

# Métagénomique

Février 2016

*Bérénice Batut*

✉ [berenice.batut@udamail.fr](mailto:berenice.batut@udamail.fr)

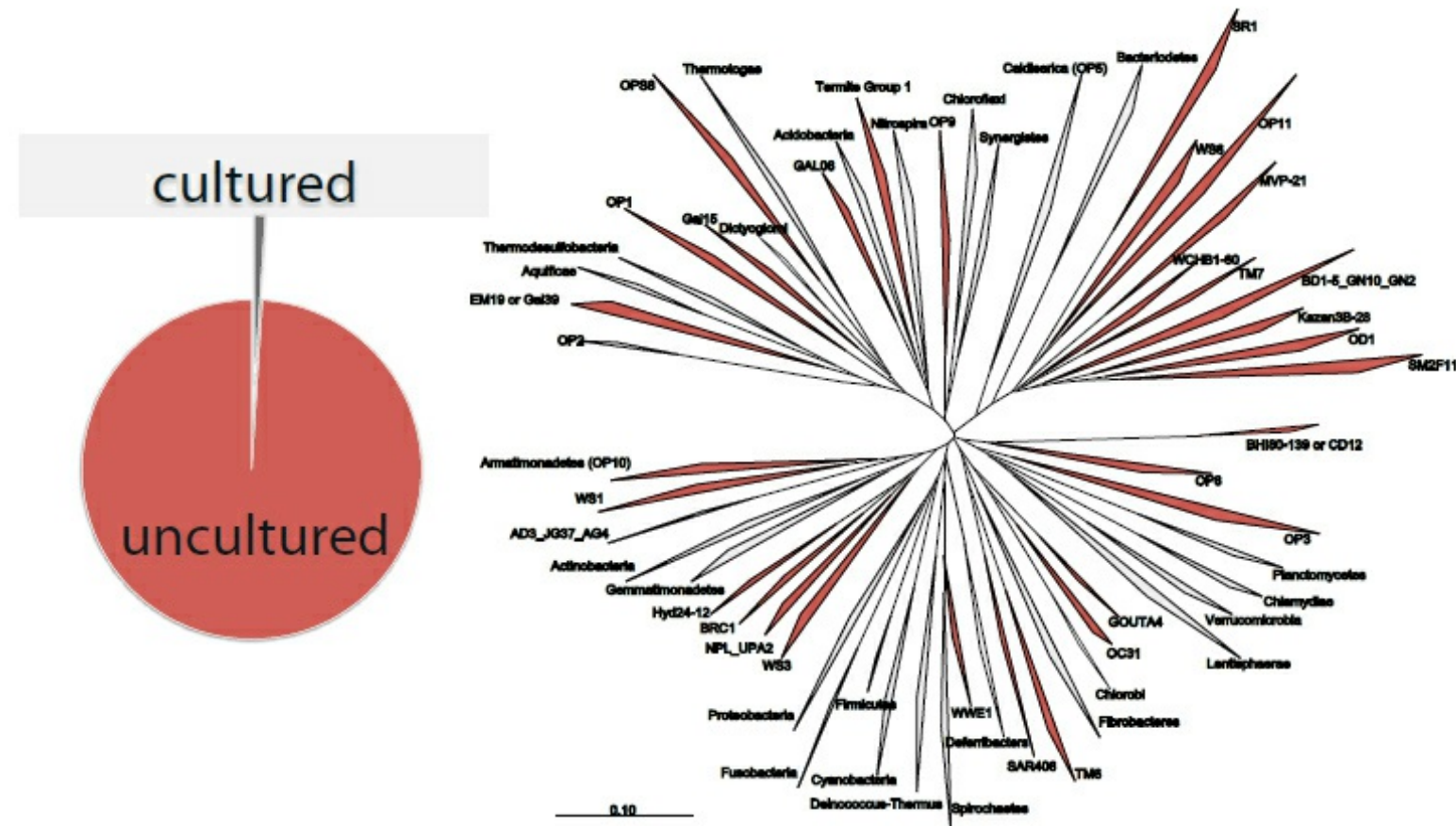
Pourquoi la  
métagénomique?

# Ecosystèmes microbiens

- Majeure partie de la biomasse sur Terre
  - 90% des cellules d'un être humain
- Colonisation de la plupart des niches écologiques

# Diversité microbienne

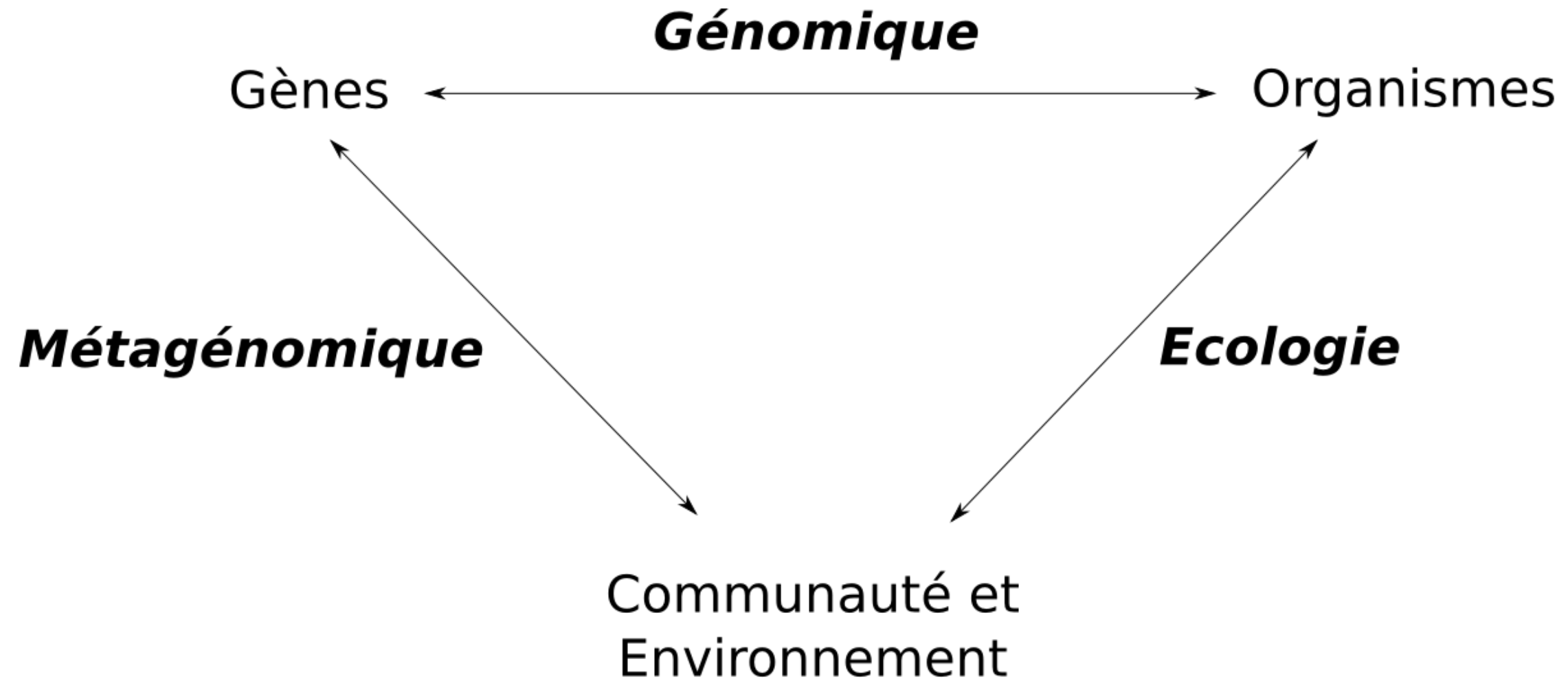
Our skewed view of the microbial world



16S rRNA tree of known bacterial phyla

Crédits : Christian Rinke / Tanja Woyke, DOE JGI

# 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

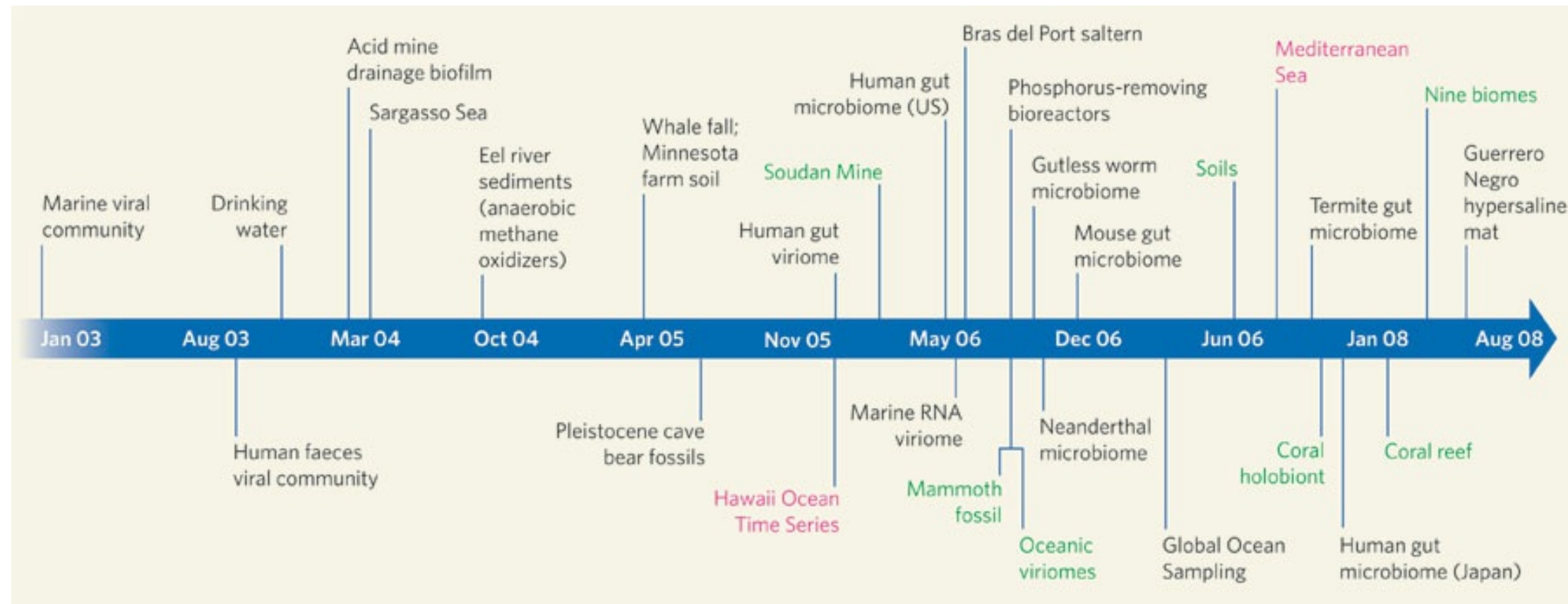
Compréhension du rôle écologique, du métabolisme et de  
l'histoire évolution d'un cosystème

# Métagénomique

- Pas de culture nécessaire
- Identification pour un écosystème
  - Micro-organismes (composition)
  - Gènes
  - Fonctions métaboliques



# Projets métagénomiques (jusqu'en 2008)

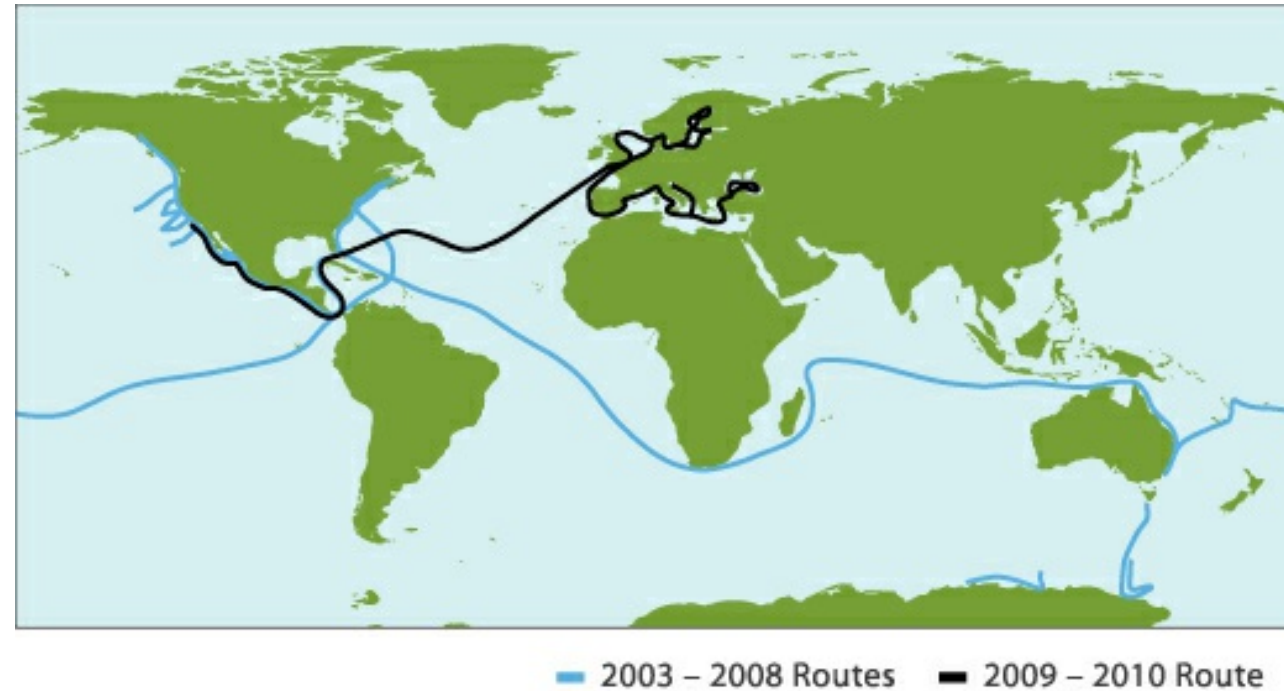


Crédits : *Hugenholtz & Tyson, Nature, 2008*

# Projets métagénomiques

- 2 165 [projets sur NCBI](#) (Janvier 2016)
  - Sable de plage
  - Moustique
  - Corail
  - Glace
  - Air de la ville de Singapour
  - Surface de la cuvette des toilettes
  - Fromages
  - ...

# Global Ocean Sampling (GOS)

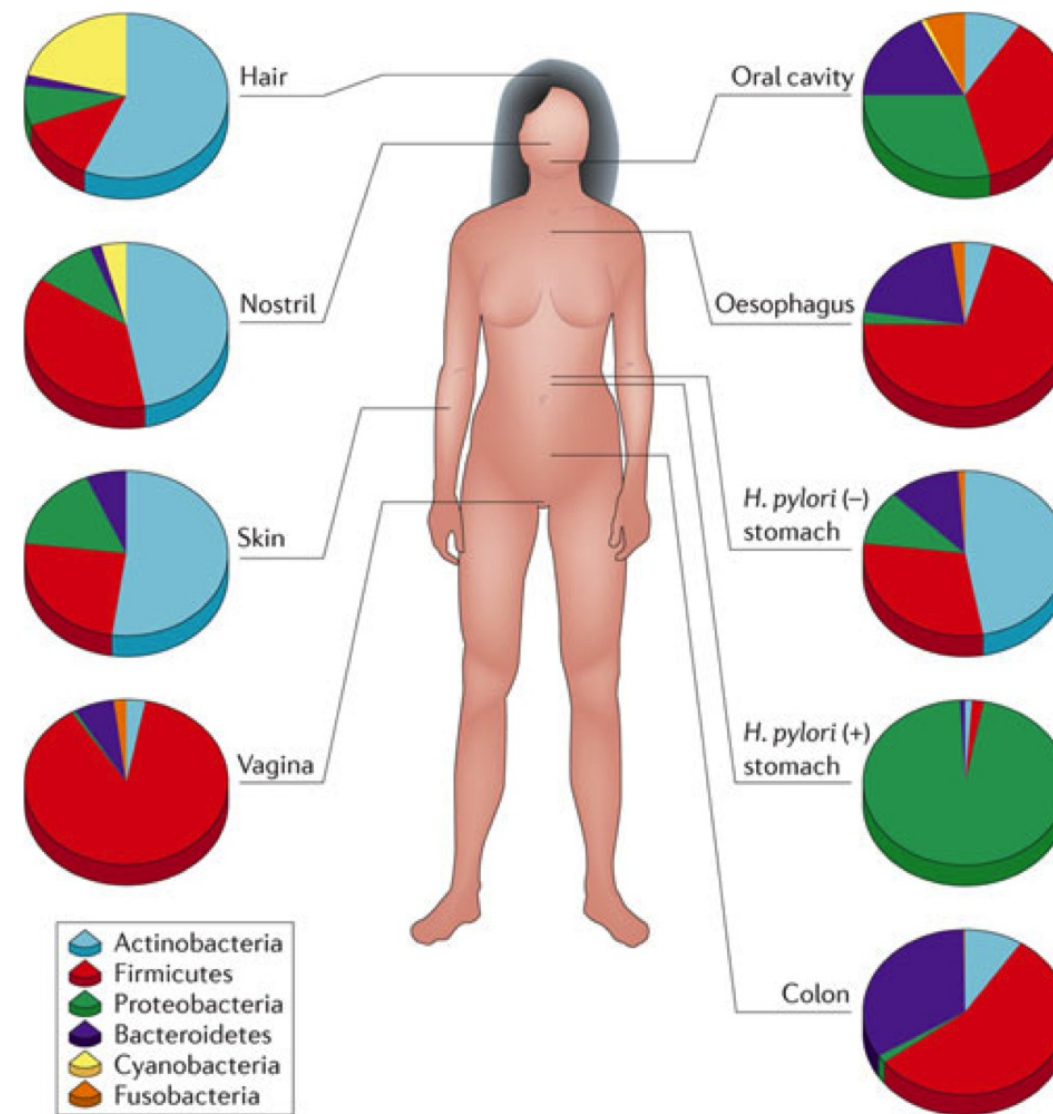


- Estimer la diversité génétique des communautés microbiennes marines
- Comprendre leur rôle dans les processus fondamentaux de la nature

# Global Ocean Sampling (GOS)

- Production du plus grand catalogue de gènes de milliers de nouvelles espèces
- Données générées
  - 3 087 469 séquences nucléotidiques
  - 6 123 395 séquences protéiques identifiées
- 8 publications et 1 édition spéciale de PLoS Biology

# Santé humaine



Crédits : *Cho & Blaser, Nature Rev Genet, 2012*

# Santé humaine

- Human Microbiome Project (HMP)
  - 2,3 Tb de données métagénomique
  - 35 milliard de reads
    - 690 échantillons
    - 300 individus
    - 15 sites du corps

# Santé humaine



- Metagenomics of the Human Intestinal Tract (MetaHIT)
  - 540 Gb de séquences ADN
  - 134 individus

# Santé humaine

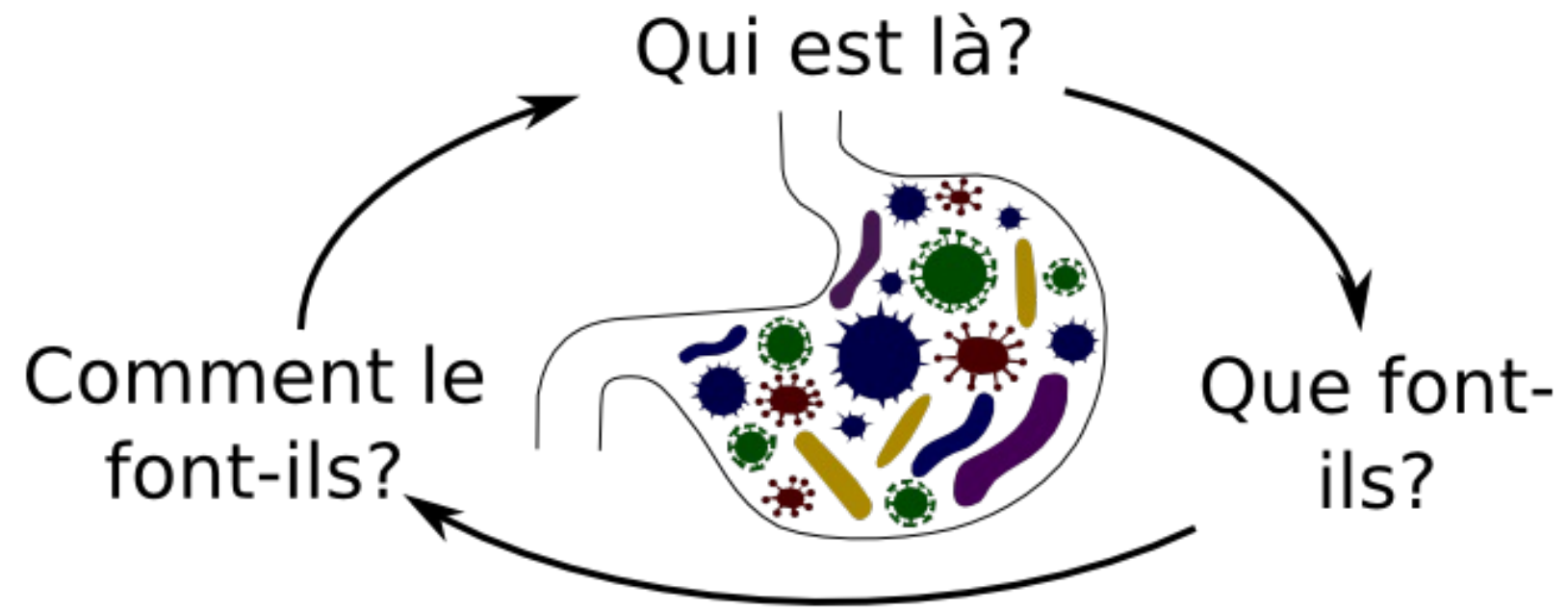


- **American Gut**
  - Projet participatif
  - 3,238 participants
  - 101 million de séquences ADN
  - 27 Gb



# Principes de la métagénomique

# Métagénomique

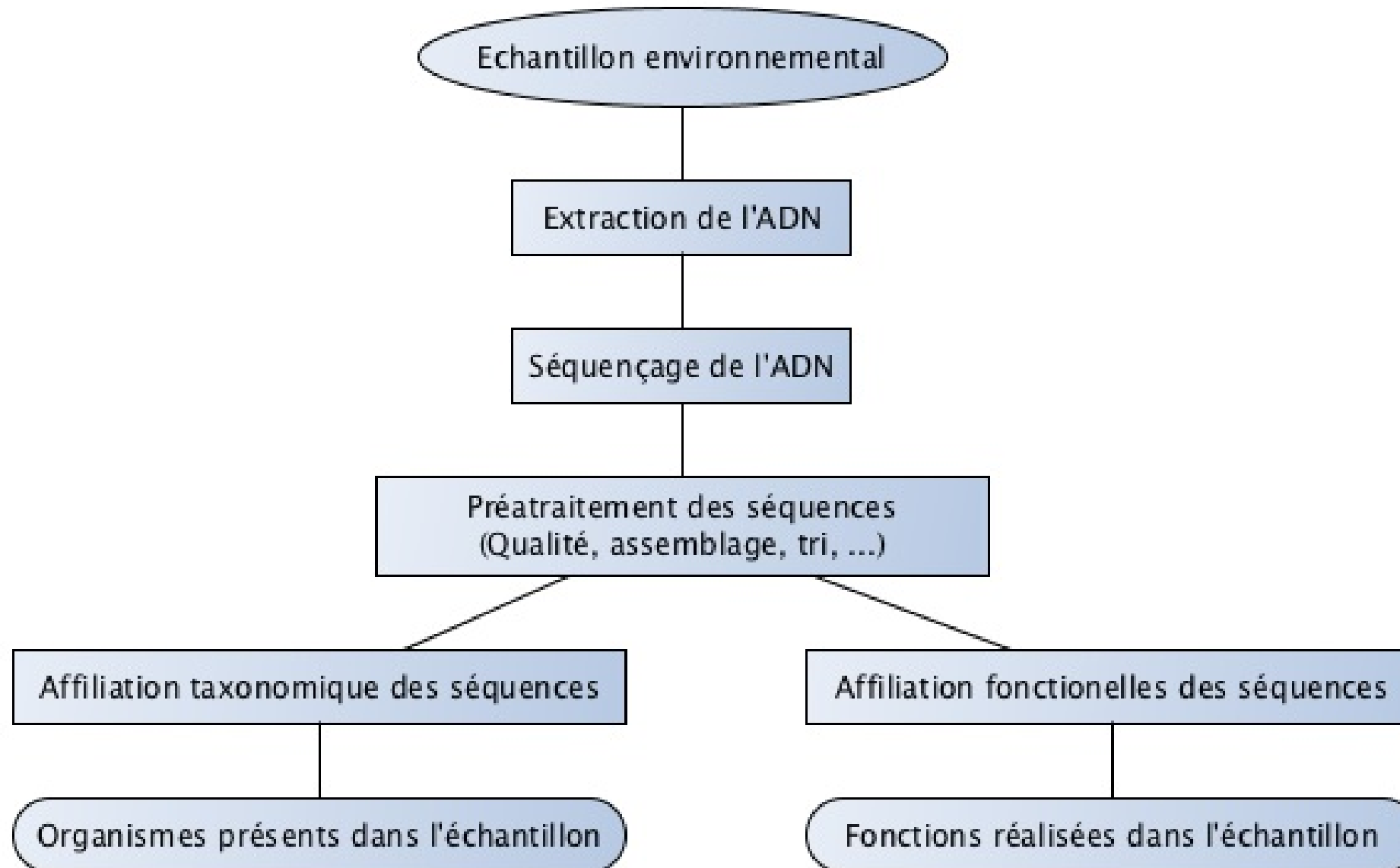


# Métagénomique

## Objectifs:

- Identifier les organismes présents dans un échantillon
- Identifier les fonctions principales réalisés par les organismes d'un échantillon

# Métagénomique



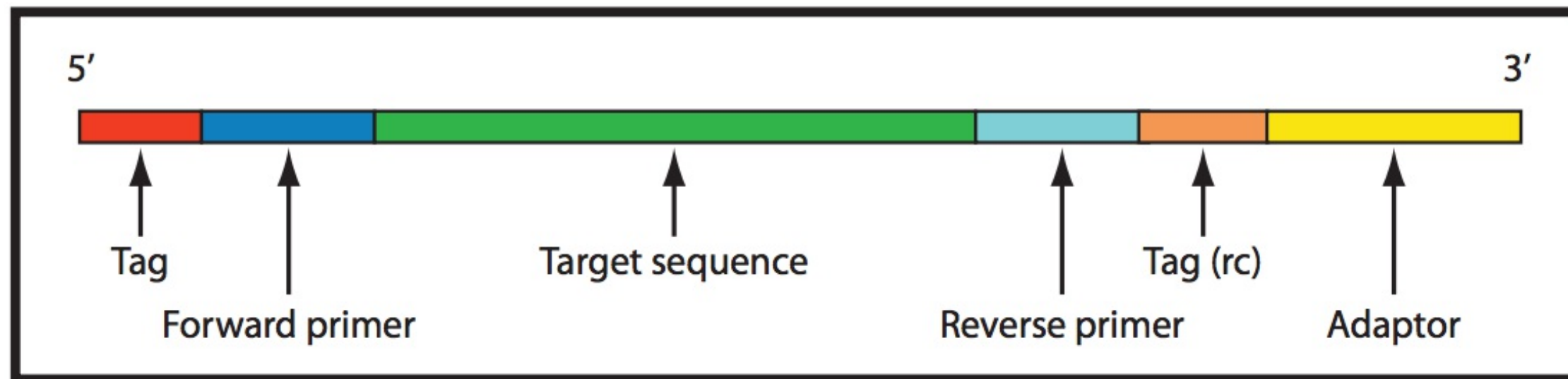
# Extraction et séquençage

## 2 types d'expériences métagénomiques

- Amplicon
- "Environmental shotgun metagenomics", WGS ou métagénomique

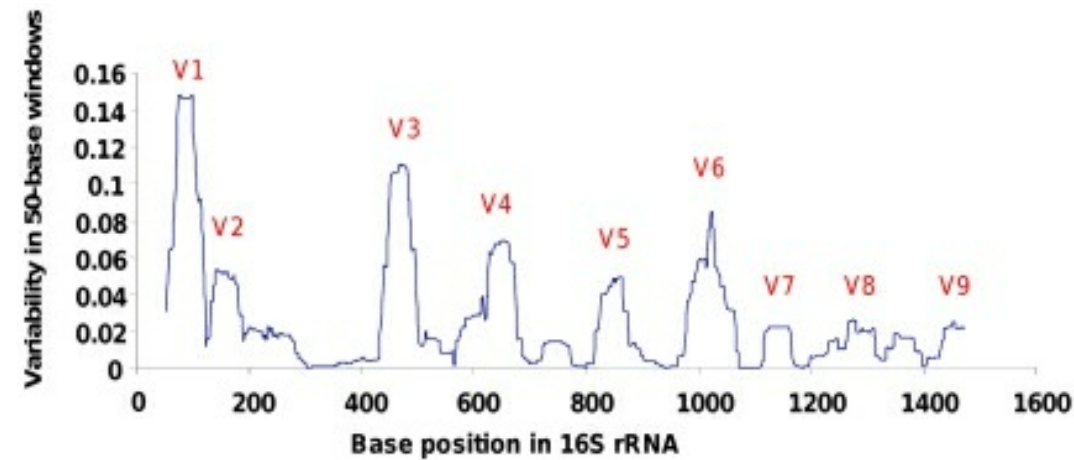
# Amplicon

Morceau d'ADN ou ARN amplifié par PCR



Crédits : *Kumar et al, BMC Bioinformatics, 2011*

# Utilisation de l'ARNr 16S



Crédits : *Bodilis et al, PLoS One, 2012*

## Séquençage des régions

- V1.3
- V3.4
- V4



# Possibilité de cibler d'autres gènes

  
**functional gene pipeline & repository**

[ Home | Display Options | Help | FunGenePipeline | RDP Home ]

Begin with these gene links: Version 8.3 -- GenBank 211 (as of 1/5/2015)  
Process your own Functional Gene data using our new [FunGene Pipeline](#)

If you use RDP's FunGene, [please cite our most recent article](#).

**Antibiotic resistances**  
  
gene—contributor  
[ACT](#)—Syed Hashsham  
[BEL](#)—Syed Hashsham  
[beta\\_IS6](#)—Robert Stedtfield  
[beta\\_tnpA](#)—Robert Stedtfield  
[beta\\_tnpA2](#)—Robert Stedtfield  
[bet\\_blaSHV](#)—Robert Stedtfield  
[bet\\_tnpA](#)—Robert Stedtfield  
[CARB](#)—Syed Hashsham  
[cefa\\_qacEdelta](#)—Robert Stedtfield  
[chl\\_cmlA](#)—Robert Stedtfield  
[CMY](#)—Syed Hashsham  
[cprA](#)—Tamara Tsoi Cole  
[cprB](#)—Tamara Tsoi Cole  
[CTX-M](#)—Syed Hashsham  
[dfra1](#)—Syed Hashsham  
[dfra12](#)—Syed Hashsham  
[FOX](#)—Syed Hashsham  
[gapA](#)—Tim Johnson  
[GES](#)—Syed Hashsham  
[IMI](#)—Syed Hashsham  
[IMP](#)—Syed Hashsham  
[IncW\\_trwA](#)—Tim Johnson  
[IncW\\_trwB](#)—Tim Johnson  
[IND](#)—Syed Hashsham  
[intl](#)—Carlos Rodriguez-Minguela  
[intl1\\_sub1](#)—Tim Johnson  
[intl2](#)—Tim Johnson  
[intl3](#)—Tim Johnson  
[KPC](#)—Syed Hashsham  
[mdh\\_sub1](#)—Tim Johnson  
[mdh\\_sub2](#)—Tim Johnson  
[MIR](#)—Syed Hashsham  
[MOX](#)—Syed Hashsham  
[NDM](#)—Syed Hashsham  
[OXA](#)—Syed Hashsham  
[pec\\_aad2](#)—Robert Stedtfield  
[PER](#)—Syed Hashsham  
[repA](#)—Tim Johnson

**Plant Pathogenicity**  
  
gene—contributor  
[avrE](#)—James Kremer  
[txtA](#)—RDP  
[txtB](#)—RDP  
  
**Biogeochemical cycles**  
  
gene—contributor  
[amoA\\_AOA](#)—Felfei Liu  
[amoA\\_AOB](#)—RDP  
[buk](#)—RDP  
[but](#)—RDP  
[cbh1](#)—Cheryl Kuske  
[chb](#)—Fan Yang  
[cooS](#)—Fan Yang  
[cydA](#)—Rachel Morris  
[dsrA](#)—Alexander Loy/Michael Wagner  
[dsrB](#)—Alexander Loy/Michael Wagner  
[exc1](#)—Fan Yang  
[fixN](#)—Rachel Morris  
[glx](#)—Qichao Tu  
[hydA](#)—Fan Yang  
[lcc\\_ascomycetes](#)—Chris Wright  
[lcc\\_basidiomycetes](#)—Chris Wright  
[ligE](#)—Ryan Penton  
[lip](#)—Qichao Tu  
[mcrA](#)—Blaz Stres  
[mmoX](#)—Qichao Tu  
[mnp](#)—Qichao Tu  
[nag3](#)—Fan Yang  
[napA](#)—Laurent Philippot  
[narG](#)—Laurent Philippot  
[nifD](#)—RDP  
[nifH](#)—RDP  
[nirA](#)—RDP  
[nirB](#)—RDP  
[nirK](#)—Tracy Teal  
[nirS](#)—Veronica Gruntzig  
[nirX](#)—Eunha Beekun

**Phylogenetic markers**  
  
gene—contributor  
[EF-Tu](#)—James Kremer  
[fusA](#)—Scott Santos/Howard Ochman  
[gyrB](#)—Zarraz May-Ping Lee  
[ileS](#)—Scott Santos/Howard Ochman  
[lepA](#)—Scott Santos/Howard Ochman  
[leuS](#)—Scott Santos/Howard Ochman  
[pyrG](#)—Scott Santos/Howard Ochman  
[recA](#)—Scott Santos/Howard Ochman  
[recG](#)—Scott Santos/Howard Ochman  
[rplB](#)—Scott Santos/Howard Ochman  
[rpoB](#)—Scott Santos/Howard Ochman  
  
**Biodegradation**  
  
gene—contributor  
[alkB](#)—Gerben Zylstra/Elyse Rodgers-Vieira  
[benA](#)—Stephan Gantner  
[bph](#)—Gerben Zylstra  
[bphA1](#)—Stephan Gantner  
[bphA2](#)—Stephan Gantner  
[BSH](#)—Robert Stedtfield  
[carA](#)—Shoko Iwai  
[cntA](#)—Robert Stedtfield  
[cutC](#)—Robert Stedtfield  
[dbfA1](#)—Shoko Iwai  
[dxnA](#)—Shoko Iwai  
[dxnA-dbfa1](#)—Tim Johnson  
[HSDH](#)—Robert Stedtfield  
[npah](#)—Gerben Zylstra  
[p450](#)—Gerben Zylstra/Elyse Rodgers-Vieira  
[ppah](#)—Gerben Zylstra  
[PSA](#)—Robert Stedtfield  
  
**Metal Cycling**  
  
gene—contributor  
[arsA](#)—PFAM

”Environmental shotgun metagenomics”,  
WGS ou métagénomique

Séquençage de l'ensemble des séquences d'un échantillon

# Comparaison

Amplicon	Métagénomique
Un fragment de gène	Toutes les séquences
Technique simple	Technique plus complexe
Biais de PCR	-
Information phylogénétique limitée	Information phylogénétique "complète"
Pas d'information fonctionnelle	Information fonctionnelle complète
Bases de données complètes	Bases de données incomplètes

# Prétraitements

# Prétraitements

- Contrôle de la qualité des séquences
- Assemblage
- Tri des séquences
- Prédiction des séquences d'intérêt

# Assemblage



Crédit : Drew Sheneman

*Metagenomics is like a disaster in jigsaw shop*

# Assemblage

- Objectif
  - Construire des séquences plus longues, voir des organismes entiers
- Problème
  - Séquences courtes
  - Séquences issues d'organismes différents

# Assemblage

2 types d'outils d'assemblage

- A l'aide de bases de données
- *De novo*

⚠ Chimères : assemblage de séquences issues d'espèces différentes

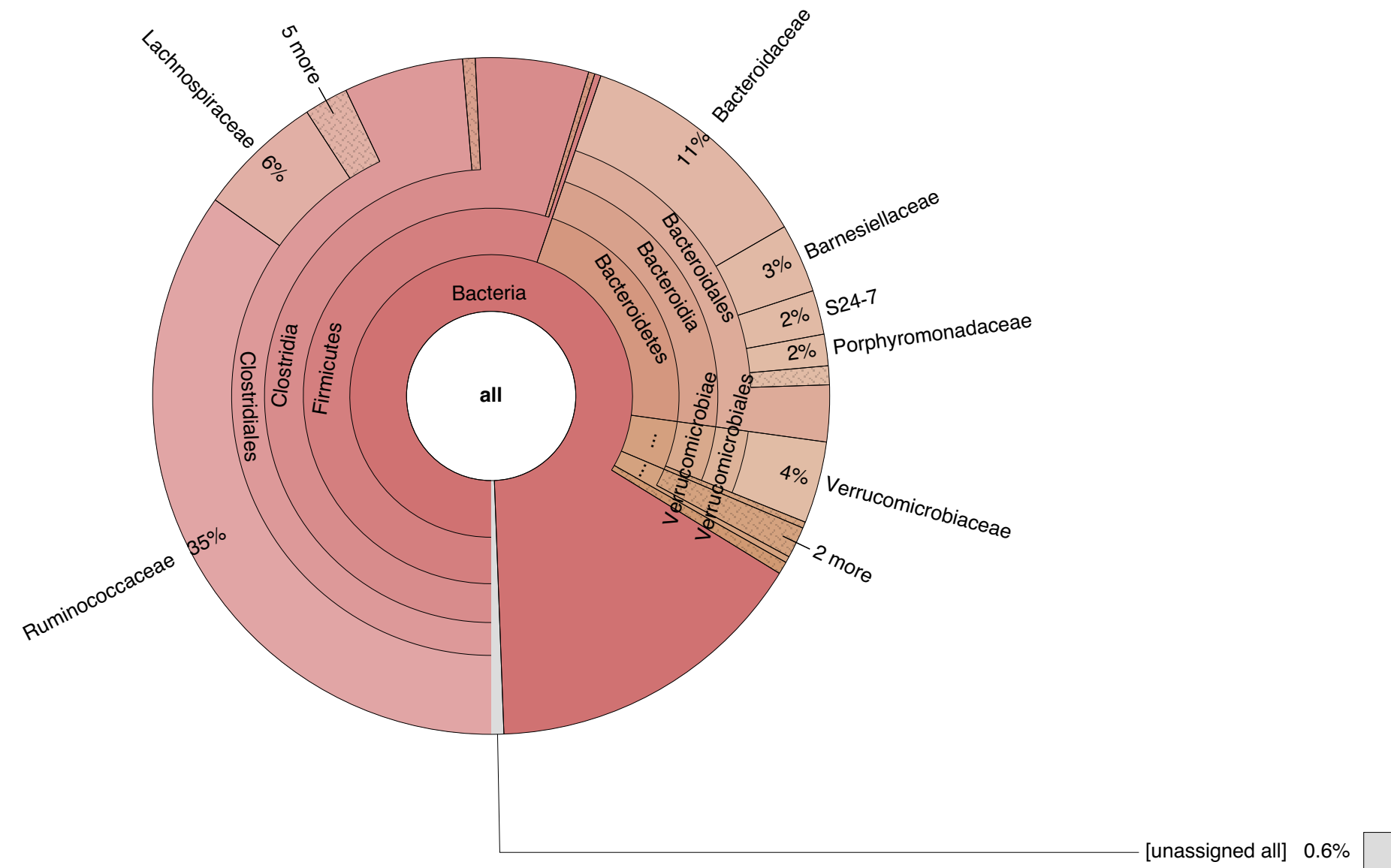


# Tri des séquences

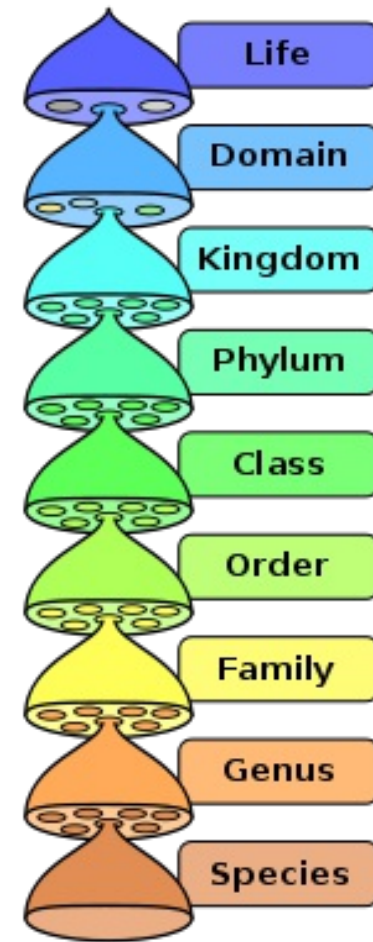
- Utilité
  - Assignment des séquences
  - Recherche dans des bases de données spécifiques
- Caractéristiques d'intérêt
  - Gènes ARNr, ARNt
  - ORF
  - Unités taxonomiques opérationnelles (OTU)
- Quelques outils
  - SortMeRNA, rRNA selector
  - MetaGeneAnnotator, Prodigal, Orphelia
  - tRNAscanSE

# Affiliation taxonomique des séquences

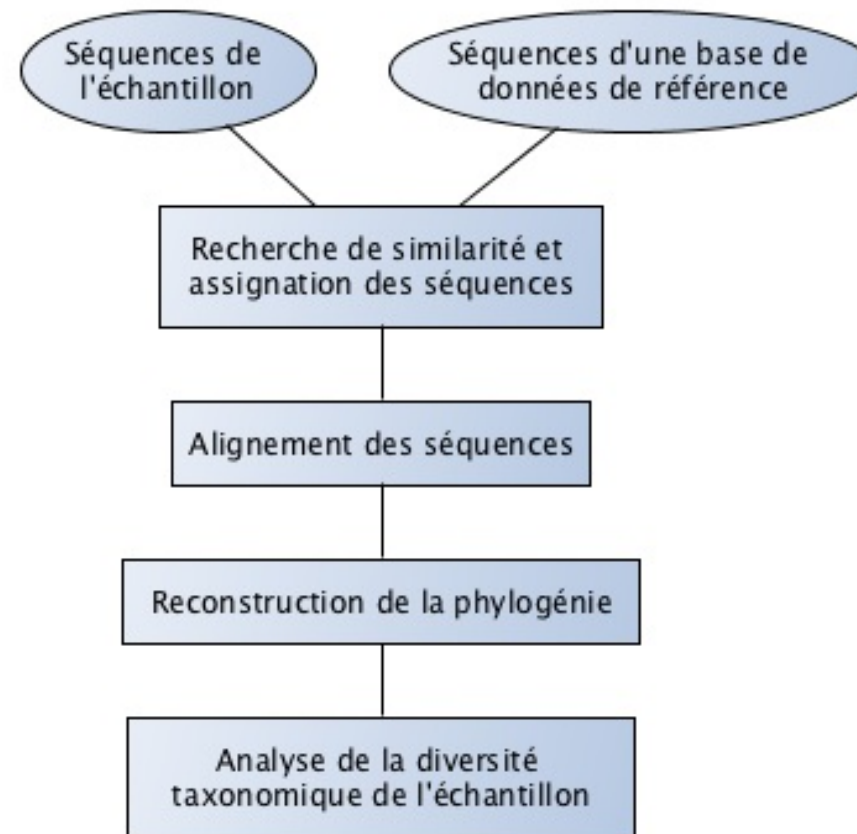
# Pourquoi l'affiliation taxonomique?



# Pourquoi l'affiliation taxonomique?



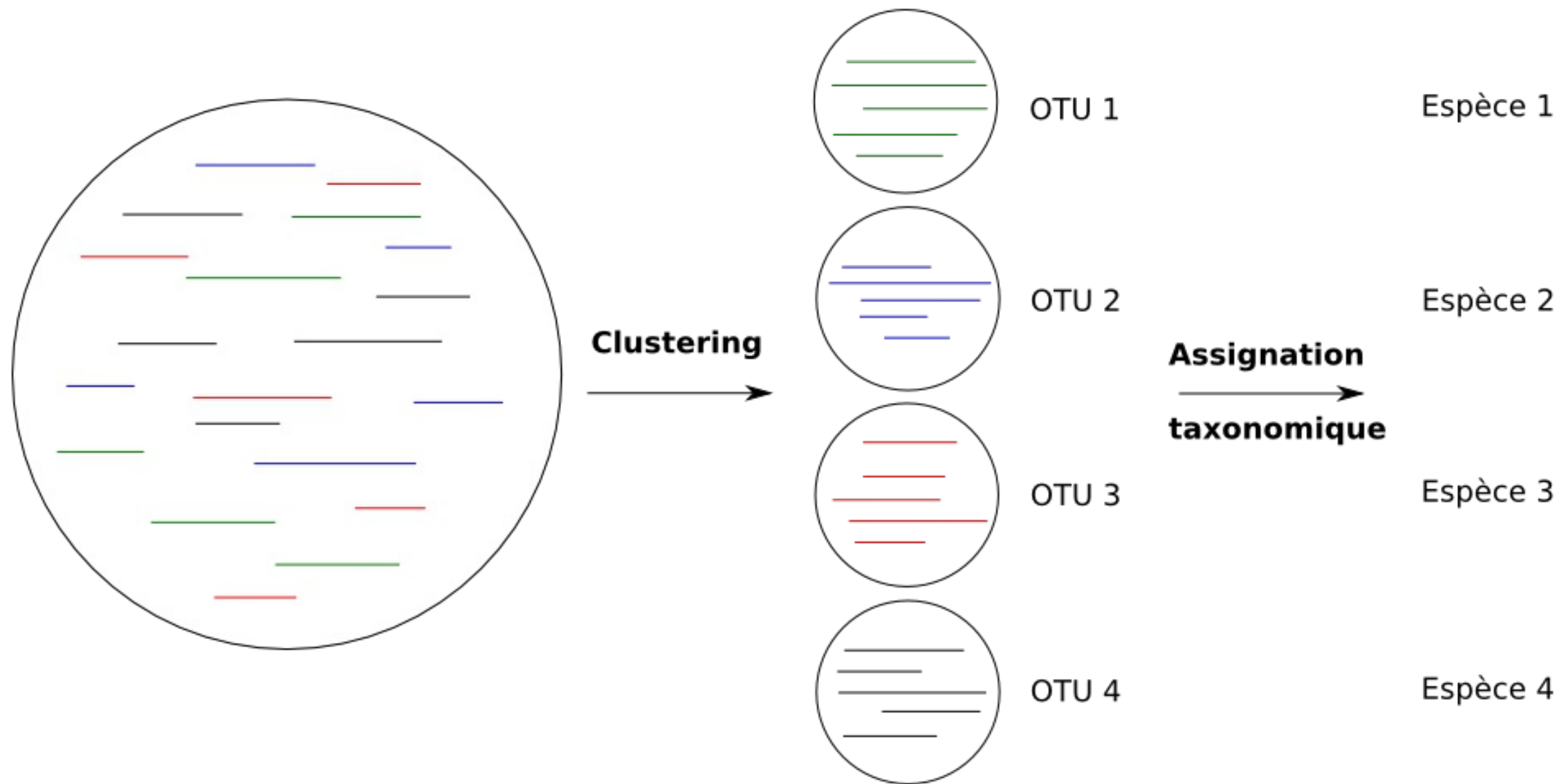
# Procédure standard



# Limites

- Longueur des séquences
- Nombre de séquences
- Absences d'homologues dans les banques
- Faible abondance des marqueurs phylogénétique classiques
  - Boues d'épuration : 0.17%
  - Mer des Sargasses : 0.06%
  - Sol acide, mines du Minnesota : 0.017%

# Clustering des séquences en OTU (Operational taxonomic unit)



# Recherche de similarité

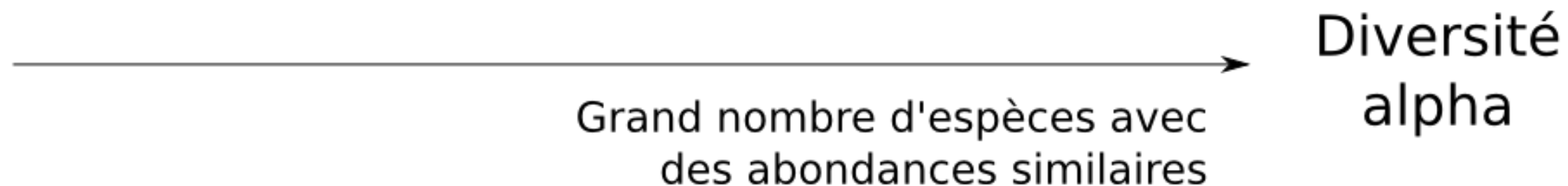
- Recherche de similarité par fréquence de k-mer
- Recherche de similarité intrinsèques
  - Contenu en GC
  - Usage des codons



# Analyse de la diversité d'un échantillon

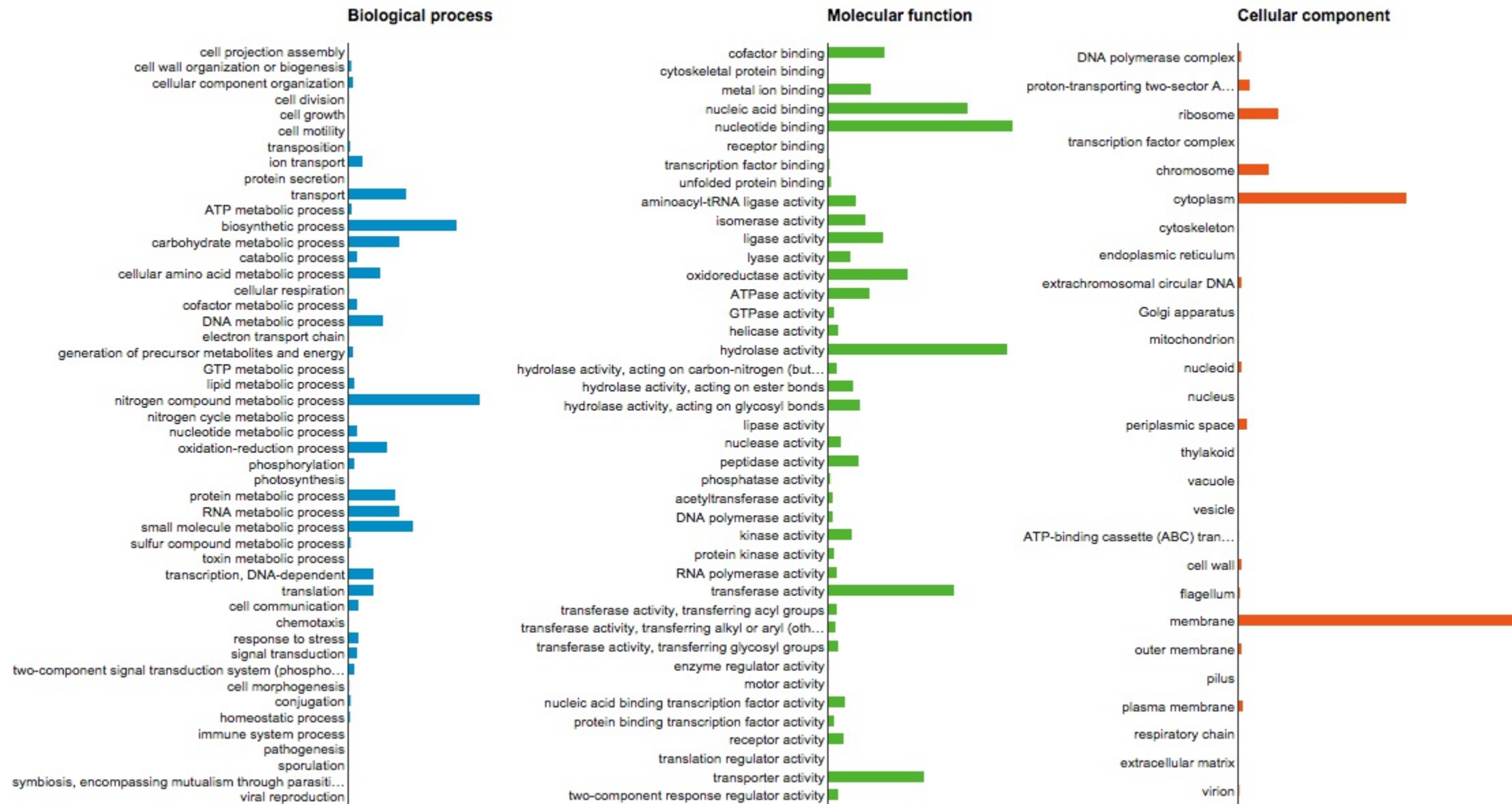
## Alpha diversité

- 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)
  - Égalité
    - Proportion d'espèces présentes

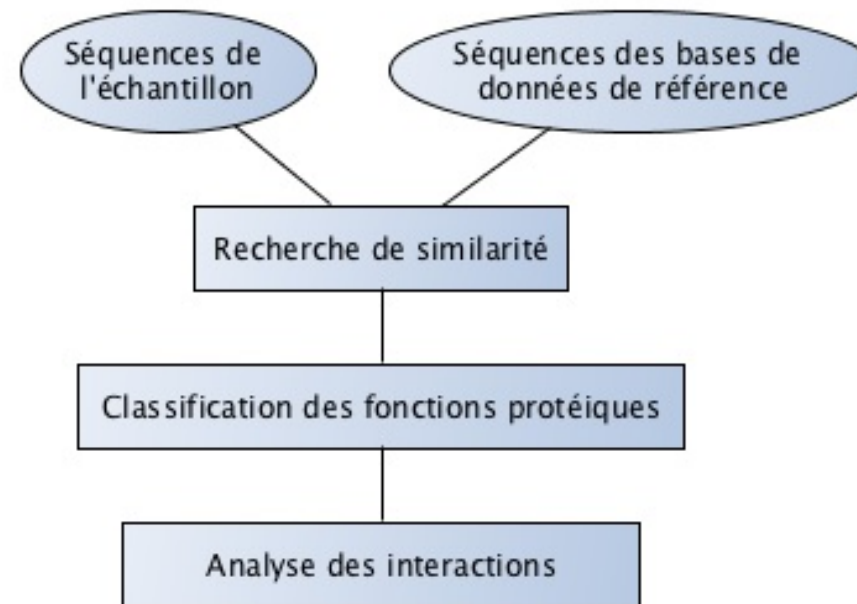


# Affiliation fonctionnelle

# Pourquoi l'affiliation fonctionnelle ?



# Procédure



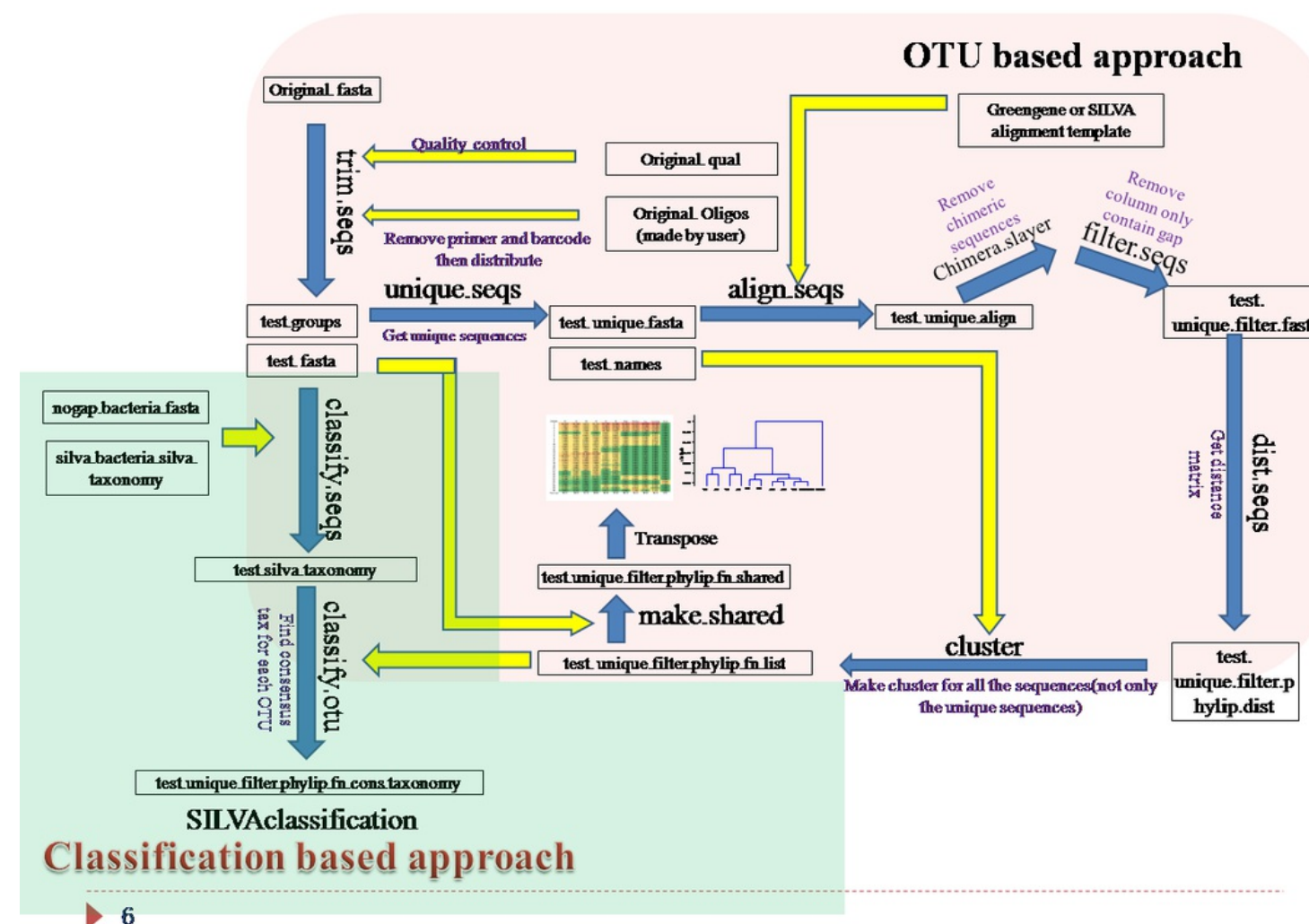
# Classification des fonctions protéiques

- Recherche contre bases de données de références
  - Cluster of orthologous groups of proteins (COG)
  - SEED
  - Kyoto Encyclopedia of Genes and Genomes (KEGG)
  - ...
- Recherche de motifs fonctionnels
  - InterPro
  - Prosite
  - ...

# Quelques pipelines

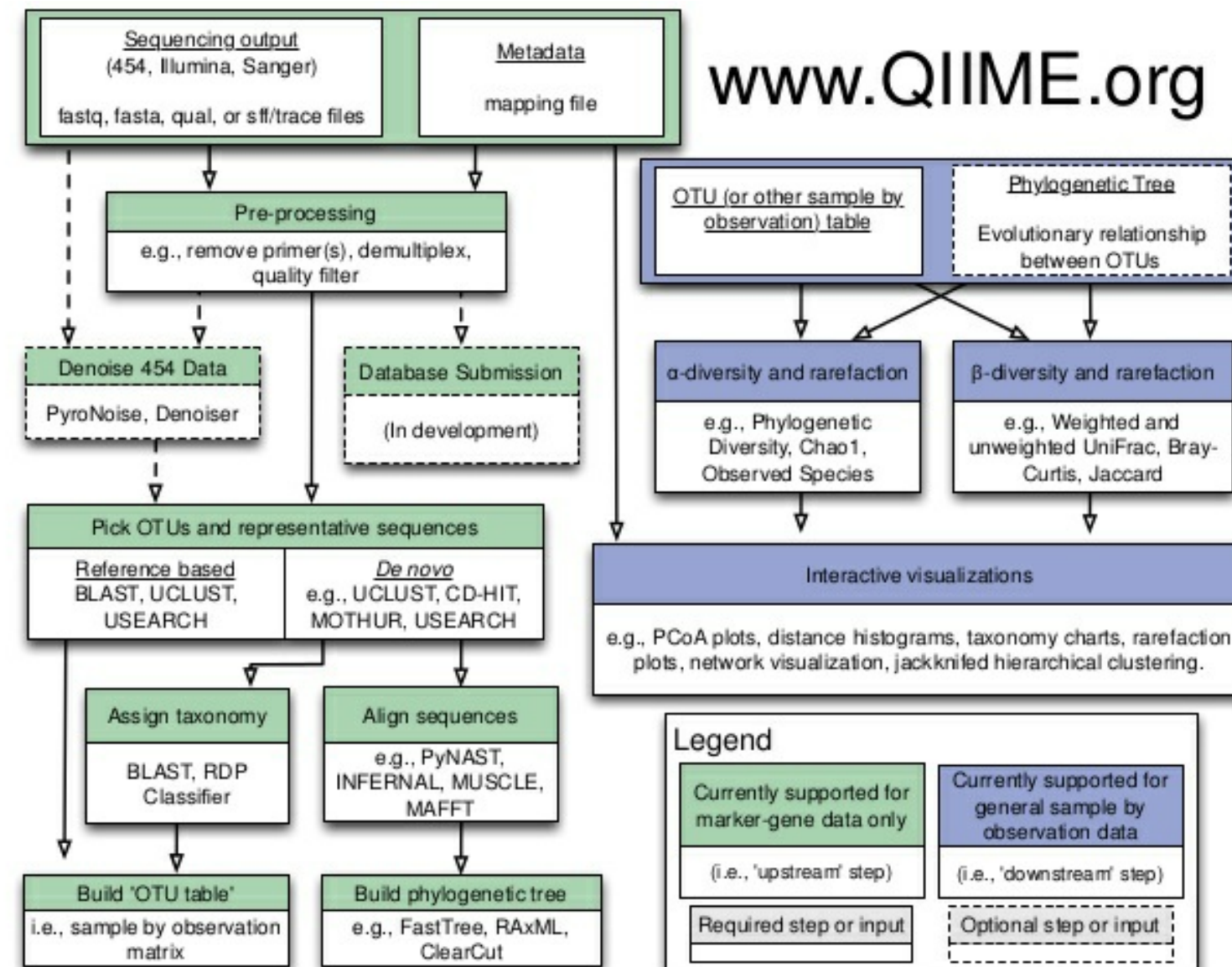
# Mothur

Schloss et al, Applied Env Microbiol, 2009



# QIIME

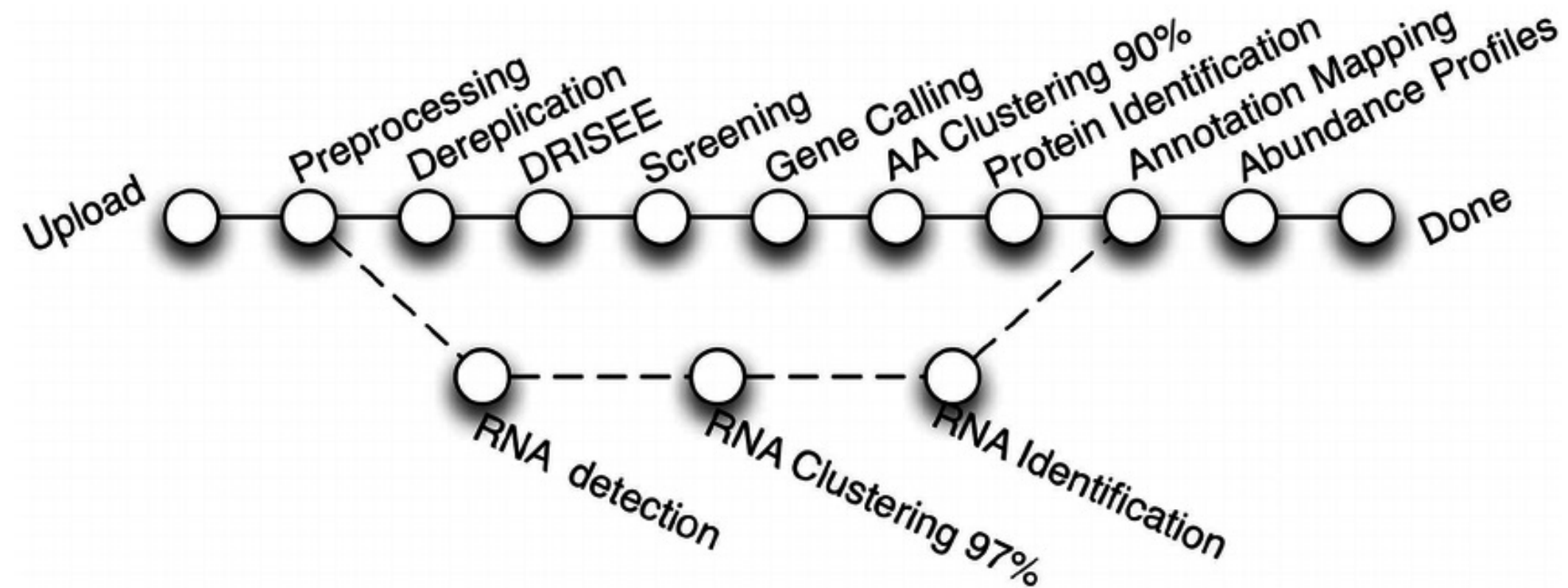
Caporaso et al, Nature Methods, 2010





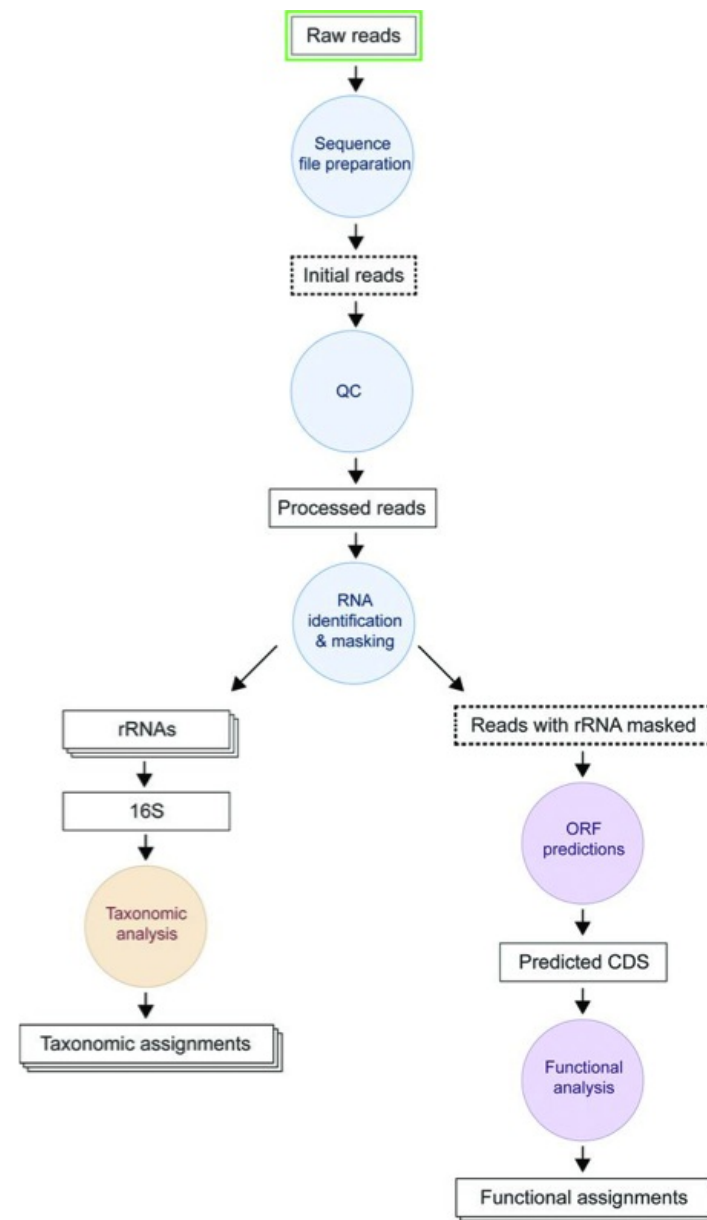
# MG-Rast

*Meyer et al, BMC Bioinformatics, 2008*



# EBI metagenomics

*Mitchell et al, Nucleic Acids Res, 2015*



# Exemple

Lean human gut

# Importance des métadonnées

# Que sont les métadonnées?

Description en profondeur et contrôlée de l'échantillon

- Où?
  - Latitude et longitude
  - Profondeur
  - ...
- Quand?
- Quoi?
- Comment?
  - Processus de conservation
  - Méthode d'extraction
  - Plateforme de séquençage
  - ...

# Importance des métadonnées

Indispensable pour partager et intégrer les données dans la communauté

Besoin d'un vocabulaire standardisé minimal

# Genome Standard Consortium (GSC)

Organisme international, ouvert, formé en 2005

## Objectifs

- Promouvoir les mécanismes de standardisation de la description des génomes
- Promouvoir l'échange et l'intégration des données génomiques

# Genome Standard Consortium (GSC)

## Missions

- Développement
  - de nouveaux standards (méta)génomiques
  - des méthodes de capture et d'échange des métadonnées
- Harmonisation de la collection des métadonnées 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

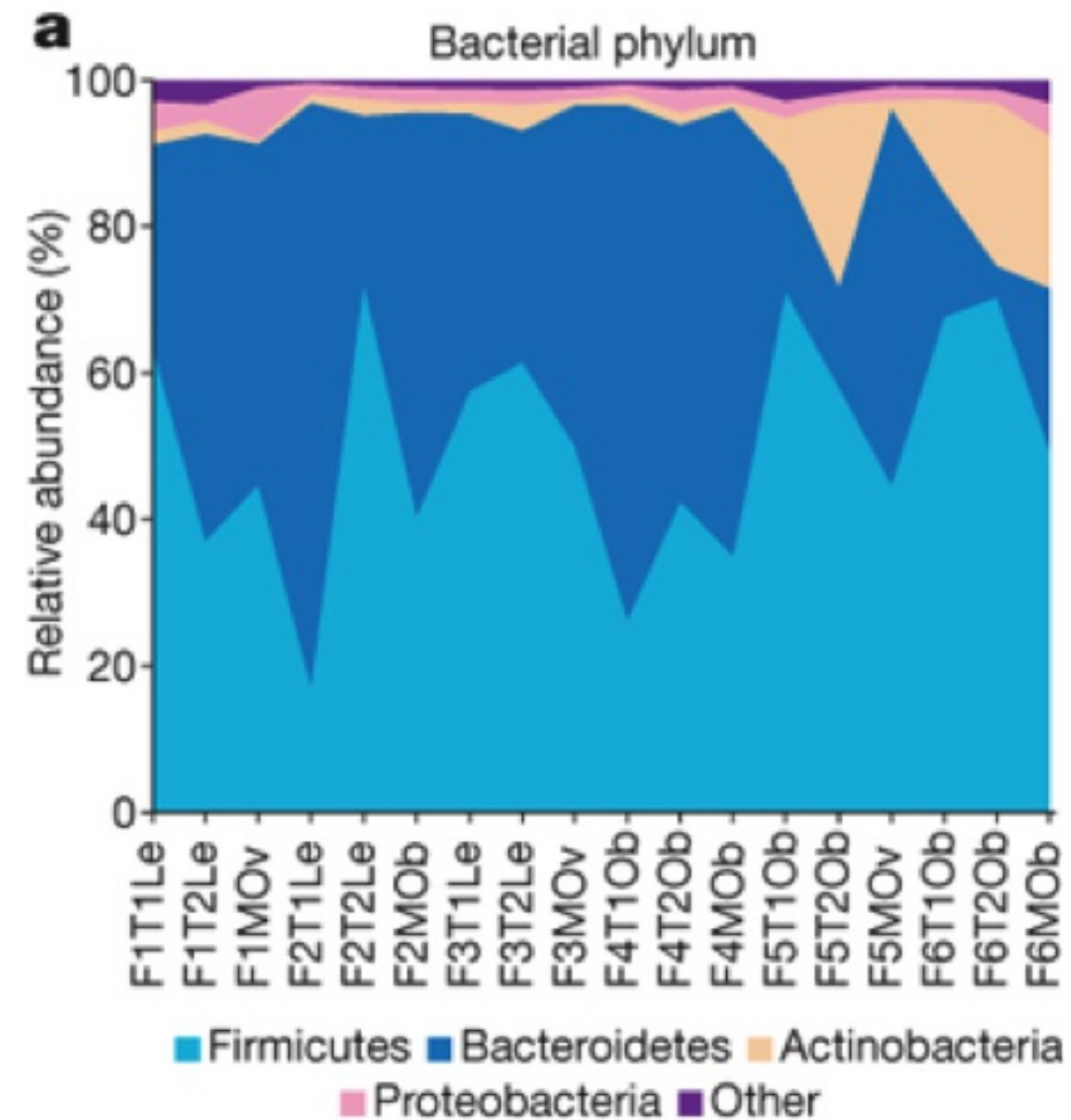
**MIMARKS** : Minimum Information about a MARKer gene  
Sequence

# Métagénomique comparative

*Turnbaugh et al, Nature, 2009*

# Comparaison de la diversité

# Comparaison des organismes présents



Crédit : *Turnbaugh et al, Nature, 2009*

# Diversité beta

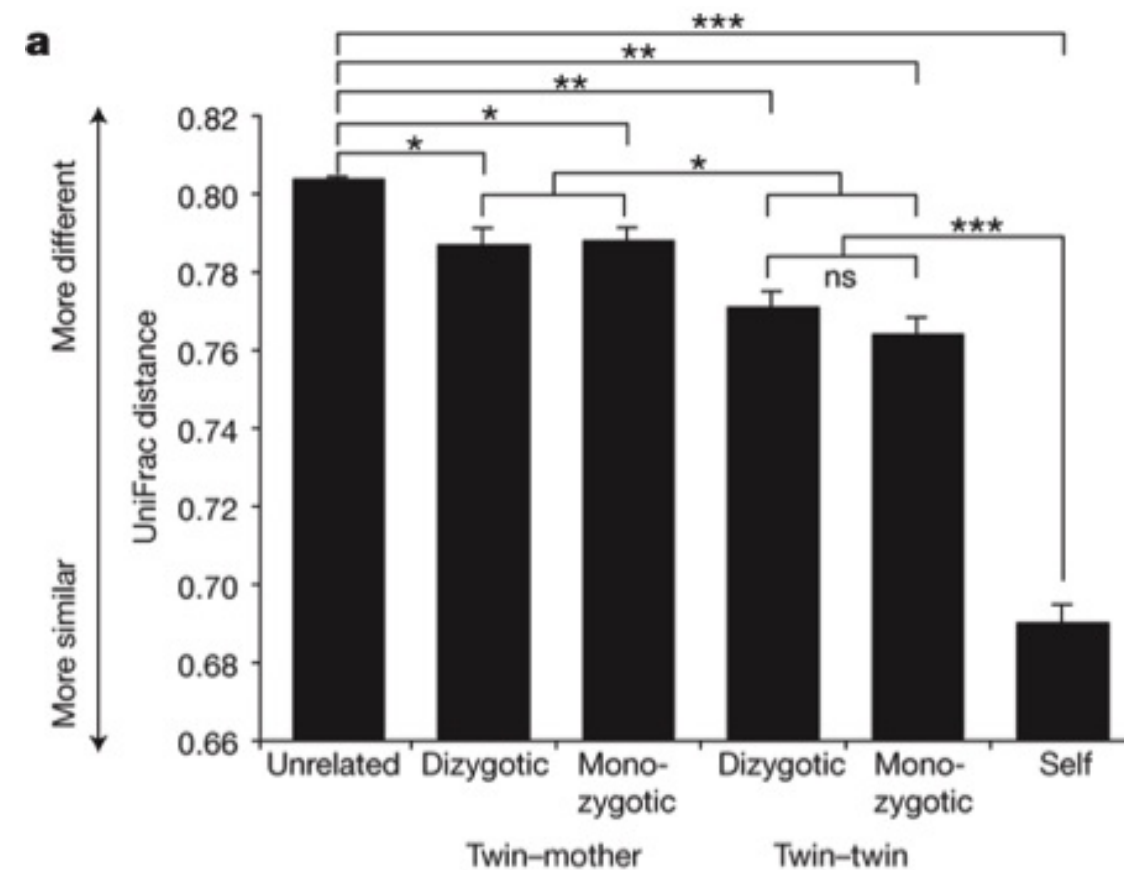
- Notion définie par *Whittaker, Evolution and Measurement of Species Diversity, 1972*
- Comparasion entre 2 échantillons de la différence de communauté (difference des espèces ou OTU)

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

S : nombre total d'espèces enregistrées dans les 2 échantillons

$\alpha$  : moyenne du nombre d'espèces trouvées au sein des 2 échantillons

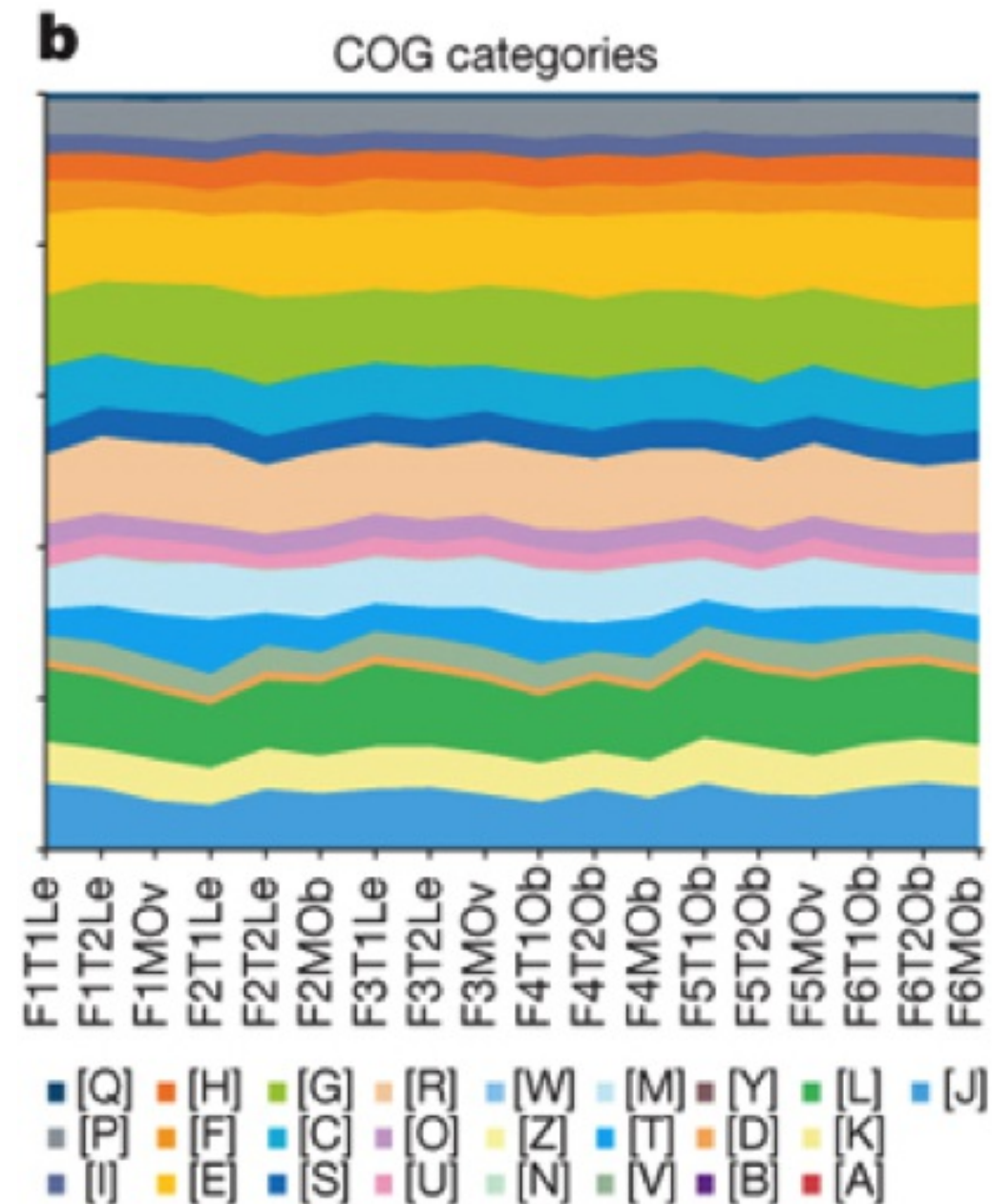
# Comparaison de la diversité avec la distance Unifrac



Crédit : [Turnbaugh et al, Nature, 2009](#)

# Comparaison des fonctions réalisées

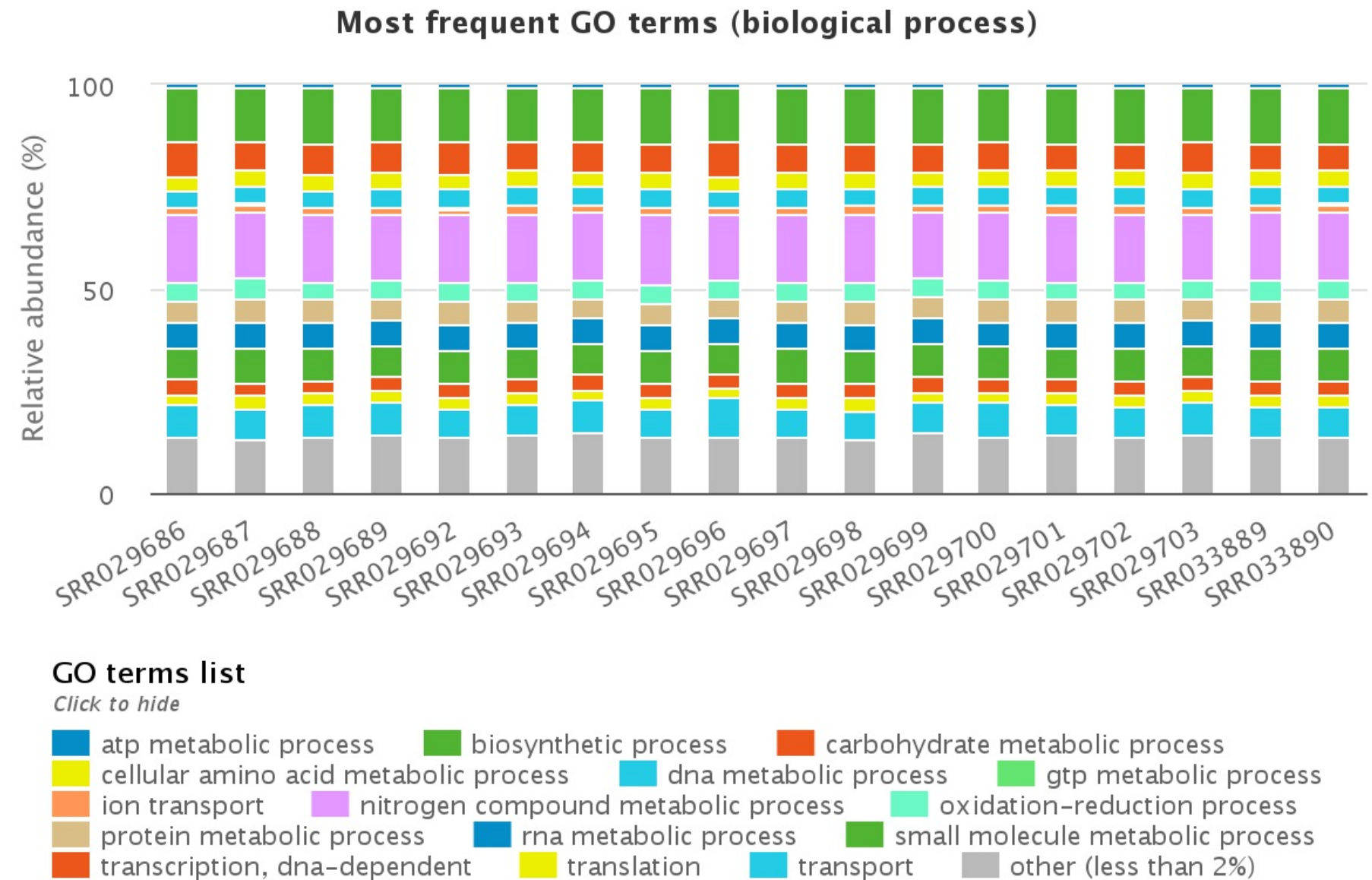
# Comparaison des catégories COG



Crédit : *Turnbaugh et al, Nature, 2009*

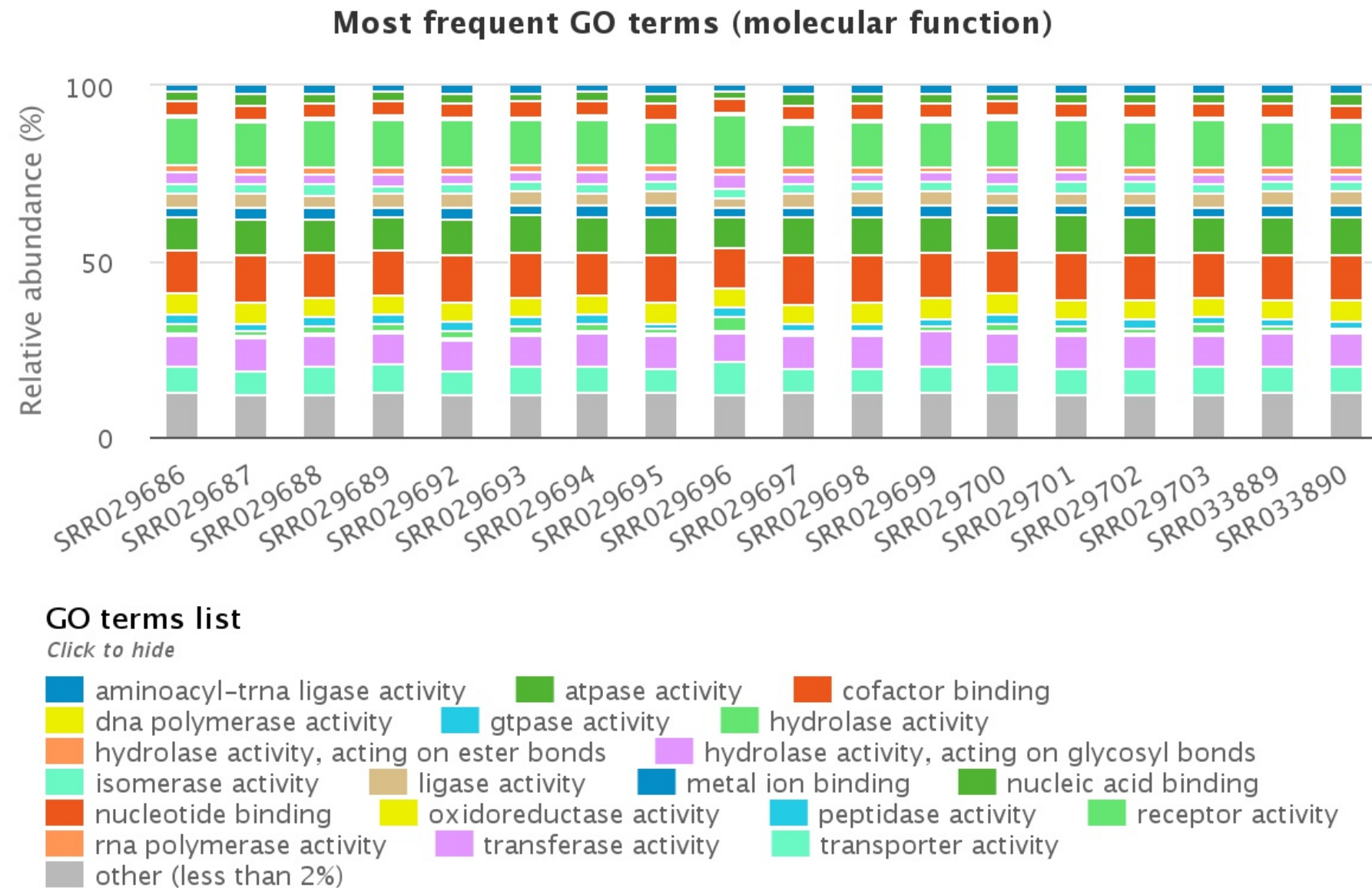


# Comparaison des processus biologiques



Crédit : [EBI metagenomics](#)

# Comparaison des fonctions moléculaires



Crédit : [EBI metagenomics](#)

# Limites et freins

# Conservation des données

# Données

- Besoin d'une profondeur de séquençage importante en métagénomique
- Augmentation exponentielle des données générées
- 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?

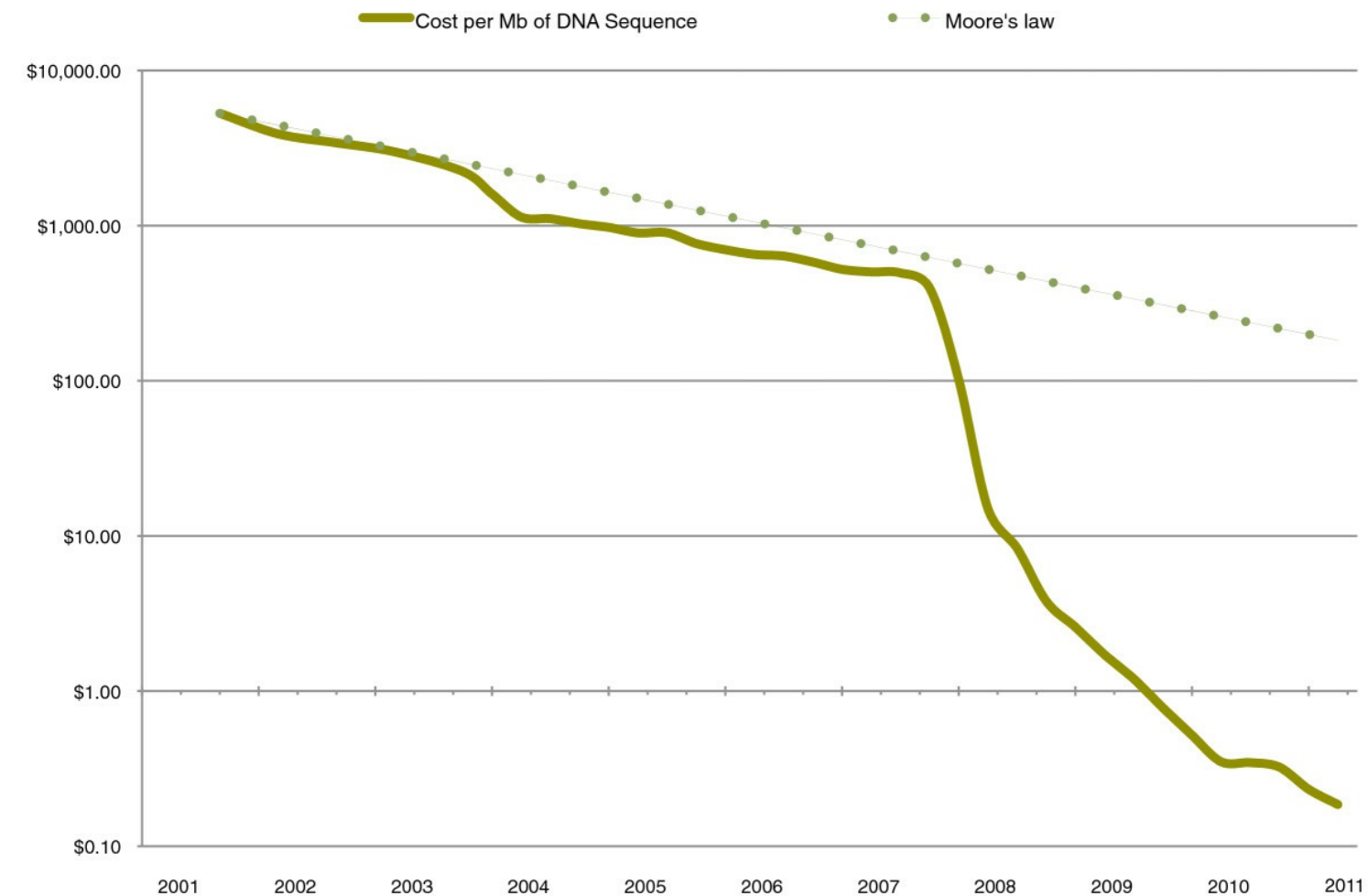
# Principaux dépôts publics

- EBI : ENA, EBI metagenomics,...
- NCBI
- DDBJ

# Traitement des données

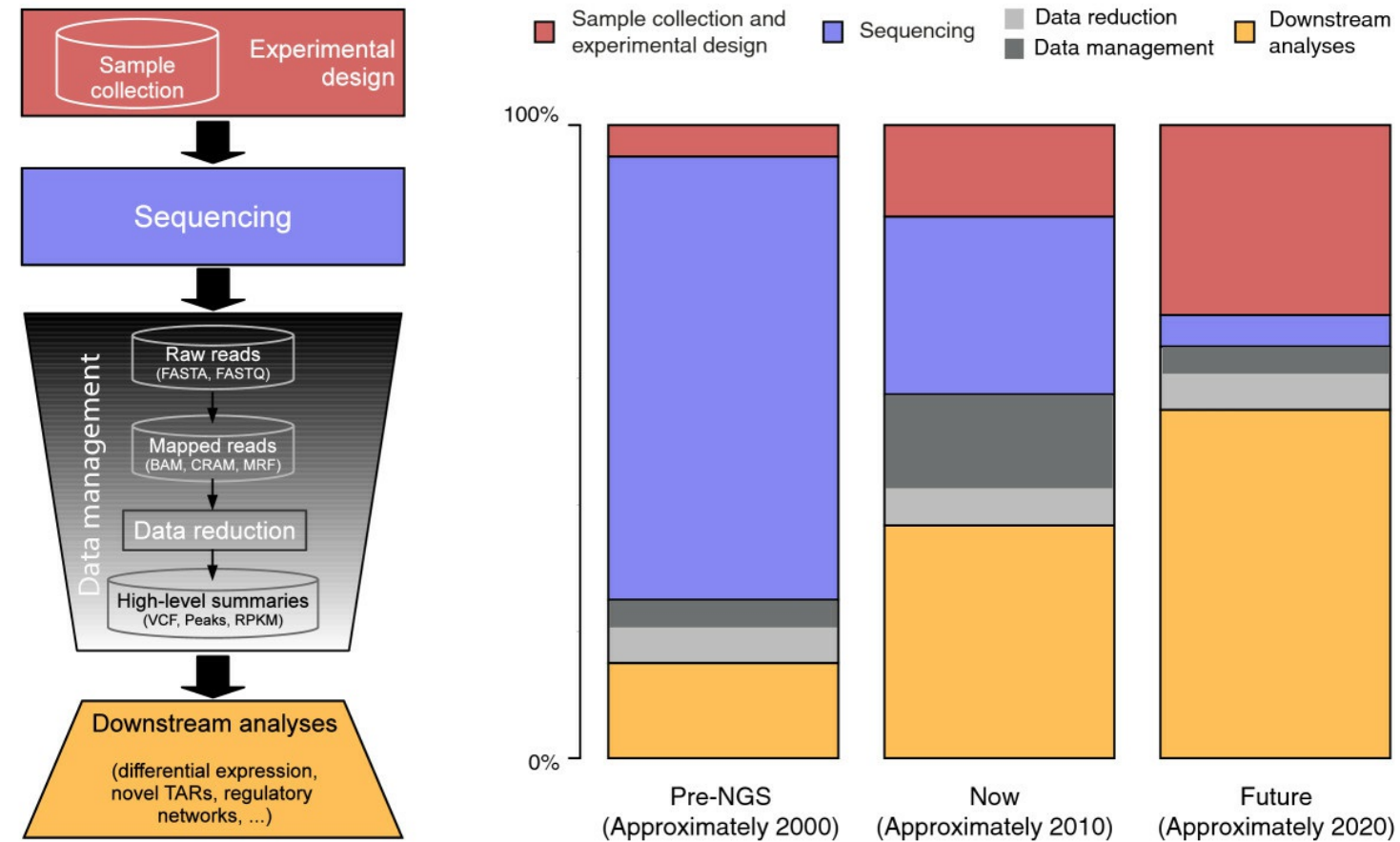


# Fin de la loi de Moore et baisse des prix des séquençage



Crédit : *Sboner et al, Genome Biol, 2011*

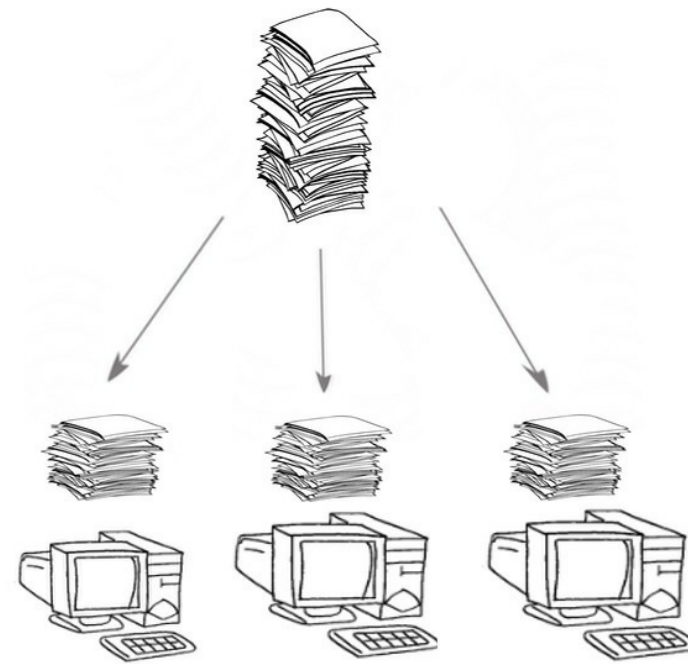
# Part plus importante des coûts de traitement des données



Crédit : *Sboner et al, Genome Biol, 2011*

# 1 solution : parallélisation

## **Divide and Conquer**



- Données
- Tâches
- Algorithme

# Solution pour la parallélisation

Clusters, grilles, cloud

