Métagénomique

Février 2016

Bérénice Batut

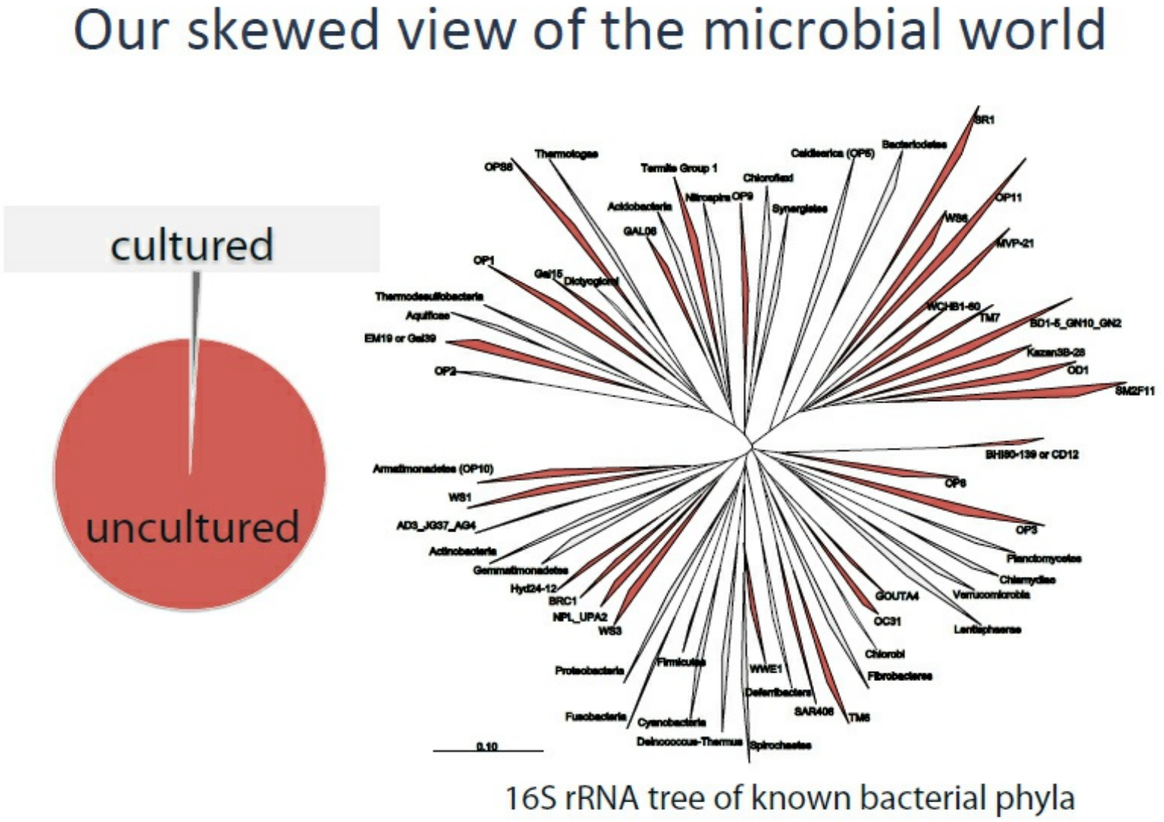
berenice.batut@udamail.fr

Pourquoi la métagénomique?

Ecosystèmes microbiens

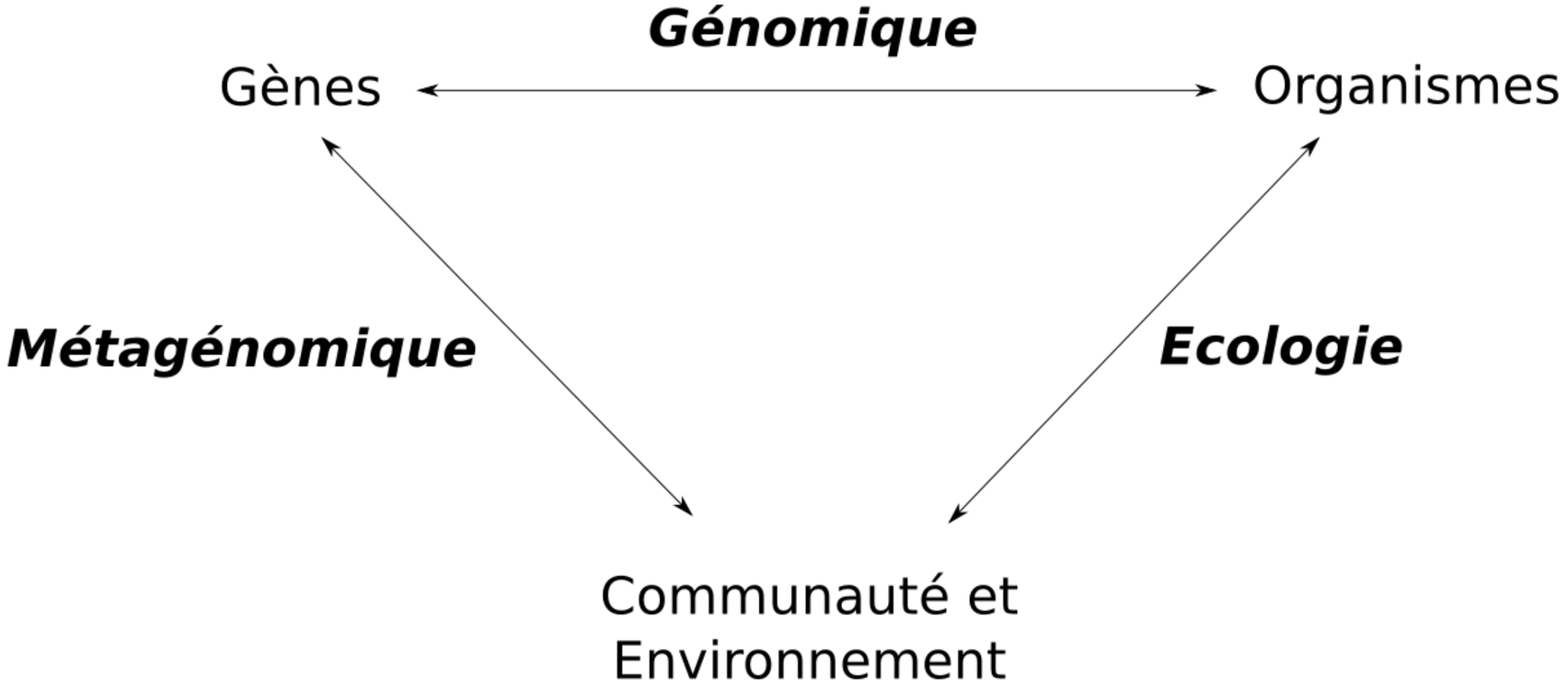
- Majeure partie de la biomasse sur Terre
  - 90% des cellules d'un être humain
- Colonisation de la plupart des niches écologiques

Diversité microbienne



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

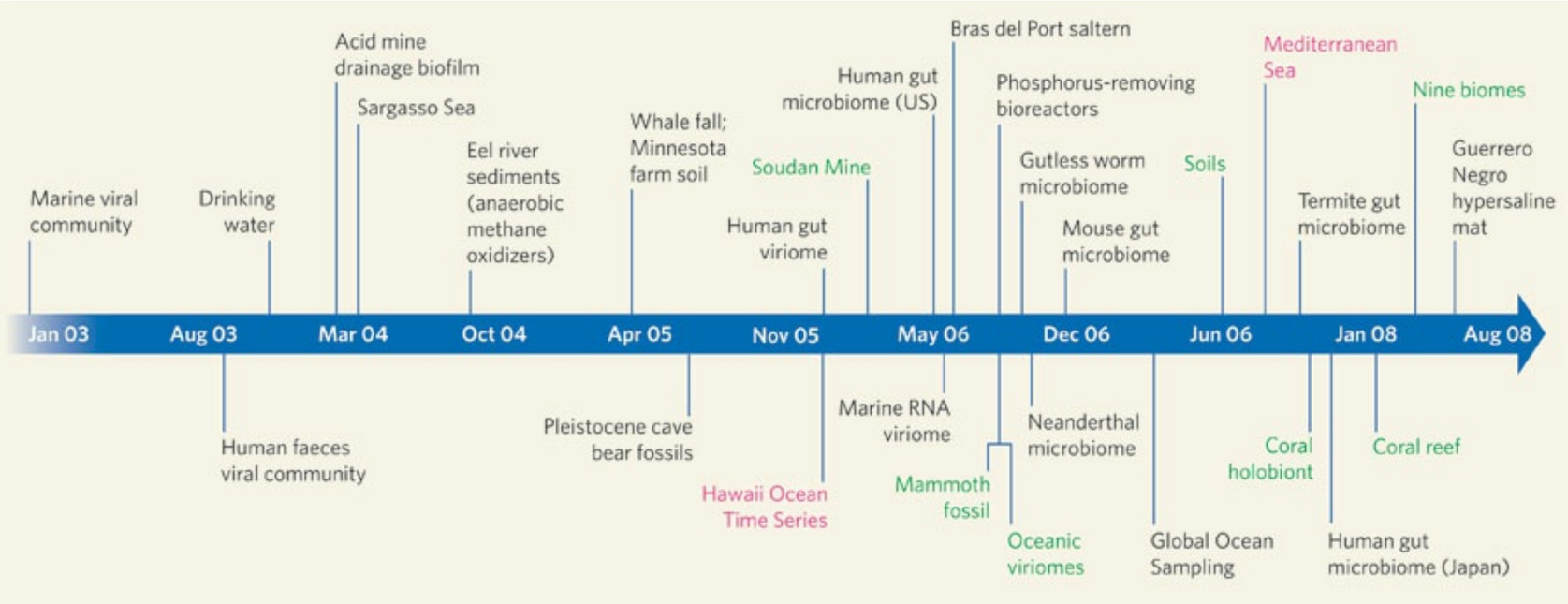
Métagénomique

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)

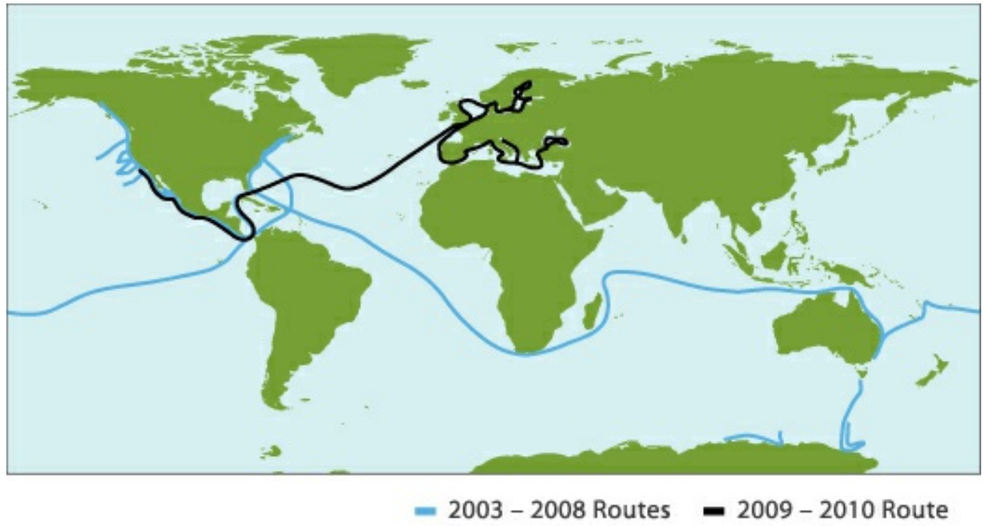


Crédits : *Hugenholtz & Tyson, Nature, 2008*

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 Ocen Sampling (GOS)

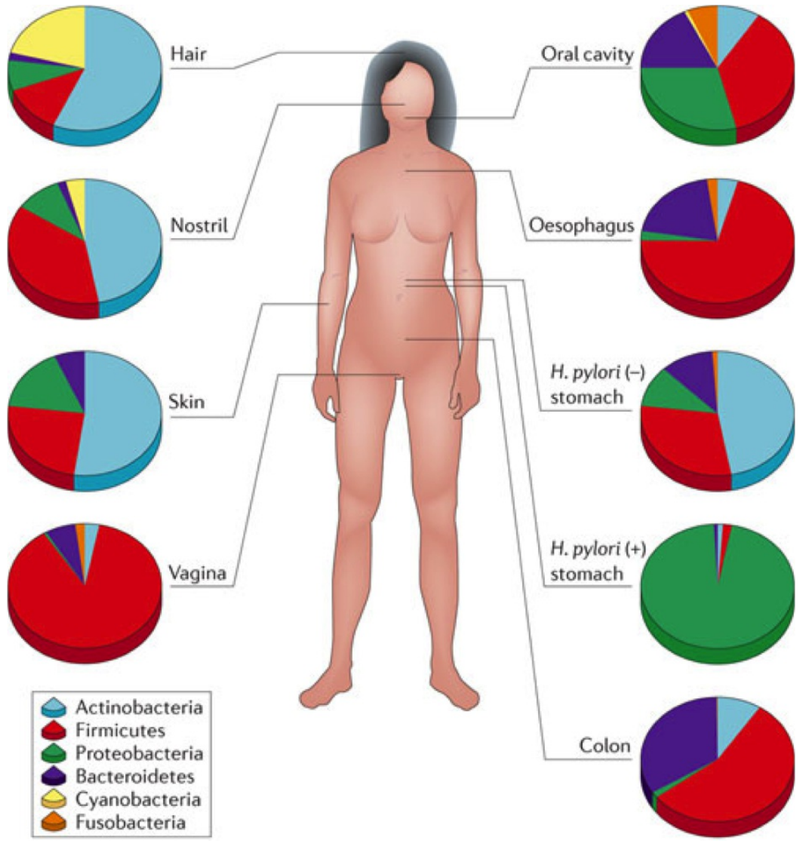


- Estimer la diversité génétique des communautés microbiennes marines
- Comprendre leur rôle dans les processus fondamentaux de la nature

Global Ocen Sampling (GOS)

- Production du plus grand catalogue de gènes de milliers de nouvelles espèces
- Données généréés
  - 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

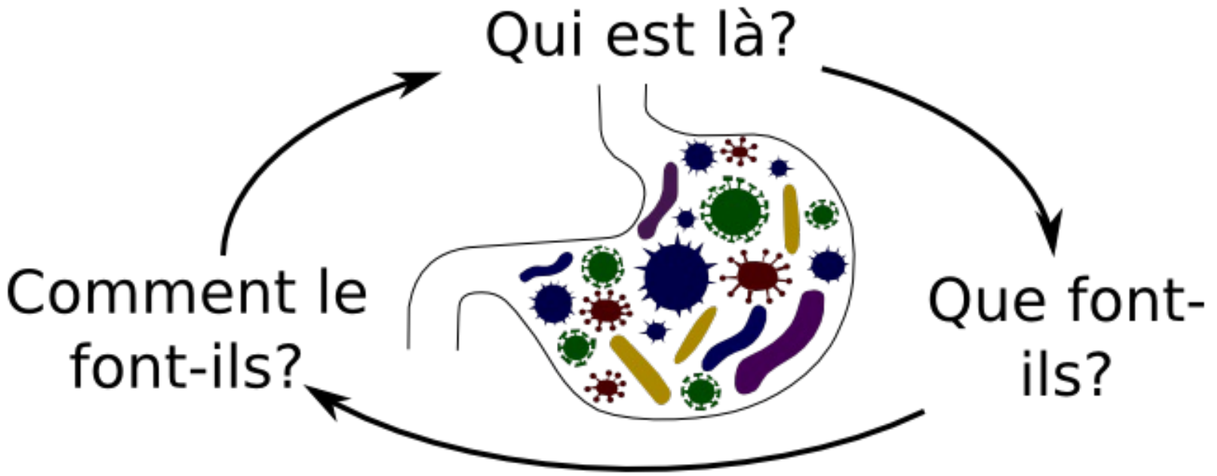
- Metagenomics of the Human Intestinal Tract (MetaHIT)
  - 540 Gb de séquences ADN
  - 134 individus

Santé humaine

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

Principes de la métagénomique

Métagénomique

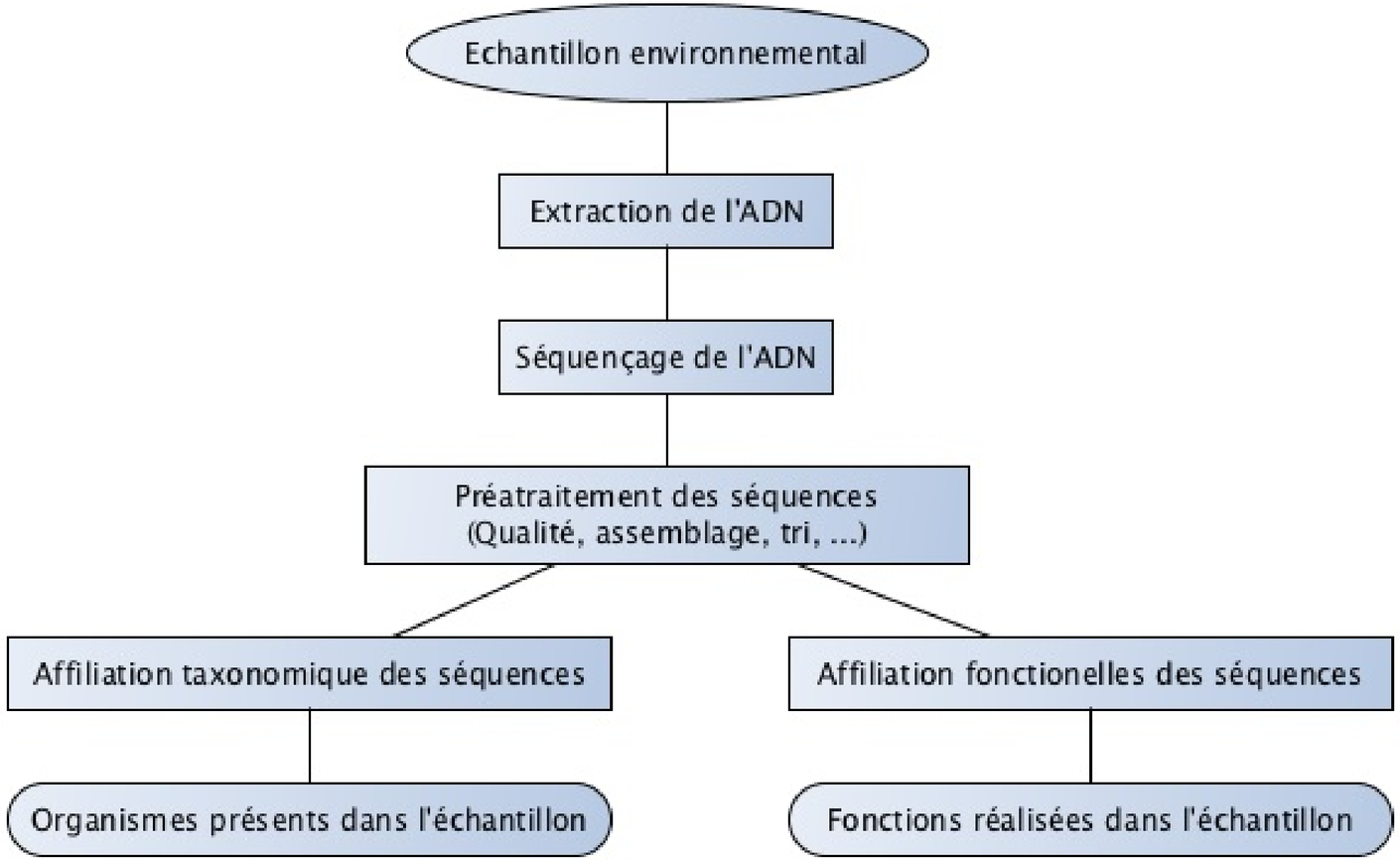


Métagénomique

Objectifs:

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

Métagénomique



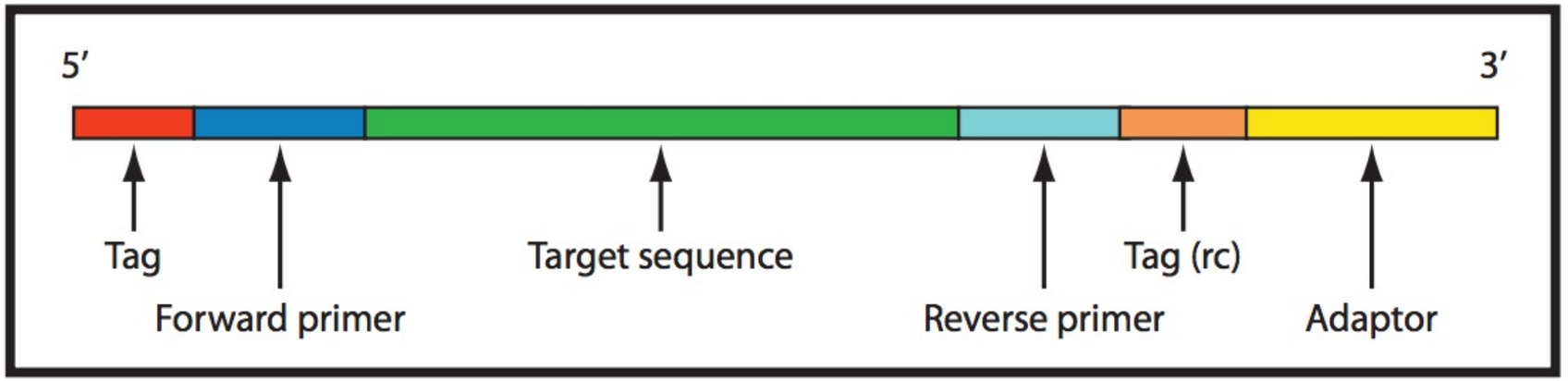
Extraction et séquençage

2 types d’expériences métagénomiques

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

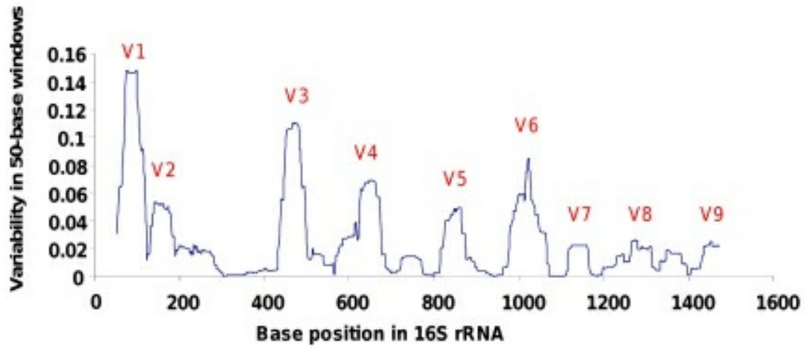
Amplicon

Morceau d'ADN ou ARN amplifié par PCR



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

Utilisation de l'ARNr 16S

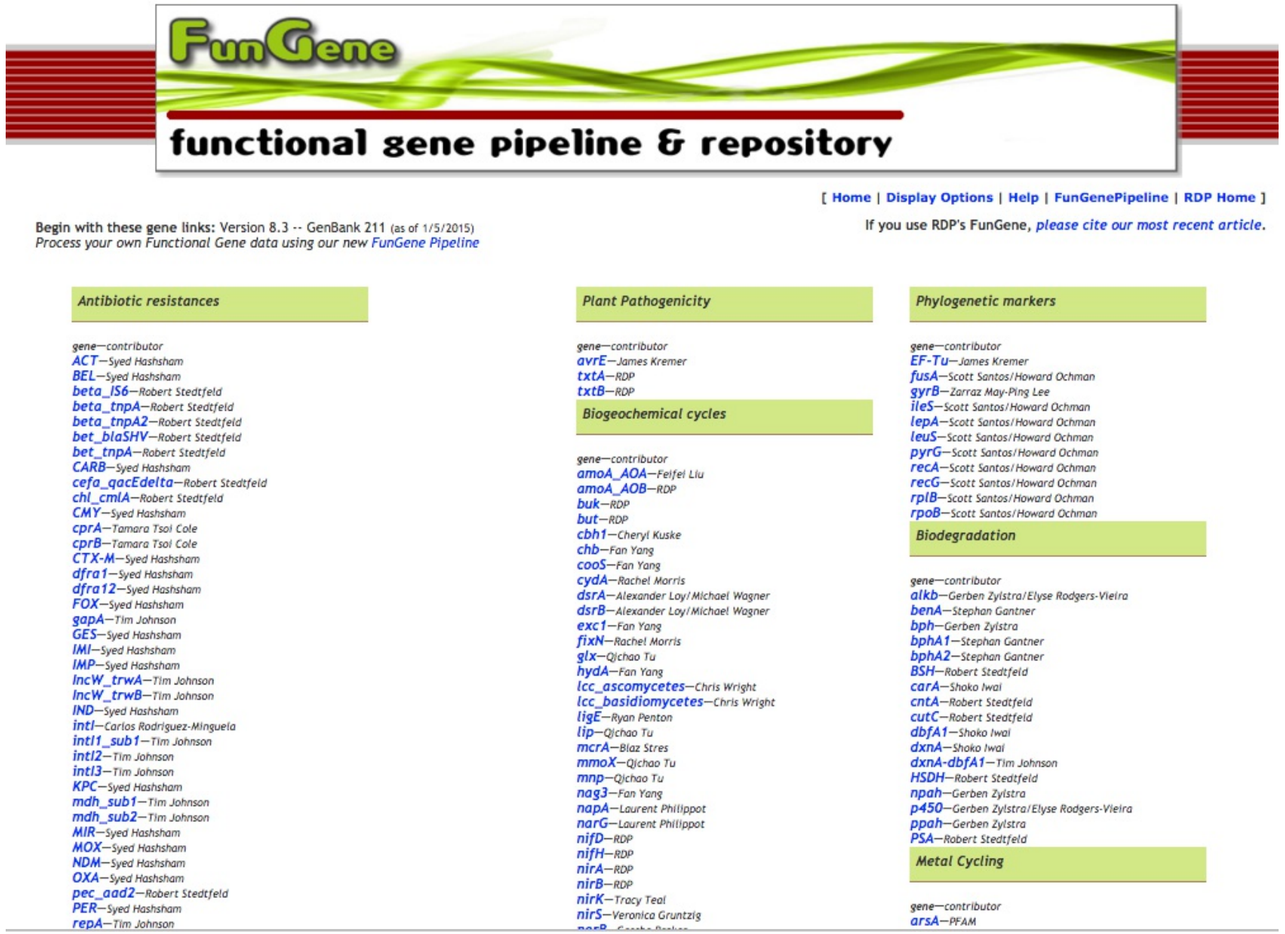


Crédits : *Bodilis et al, PLoS One, 2012*

Séquençage des régions

- V1.3
- V3.4
- V4

Possibilité de cibler d'autres gènes



”Envionmental 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 fonctionelle	Information fonctionelle 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*

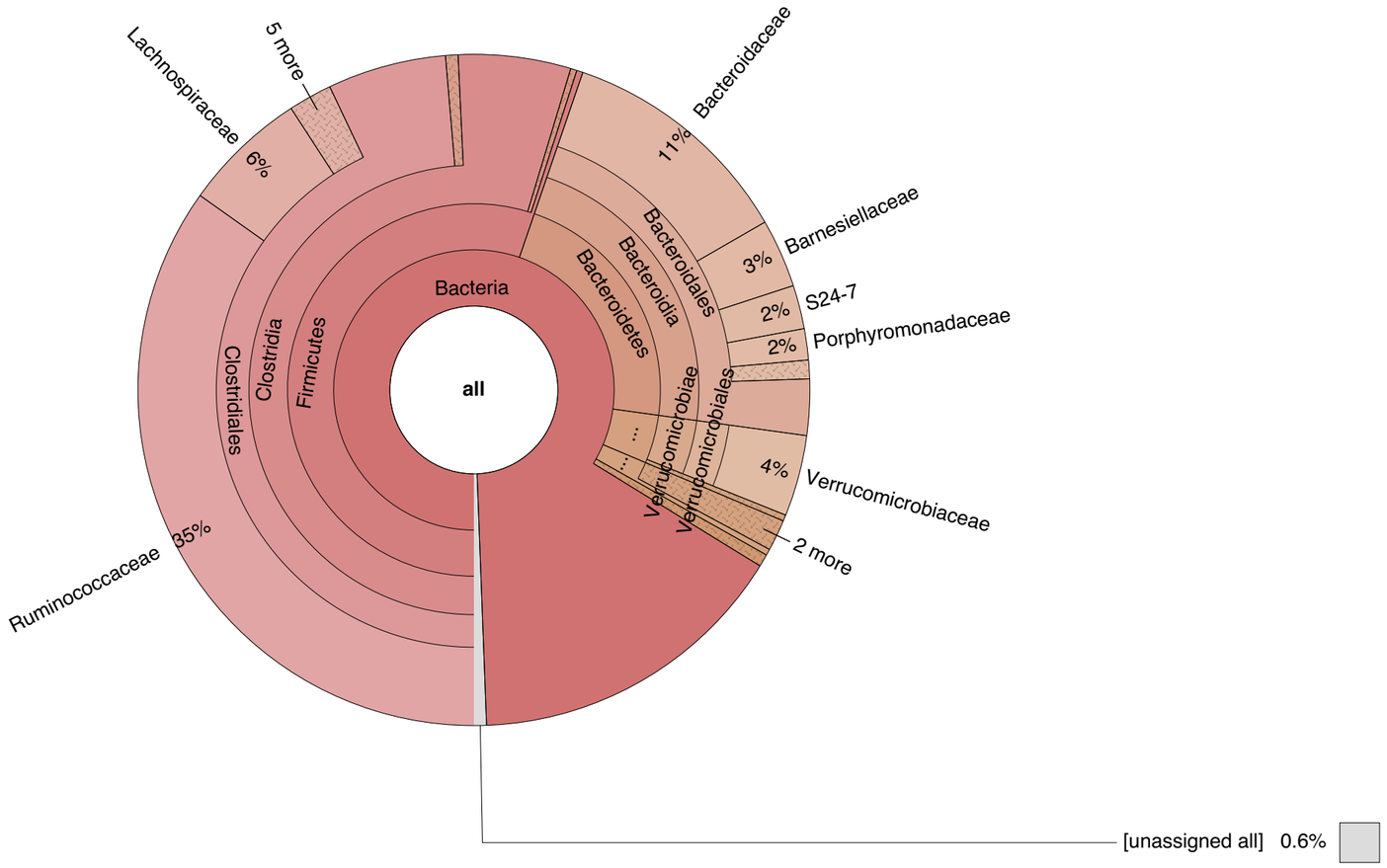
⚠ Chimères : assemblage de séquences issues d'espèces différentes

Tri des séquences

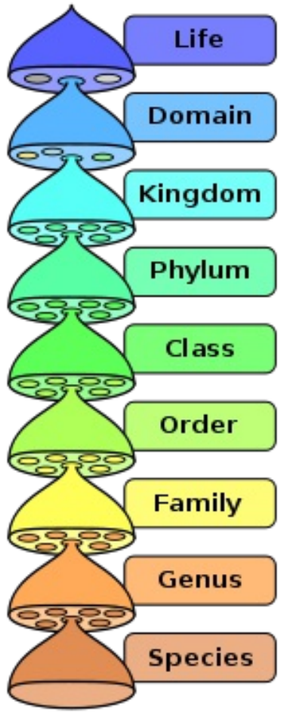
- Utilité
  - Assignation des séquences
  - Recherche dans des bases de données spécifiques
- Caractéristiques d'intérêt
  - Gènes ARNr, ARNt
  - ORF
  - Unités taxonomiques opérationnelles (OTU)
- Quelques outils
  - SortMeRNA, rRNA selector
  - MetaGeneAnnotator, Prodigal, Orphelia
  - tRNAscanSE

Affiliation taxonomique des séquences

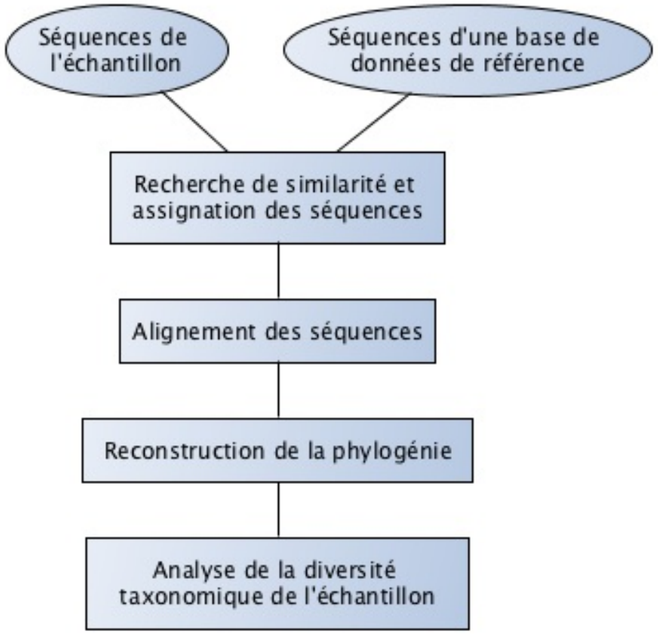
Pourquoi l'affiliation taxonomique?



Pourquoi l'affiliation taxonomique?



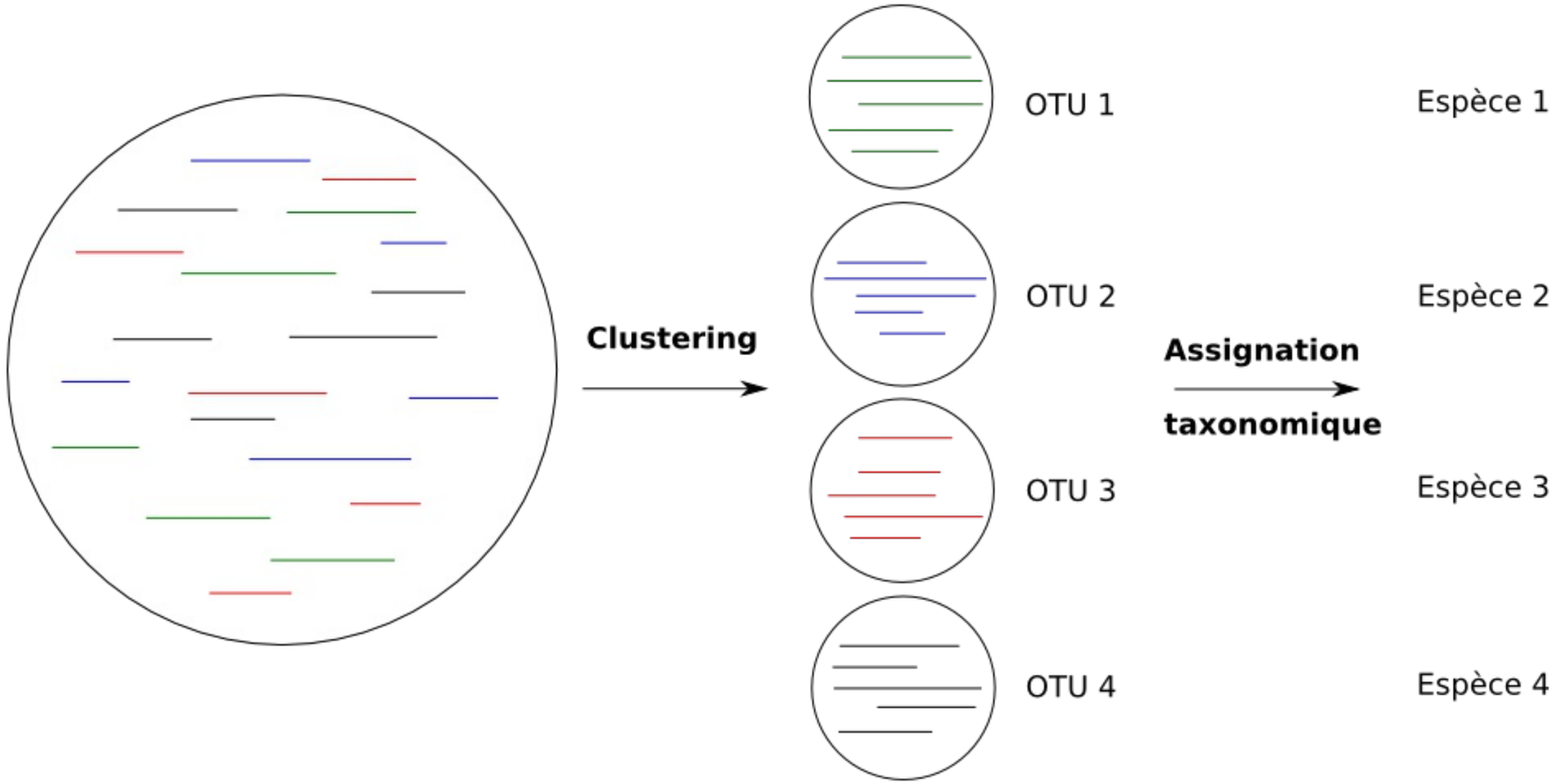
Procédure standard



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)



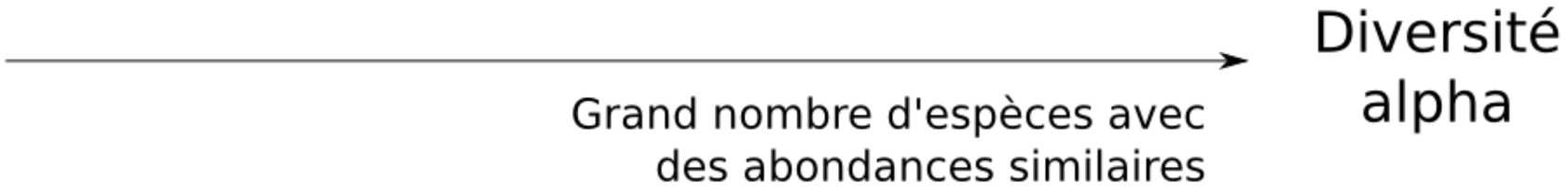
Recherche de similarité

- 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

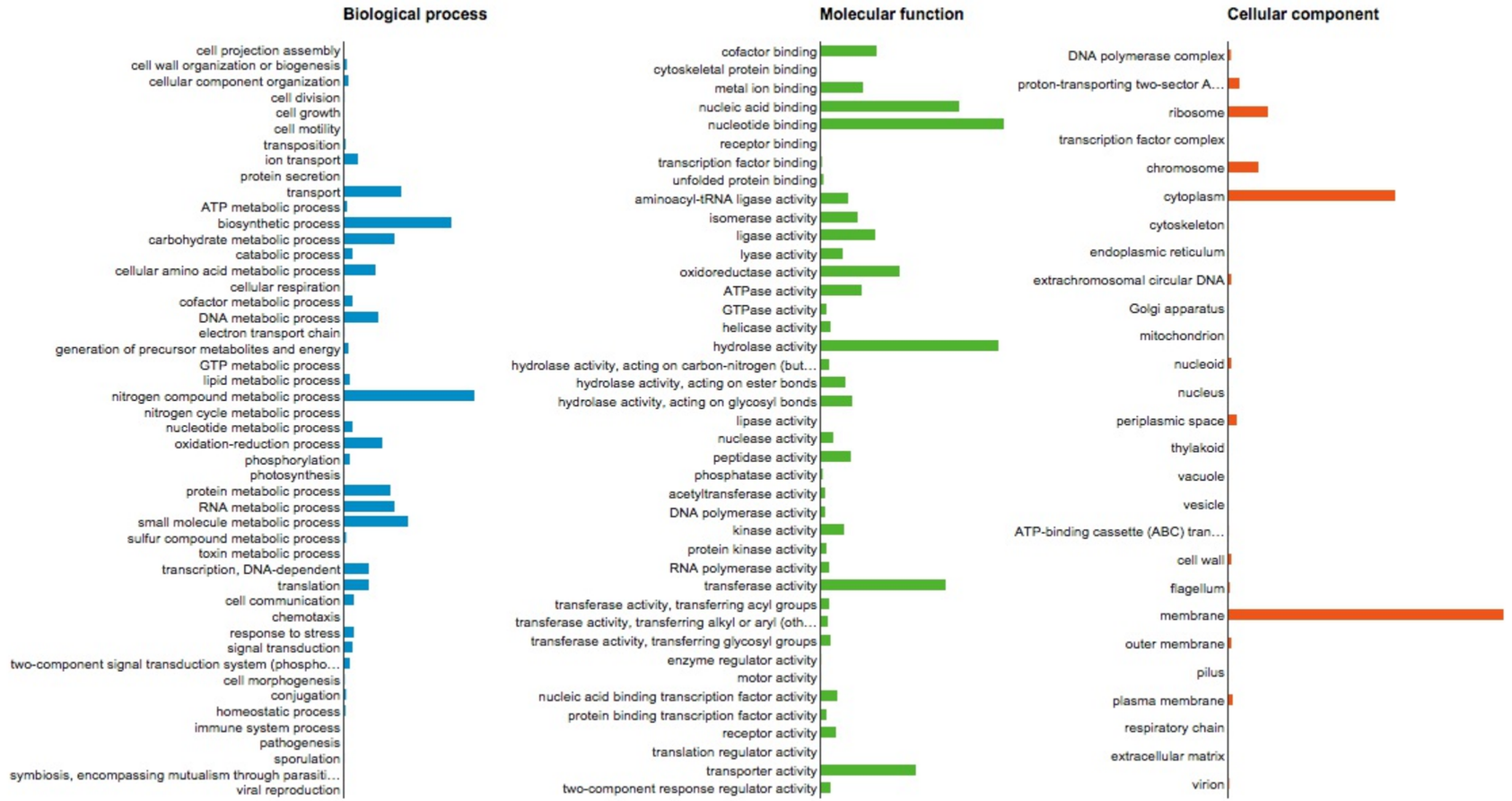
Alpha diversité

- 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)
  - Egalité
    - Proportion d'espèces présentes

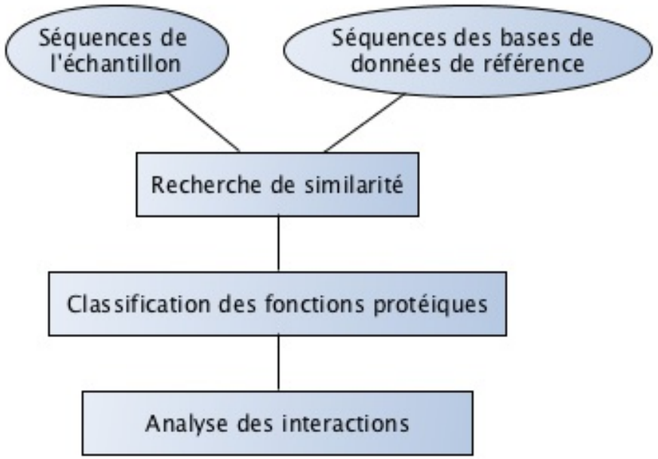


Affiliation fonctionnelle

Pourquoi l'affiliation fonctionnelle ?



Procédure


Séquences de l'échantillon
Séquences des bases de données de référence
Recherche de similarité
Classification des fonctions protéiques
Analyse des interactions

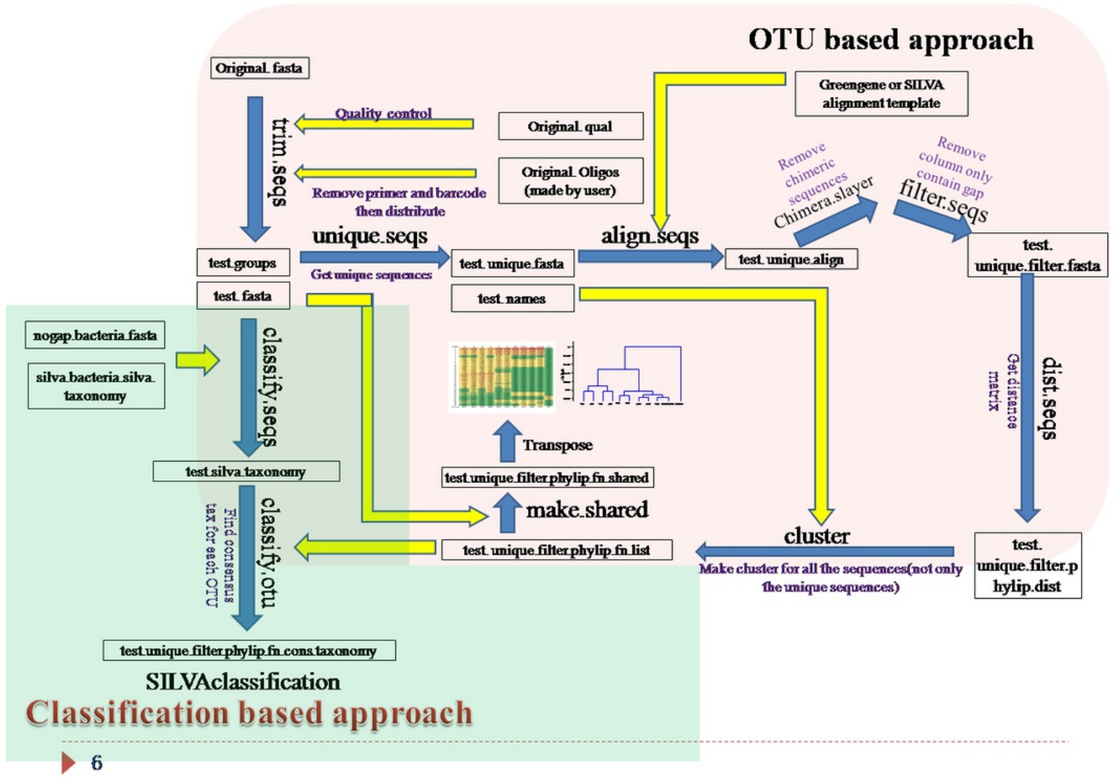
Classification des fonctions protéiques

- 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

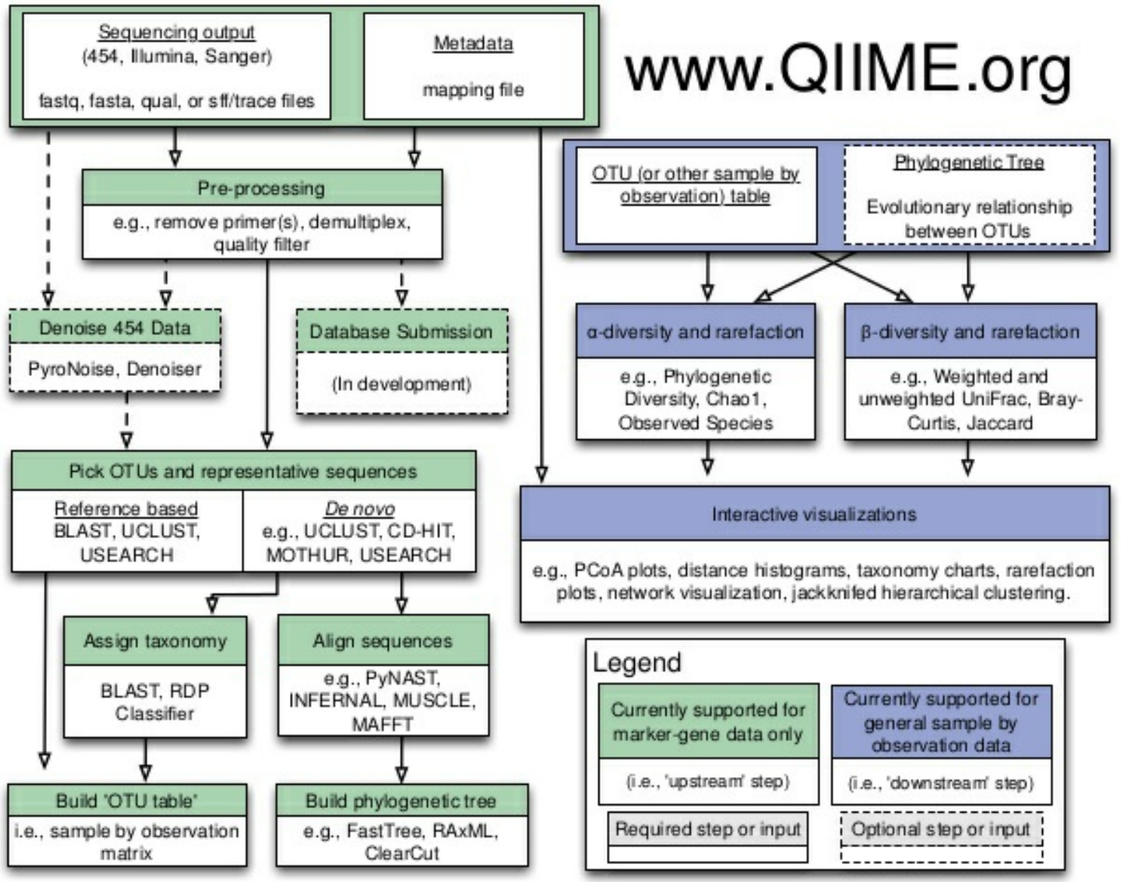
Mothur

Schloss et al, Applied Env Microbiol, 2009



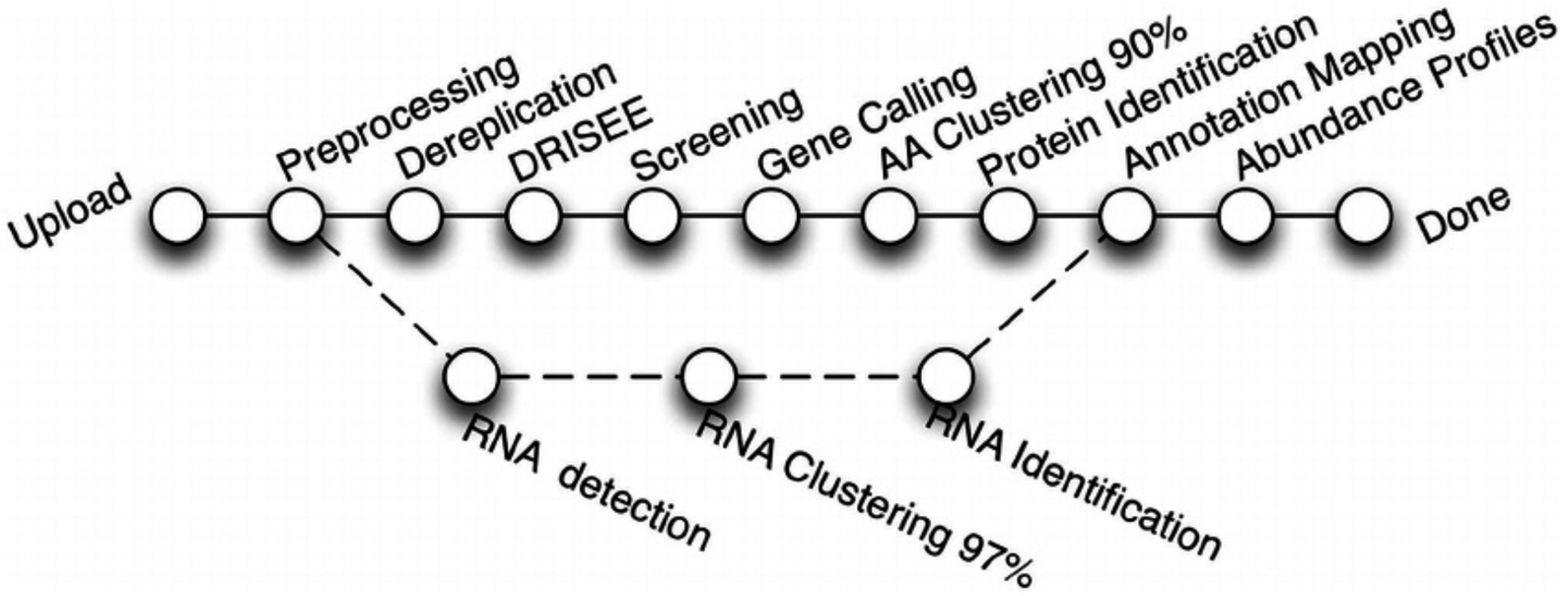
QIIME

Caporaso et al, Nature Methods, 2010



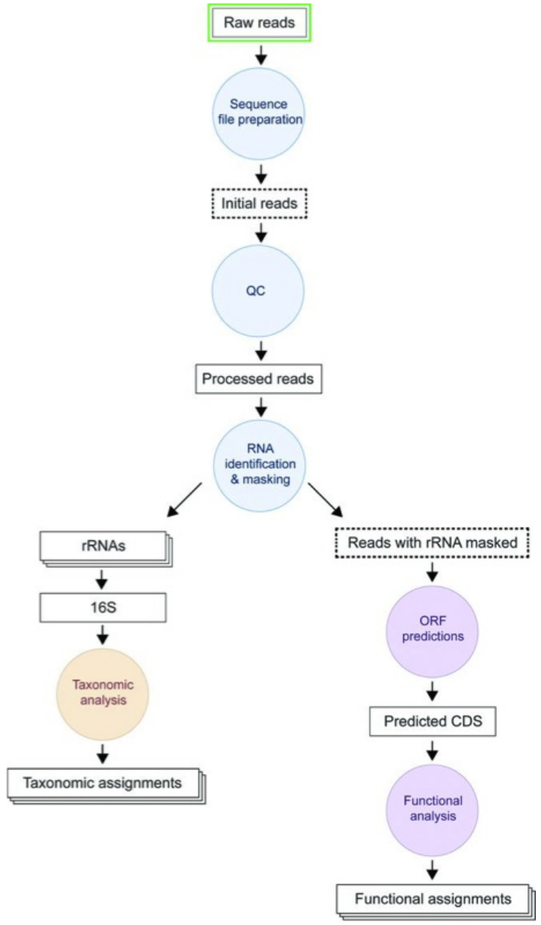
MG-Rast

Meyer et al, BMC Bioinformatics, 2008



EBI metagenomics

Mitchell et all, 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

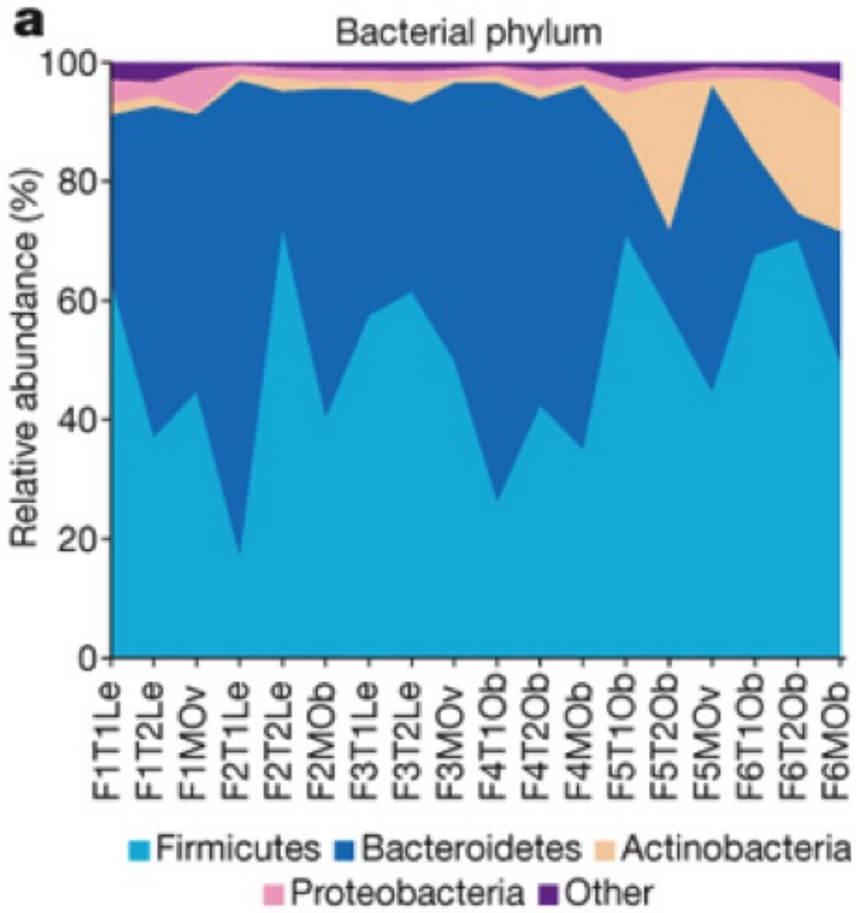
MIMARKS : Minimum Information about a MARKer gene Sequence

Métagénomique comparative

Turnbaugh et al, Nature, 2009

Comparaison de la diversité

Comparaison des organismes présents



Crédit : *Turnbaugh et al, Nature, 2009*

Diversité beta

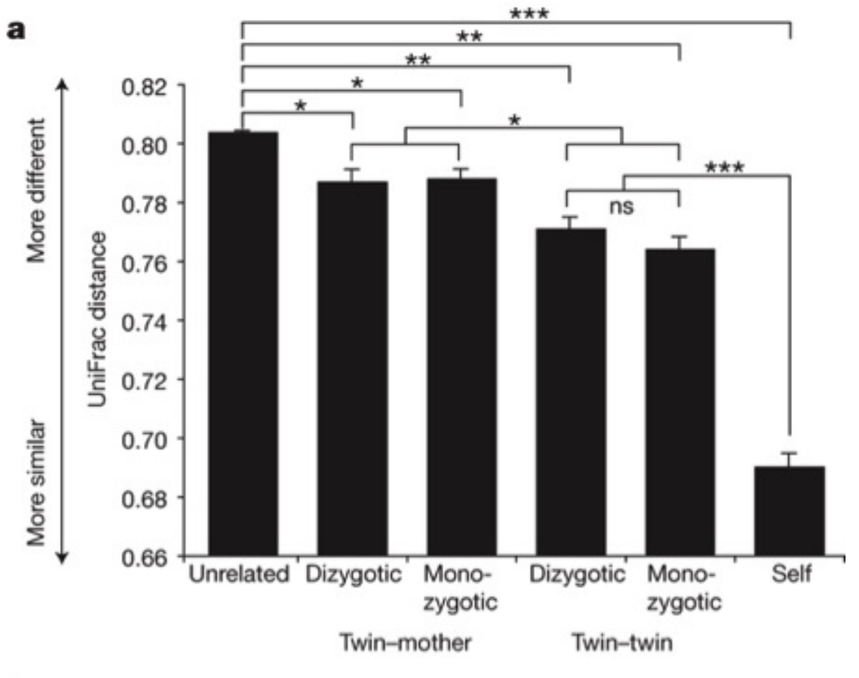
- Notion définie par *Whittaker, Evolution and Measurement of Species Diversity, 1972*
- Comparasion entre 2 échantillons de la différence de communauté (difference des espèces ou OTU)

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

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

α : moyenne du nombre d'espèces trouvées au sein des 2 échantillons

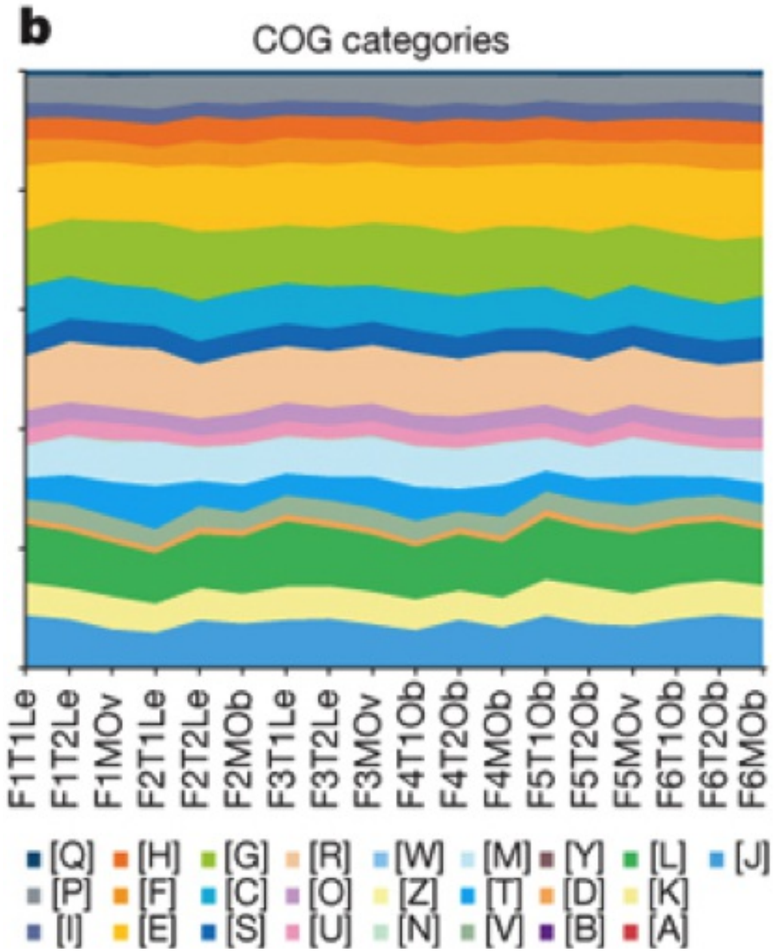
Comparaison de la diversité avec la distance Unifrac



Crédit : *Turnbaugh et al, Nature, 2009*

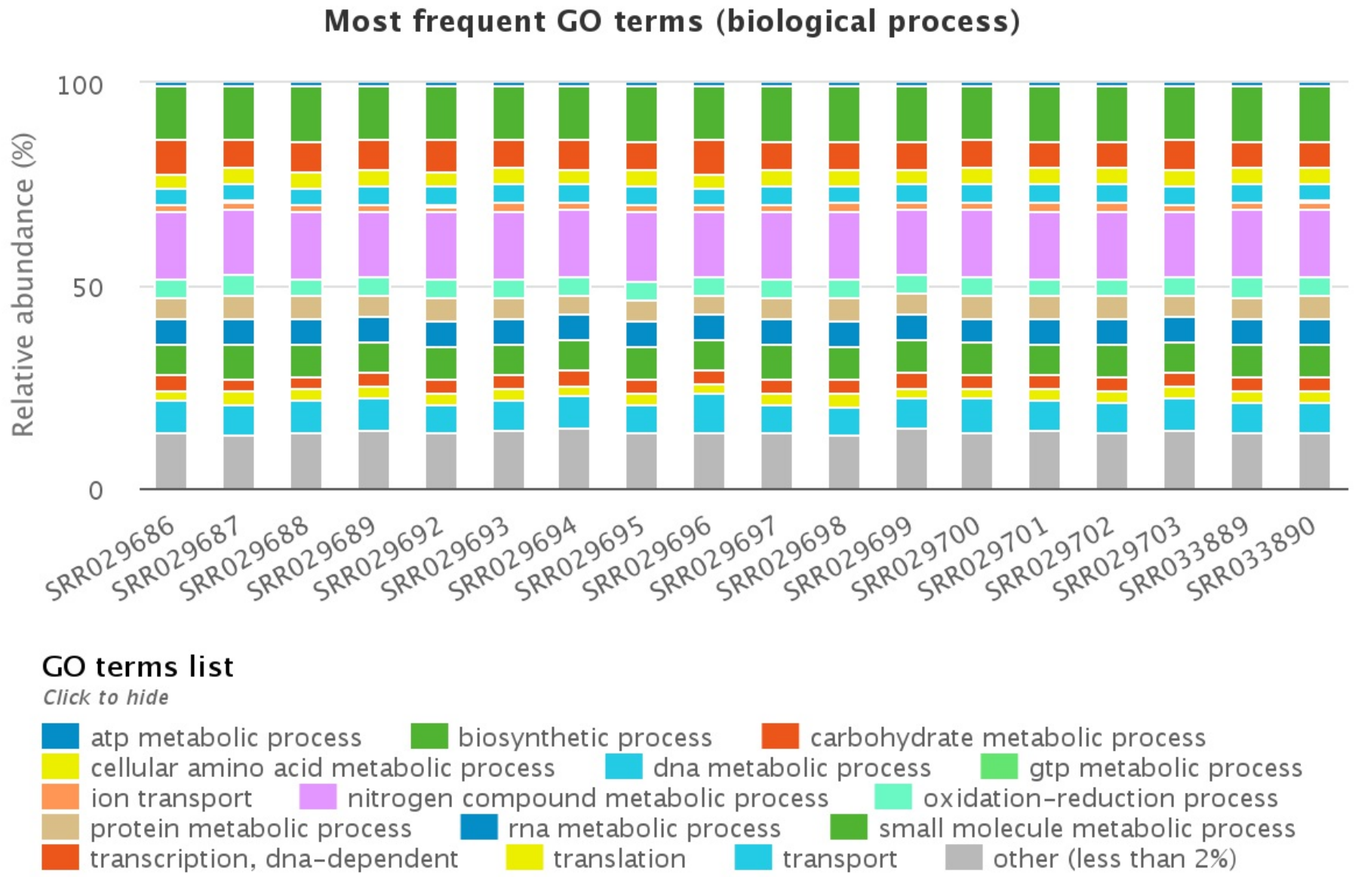
Comparaison des fonctions réalisées

Comparaison des catégories COG



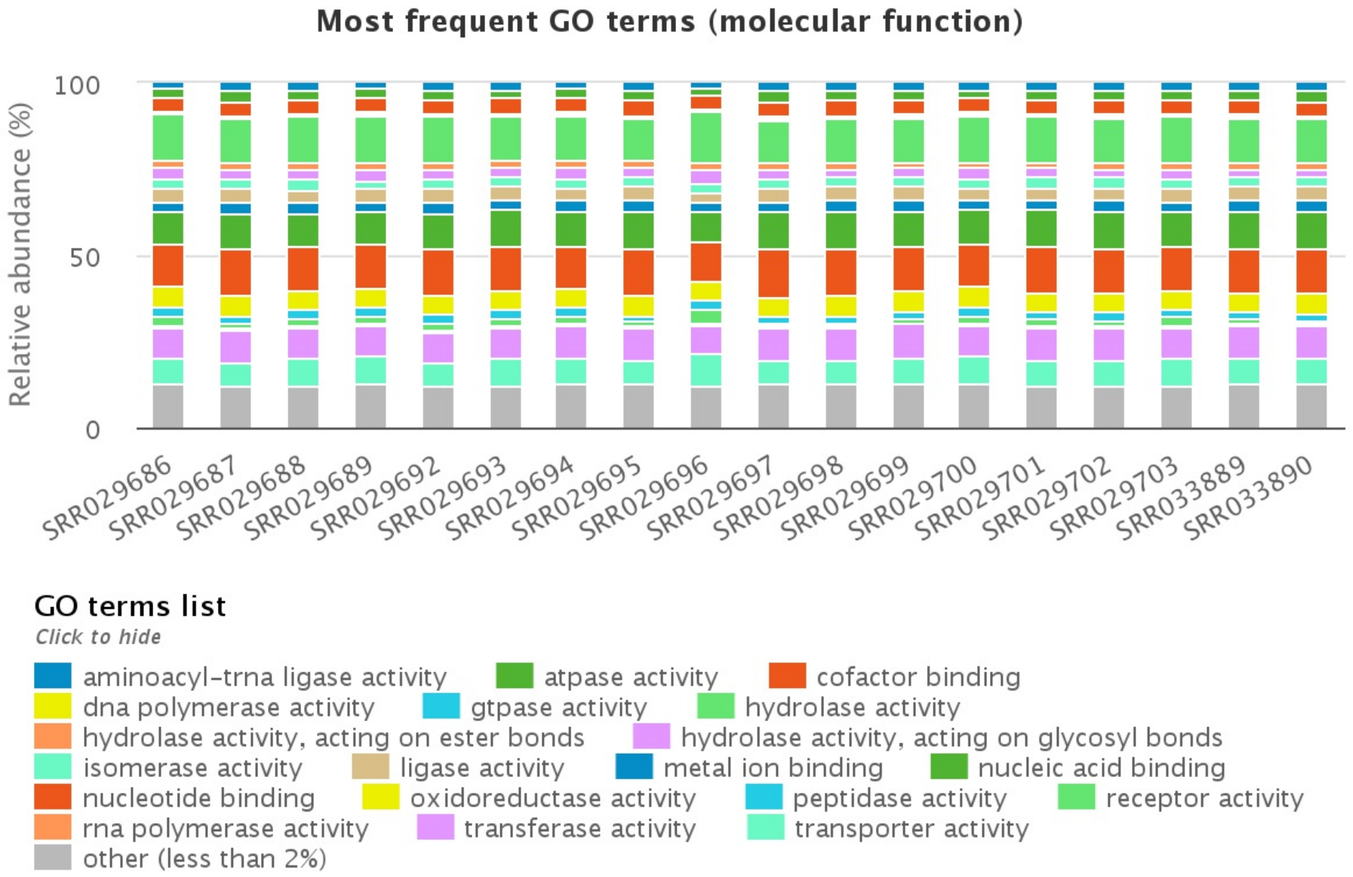
Crédit : *Turnbaugh et al, Nature, 2009*

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

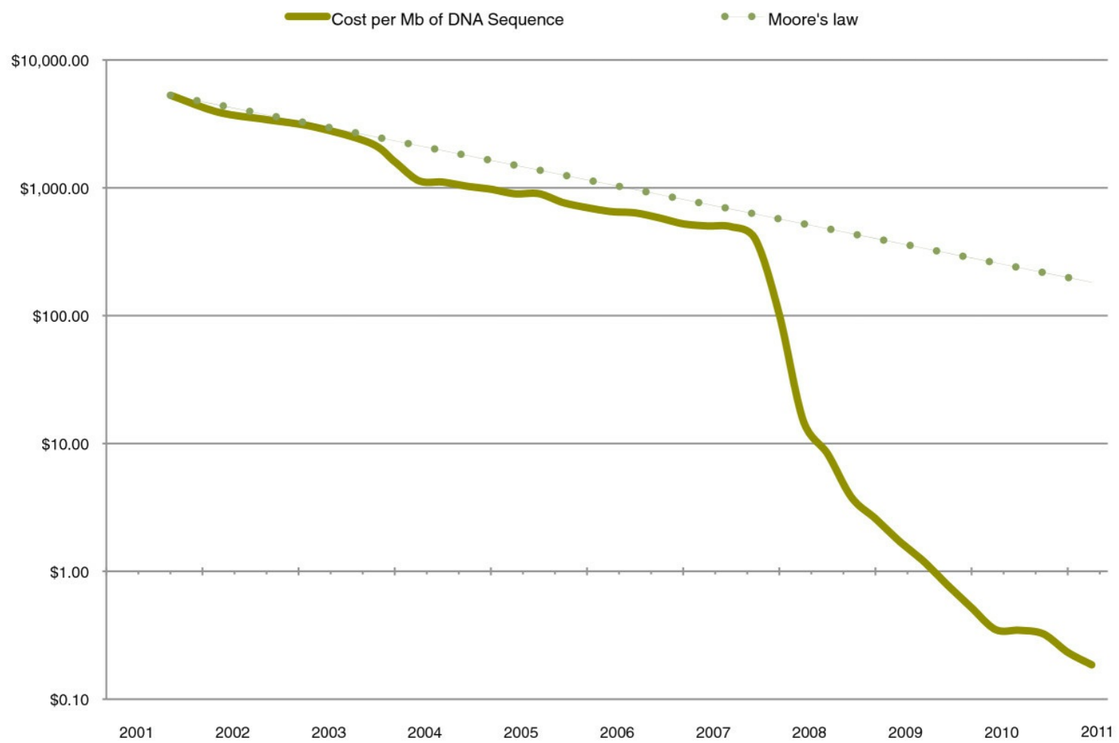
- Quoi?
  - Données brutes
  - Résultats des analyses
  - Description des données
- Où?
  - Localement?
  - Cloud?
  - Dépôts publics?

Principaux dépôts publics

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

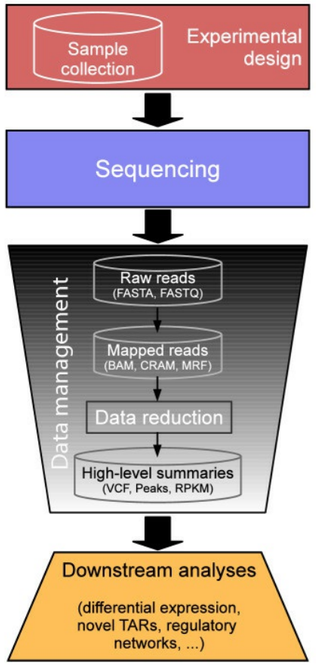
Traitement des données

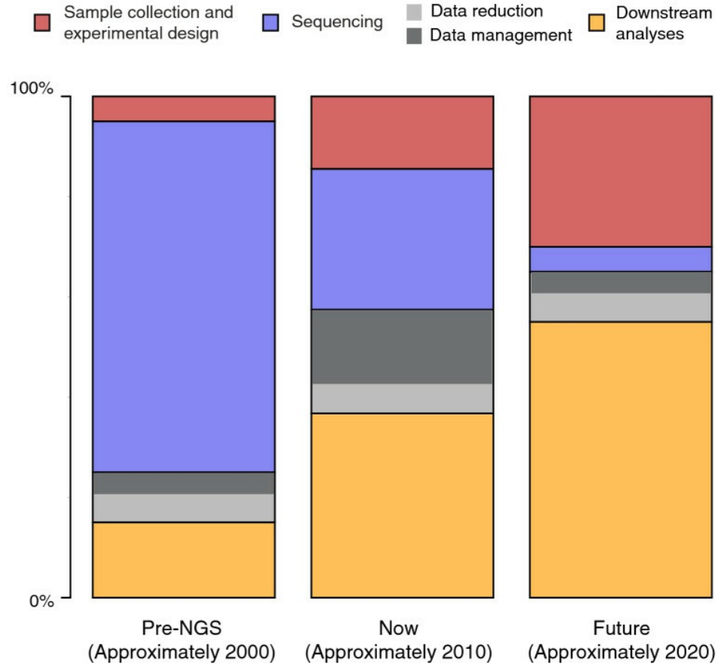
Fin de la loi de Moore et baisse des prix des séquençage



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

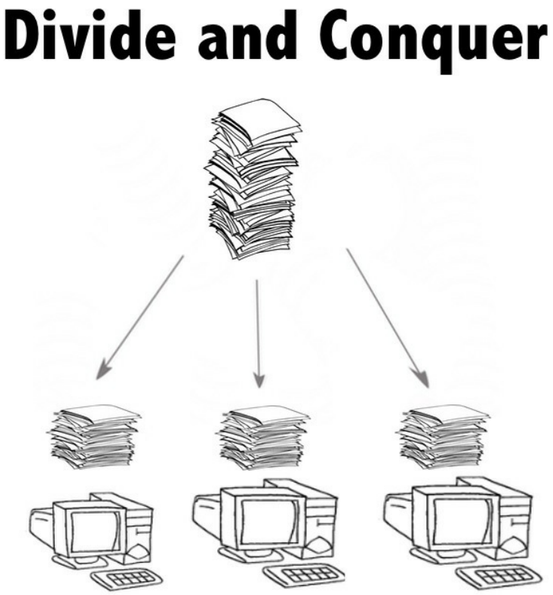
Part plus importante des coûts de traitement des données





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

1 solution : parallélisation



- Données
- Tâches
- Algorithme

Solution pour la parallélisation

Clusters, grilles, cloud