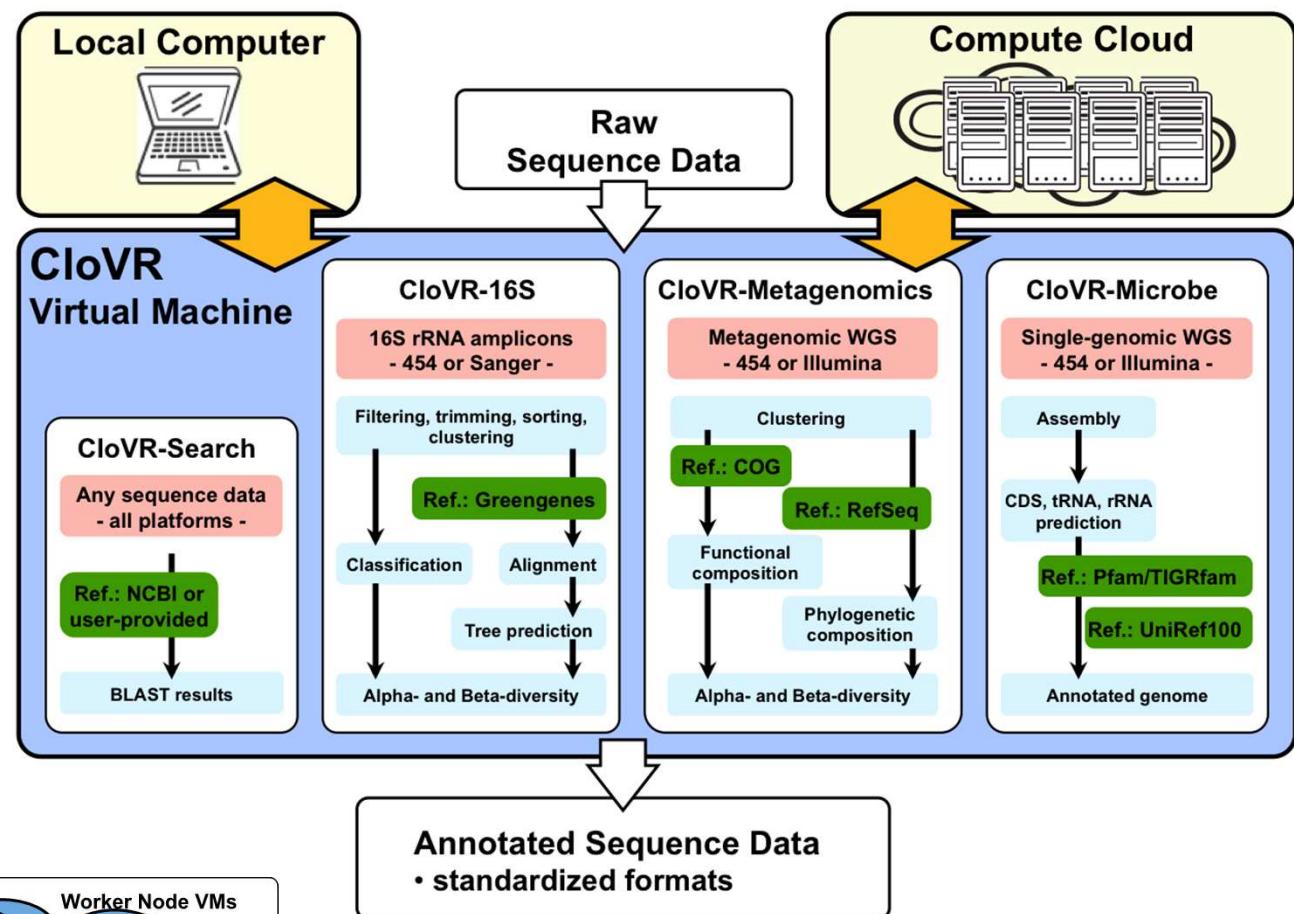


Figure 2 Architecture of the CloVR application. CloVR provides a virtual machine (VM) that is run on user's local desktop or laptop computer. The user interacts with the local VM via a command line or web interface to execute pipelines. Optionally, clusters of additional VM instances are provisioned on supported cloud platforms for increased throughput. Each cluster has a master VM instance that provides services for GridEngine [36] and Hadoop [37]. Input data and output data is transferred between the local VM and a master VM instance in the cloud over the Internet.

Métagénomique

Introduction

Principe de la métagénomique

Importance des métadonnées

Limites et freins

Métatranscriptomique

Que sont les métadonnées ?

Description en profondeur et contrôlée de l'échantillon

Où : Latitude et longiture, profondeur, salinité, ...

Quand

Quoi

Comment : Processus de conservation, condition de stockage, méthode d'extraction, plateforme de séquençage , ...

Indispensables pour partager et intégrer les données dans la communauté génomique

Besoin d'un vocabulaire standardisé minimal

Genome Standard Consortium (GSC)

Organisme international, ouvert, formé en 2005

Objectif : Promouvoir

Mécanismes de standardisation de la description des génomes

Echange et intégration des données génomiques

Mission

Implémentation de nouveaux standards (méta)génomiques

Méthodes de capture et d'échange des métadonnées

Harmonisation de la collection de métadonnée et des efforts d'analyse à travers toute la communauté génomique

Standards des métadonnées

MixS

Minimum Information about any Sequence

MIGS/MIMS

Minimum information about a (Meta)Genome Sequence

MIMARKS

Minimum Information about a MARKer gene Sequence

Métagénomique

Introduction

Principes de la métagénomique

Importance des métadonnées

Limites et freins

 Conservation des données

 Traitement des données

Métatranscriptomique

Métagénomique

Introduction

Principes de la métagénomique

Importance des métadonnées

Limites et freins

 Conservation des données

 Traitement des données

Métatranscriptomique

Données

Besoin d'une profondeur de séquençage importante en métagénomique

Augmentation exponentielle des données générées

2-10 To de données/laboratoire/an

Limites des transferts des données par les réseaux informatiques

Conservation des données

Quoi?

Données brutes

Résultats des analyses

Description des données

Où?

Localement

Cloud

Dépôts publics

Principaux dépôts publics

EBI : ENA, EBI metagenomics, ...

Priorité aux données ayant des métadonnées qui suivent les standards MixS/MIGS/MIMS

Outils web intuitifs, checklist, support de soumission, tutoriels

NCBI : GenBank, NCBI metagenomics, ...

Métadonnées suivant les standards GSC encouragés

DDJB

Pas nécessaire de suivre les standards GSC

Métagénomique

Introduction

Principes de la métagénomique

Importance des métadonnées

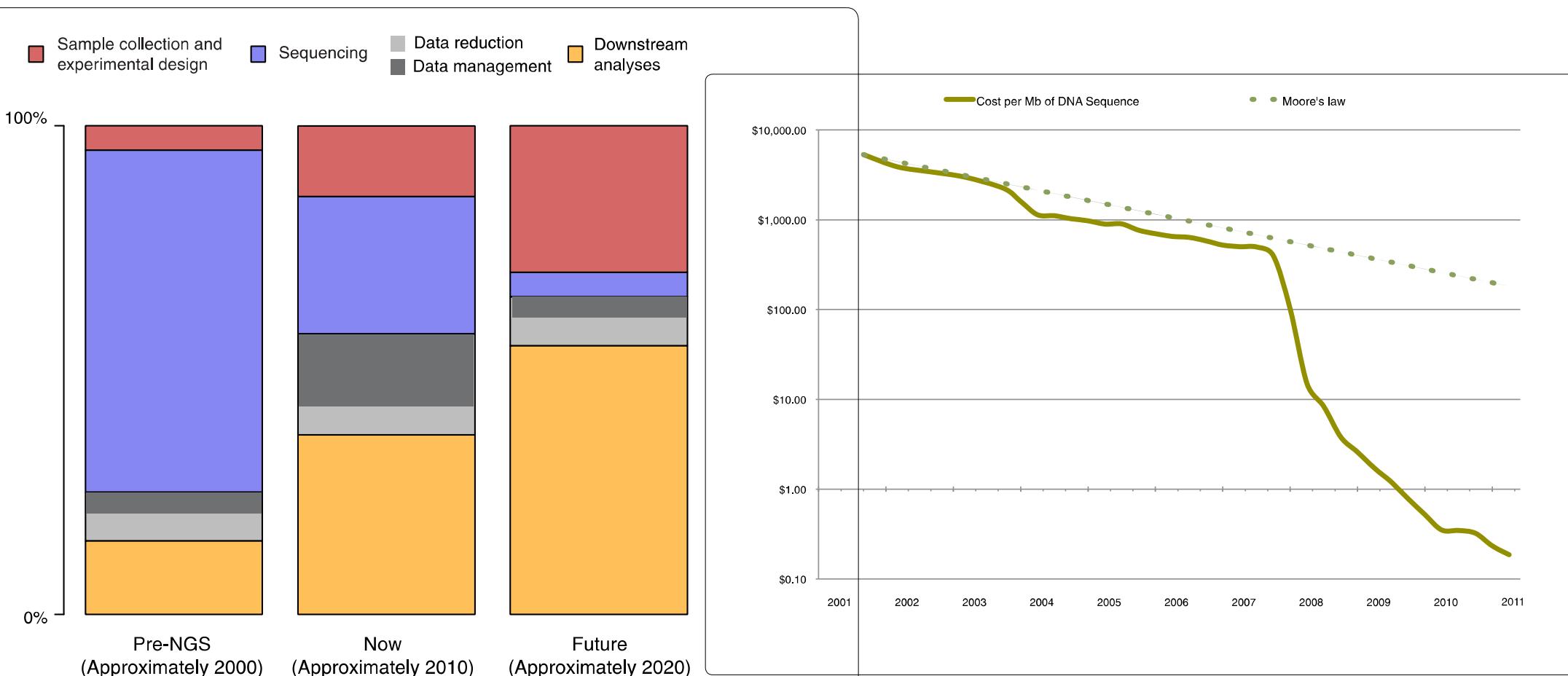
Limites et freins

Conservation des données

Traitements des données

Métatranscriptomique

Explosion des besoins en traitement des données et fin de la loi de Moore



Parallélisation des calculs

Une solutions pour traiter les données

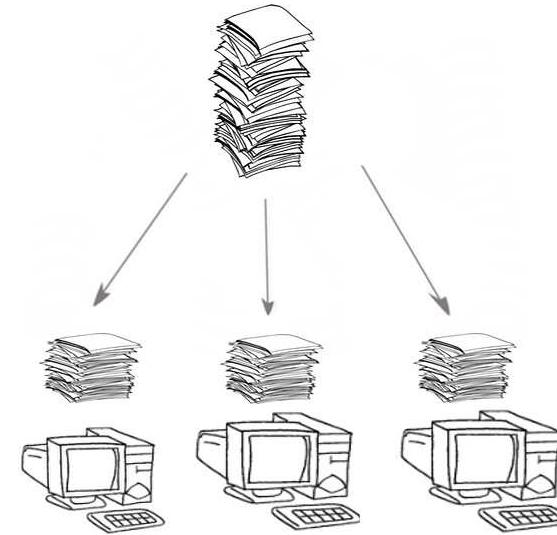
Cibles

Données

Modification des algorithmes



Divide and Conquer



Solutions

Clusters de calcul
Grilles de calculs
Cloud computing

A prendre en compte dans la planification d'expérience

Volume et stockage des données

Anticipation

Echelles de temps

Temps d'analyses >> temps de génération des données

Coûts en temps et argent

Métagénomique

Introduction

Principes de la métagénomique

Importance des métadonnées

Limites et freins

Métatranscriptomique

Métagénomique

Objectif

Etude les profils d'expression des gènes

Déterminer

quels microbes sont actifs

quels gènes sont transcrits

Méthodes

Séquençage des ARN

Limites

Courte demi-vie

Faible présence de l'ARNm

Principe

