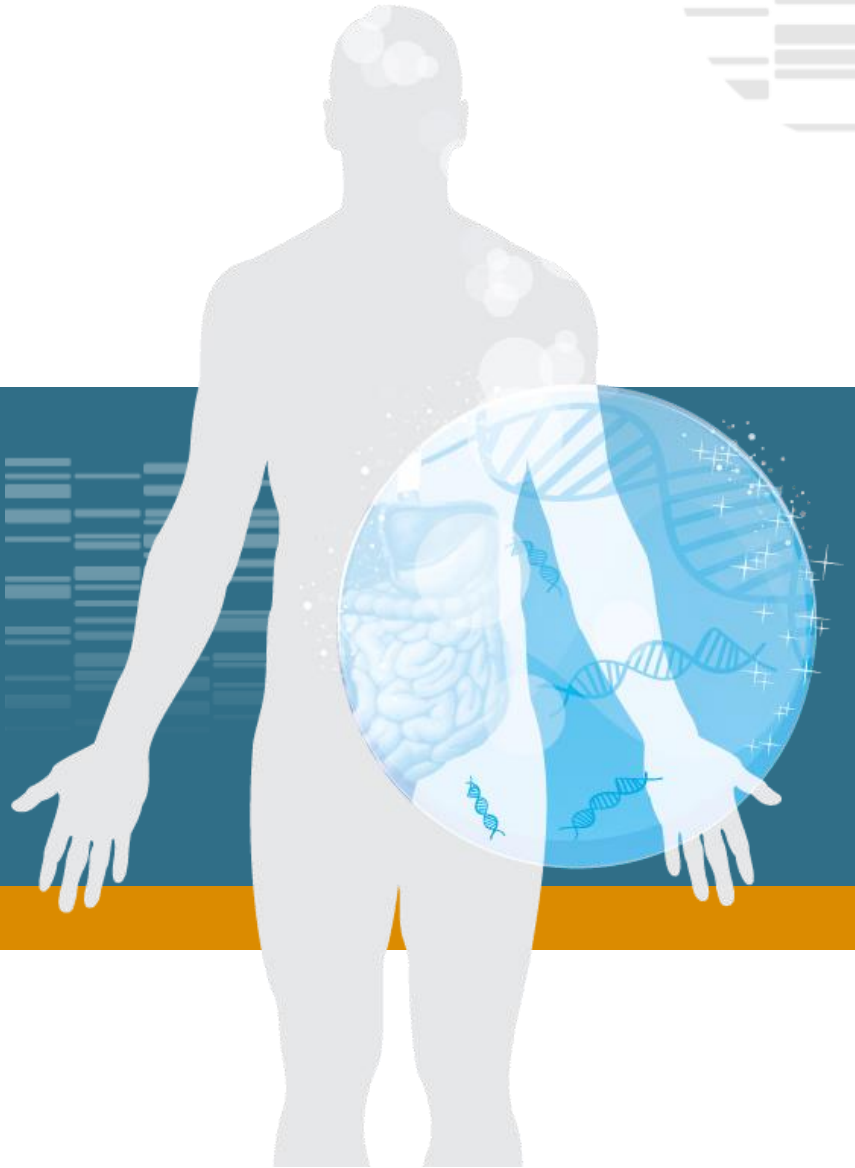


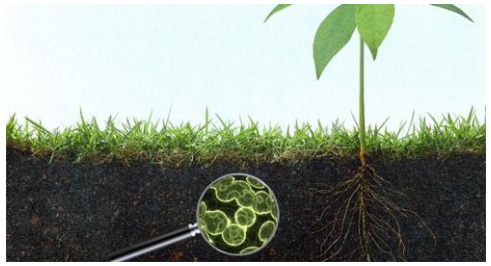
Introduction à l'analyse bioinformatique de données de séquençage métagénomique shotgun

Florian Plaza Oñate – florian.plaza-onate@inra.fr

PEPI IBIS – 07/06/2019

Archées, bactéries, eucaryotes unicellulaires, virus… des microorganismes ubiquitaires

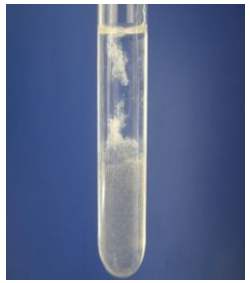
- Les microorganismes colonisent tous types d'habitats sur terre

- Jouent des rôles cruciaux
  - Cycle de l'azote, du phosphore…
  - Rôle clef dans le développement et la santé de l'hôte

- Comment caractériser ces communautés microbiennes?

Caractérisation de communautés microbiennes par culture



Prélèvement d'échantillons → Isolement & culture → Extraction de l'ADN → Séquençage du génome

- Nombreux microorganismes difficilement cultivables en laboratoire
- Substrats et conditions de culture inconnus / difficiles à mettre en œuvre
- Espèces interdépendantes / en symbiose

Caractérisation de communautés microbiennes par séquençage métagénomique shotgun

- Séquençage aléatoire non ciblé (*shotgun*) de fragments d'ADN (*-génomique*) provenant d'une communauté microbienne (*meta-*) sans isolement et culture préalable.
- Les protocoles expérimentaux impactent fortement la qualité des données générées

Caractérisation de communautés microbiennes par séquençage métagénomique shotgun

- Séquençage aléatoire non ciblé (*shotgun*) de fragments d'ADN (-*génomique*) provenant d'une communauté microbienne (*meta*-) sans isolement et culture préalable.

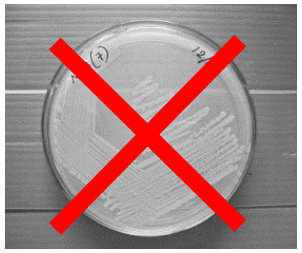
Prélèvement d'échantillons

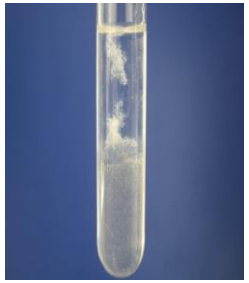
Caractérisation de communautés microbiennes par séquençage métagénomique shotgun

- Séquençage aléatoire non ciblé (*shotgun*) de fragments d'ADN (*-génomique*) provenant d'une communauté microbienne (*meta-*) sans isolement et culture préalable.

Caractérisation de communautés microbiennes par séquençage métagénomique shotgun

- Séquençage aléatoire non ciblé (*shotgun*) de fragments d'ADN (*-génomique*) provenant d'une communauté microbienne (*meta-*) sans isolement et culture préalable.

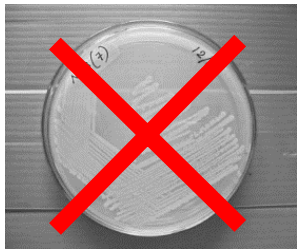


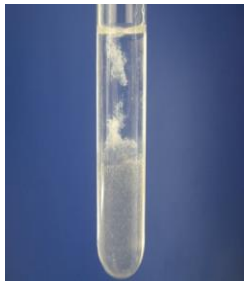


Prélèvement d'échantillons → ~~Isolement & culture~~ → Extraction de l'ADN

Caractérisation de communautés microbiennes par séquençage métagénomique shotgun

- Séquençage aléatoire non ciblé (*shotgun*) de fragments d'ADN (*-génomique*) provenant d'une communauté microbienne (*meta-*) sans isolement et culture préalable.

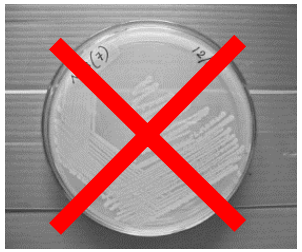


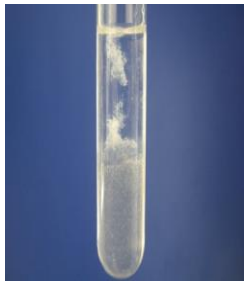


Prélèvement d'échantillons → ~~Isolement & culture~~ → Extraction de l'ADN → Séquençage

Caractérisation de communautés microbiennes par séquençage métagénomique shotgun

- Séquençage aléatoire non ciblé (*shotgun*) de fragments d'ADN (*-génomique*) provenant d'une communauté microbienne (*meta-*) sans isolement et culture préalable.





```
@ERR1397942.1 FCC34VAACXX:7:1101:1460:1822/1
NAACAATCTGGACGAGCTGTTGTACGATGTATCGCTTGCCGGAT
+
#1:DDDDDHHFFFIIJIDHHIIEHIGGCG3CDFHHJGIIGEFGE
@ERR1397942.2 FCC34VAACXX:7:1101:1495:1826/1
NCTGAAATTTTCAAAAAAGAAAATCCCAAAGCGGCAGAACGTCC
+
#1=DFFFFHHHHGJJJJJIJJJJJIJJJEIIGGHIGGGI@GCCG
```

Prélèvement d'échantillons → ~~Isolement & culture~~ → Extraction de l'ADN → Séquençage → Traitement bioinfo

L'analyse de données séquençage métagénomique, un défi bioinformatique

- Des données relativement volumineuses
- Des données complexes
  - « Plusieurs puzzles dont les pièces ont été mélangées dans une seule boite »
  - Collection de divers microorganismes parfois étroitement apparentés
  - Le lien entre un read et le génome dont il provient est inconnu
  - Abondance relative des microorganismes variant sur plusieurs ordres de grandeur
- Besoin d'outils et de bases de données dédiées

Types d'analyses de données métagénomiques

- Profilage taxonomique
  Quels sont les microorganismes présents et quelle est leur abondance?
- Assemblage métagénomique
- Métagénomique au niveau souche
- Profilage fonctionnel
  Que peuvent t'il faire ?
- Métagénomique comparative
  Quelles sont les différences/similitudes entre plusieurs métagénomes?

Pré-traitement des données

- Raccourcissement et suppression des reads mauvaise qualité
- Suppression des adaptateurs de séquençage

Trimmomatic: a flexible trimmer for Illumina sequence data

Anthony M. Bolger, Marc Lohse, Bjoern Usadel ✉ Author Notes

Bioinformatics, Volume 30, Issue 15, 1 August 2014, Pages 2114–2120, https://doi.org/10.1093/bioinformatics/btu170

- Suppression des reads provenant de l'hôte

Types d'analyses de données métagénomiques

- **Profilage taxonomique**
  Quels sont les microorganismes présents et quelle est leur abondance?

- Assemblage métagénomique

- Métagénomique au niveau souche

- Profilage fonctionnel
  Que peuvent t'il faire ?

- Métagénomique comparative
  Quelles sont les différences/similitudes entre plusieurs métagénomes?

Profilage taxonomique par classification exhaustive des lectures

- Alignement des reads contre des génomes de référence

Method

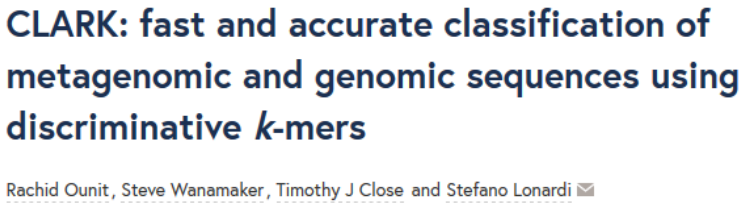
Centrifuge: rapid and sensitive classification of metagenomic sequences

Daehwan Kim,[1,4] Li Song,[1,2,4] Florian P. Breitwieser,[1,4] and Steven L. Salzberg[1,2,3]

- Approches par « pseudo »-alignement

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood and Steven L Salzberg



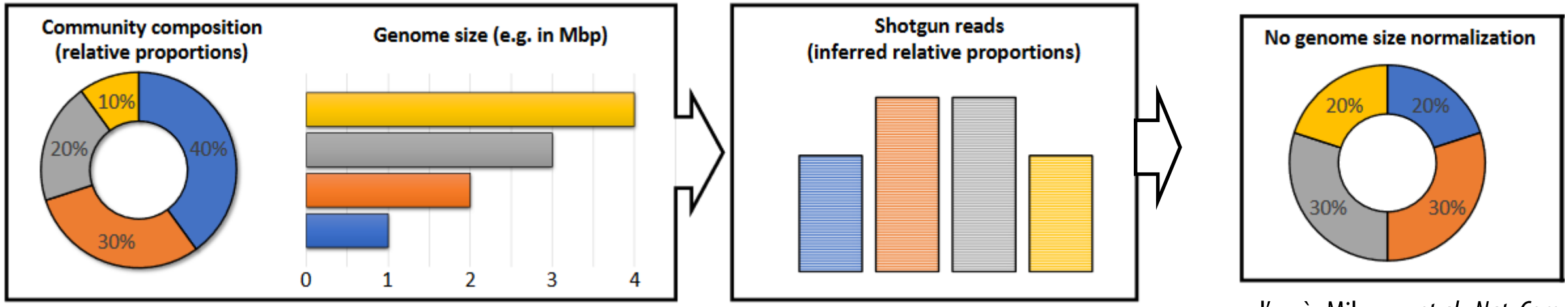
CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers

Rachid Ounit, Steve Wanamaker, Timothy J Close and Stefano Lonardi

- Assignation de chaque lecture à un taxon donné (règne→espèce)

Profilage taxonomique par classification exhaustive des lectures

Profils taxonomiques bruts biaisés par la taille des génomes



d'après Milanese *et al.*, *Nat. Com.* 2019

Normalisation possible a posteriori mais imparfaite si :

- Génomes de référence ont un contenu en gènes différent de celui des souches dans l'échantillon
- Séquences présentes en de multiples copies (plasmides, prophages)

Profilage taxonomique par utilisation de gènes marqueurs

- Alignement des reads contre des familles de gènes orthologues informatives sur le plan taxonomique
  - Gènes core
  - Nombre de copies constant
  - Peu sujet à des transferts horizontaux

- Approche précise, rapide, et peu consommatrice de RAM

Profilage taxonomique par utilisation de gènes marqueurs

❖ Utilisation de gènes marqueurs universels

Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences

Bo Liu[1,2], Theodore Gibbons[1,3], Mohammad Ghodsi[1,2], Todd Treangen[1], Mihai Pop[1,2,3*]

Microbial abundance, activity and population genomic profiling with mOTUs2

Alessio Milanese[1], Daniel R Mende[2], Lucas Paoli[3,4], Guillem Salazar[3], Hans-Joachim Ruscheweyh[3], Miguelangel Cuenca[3], Pascal Hingamp[5], Renato Alves[1,6], Paul I Costea[1], Luis Pedro Coelho[1], Thomas S.B. Schmidt[1], Alexandre Almeida[7,8], Alex L Mitchell[7], Robert D. Finn[7], Jaime Huerta-Cepas[1,9], Peer Bork[1,10,11,12], Georg Zeller[1] & Shinichi Sunagawa[3]

❖ Utilisation de gènes marqueurs clade-spécifiques

MetaPhlAn2 for enhanced metagenomic taxonomic profiling

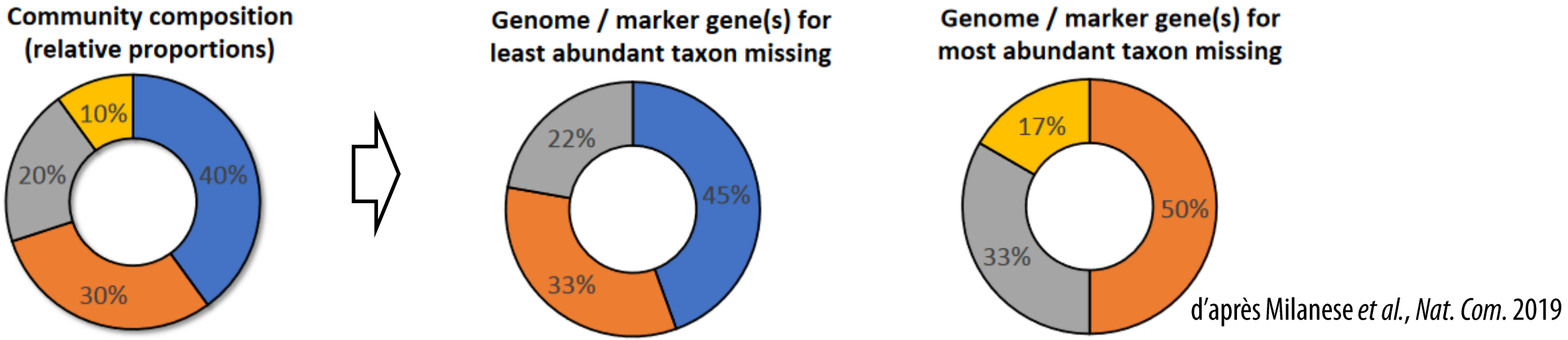
Data and text mining

MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data

Florian Plaza Oñate[1,2,*], Emmanuelle Le Chatelier[2], Mathieu Almeida[2], Alessandra C. L. Cervino[1], Franck Gauthier[2], Frédéric Magoulès[3], S. Dusko Ehrlich[2] and Matthieu Pichaud[1]

Profilage taxonomique par utilisation de gènes marqueurs

- Profils taxonomiques biaisés potentiellement biaisés par la représentativité de la base de données



d'après Milanese *et al.*, *Nat. Com.* 2019

- Nécessité de reconstruire les génomes des espèces manquantes

Types d'analyses de données métagénomiques

- Profilage taxonomique
  Quels sont les microorganismes présents et quelle est leur abondance?

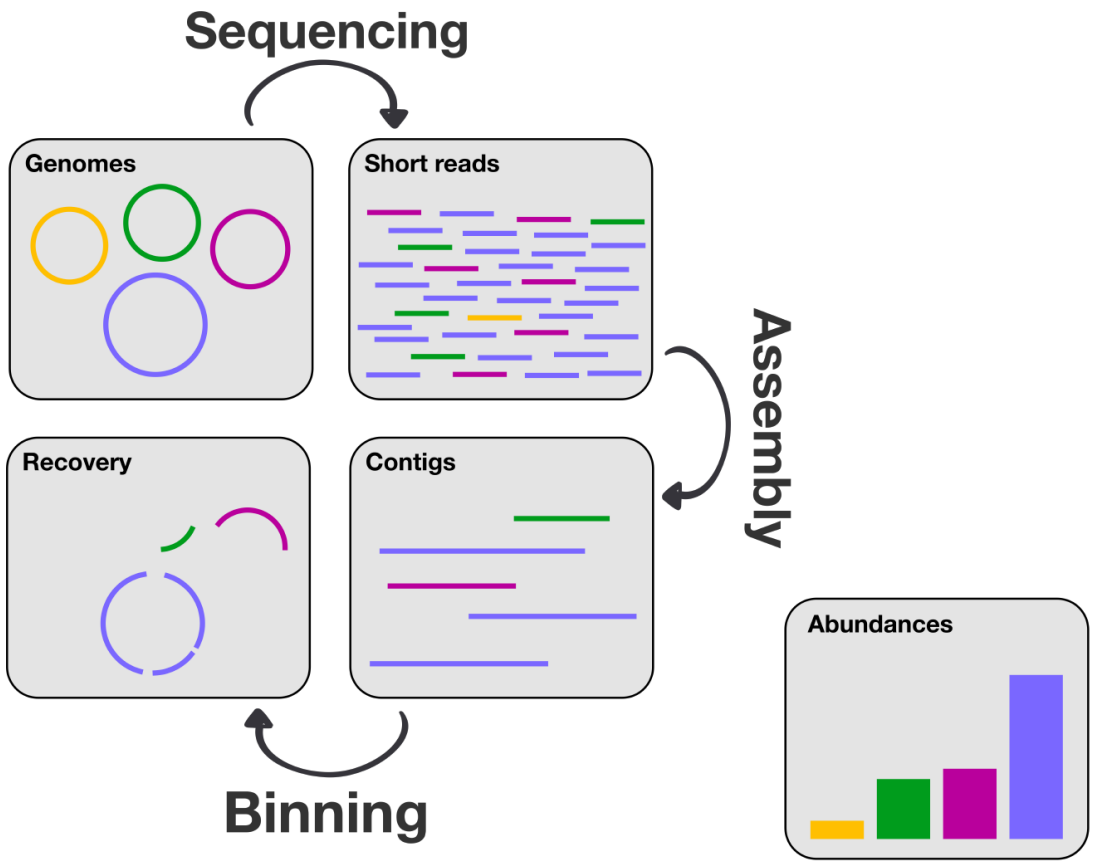
- Assemblage métagénomique

- Métagénomique au niveau souche

- Profilage fonctionnel
  Que peuvent t'il faire ?

- Métagénomique comparative
  Quelles sont les différences/similitudes entre plusieurs métagénomes?

Reconstitution de génomes par assemblage métagénomique



d'après A. Murat Eren - Intro to metagenomic binning

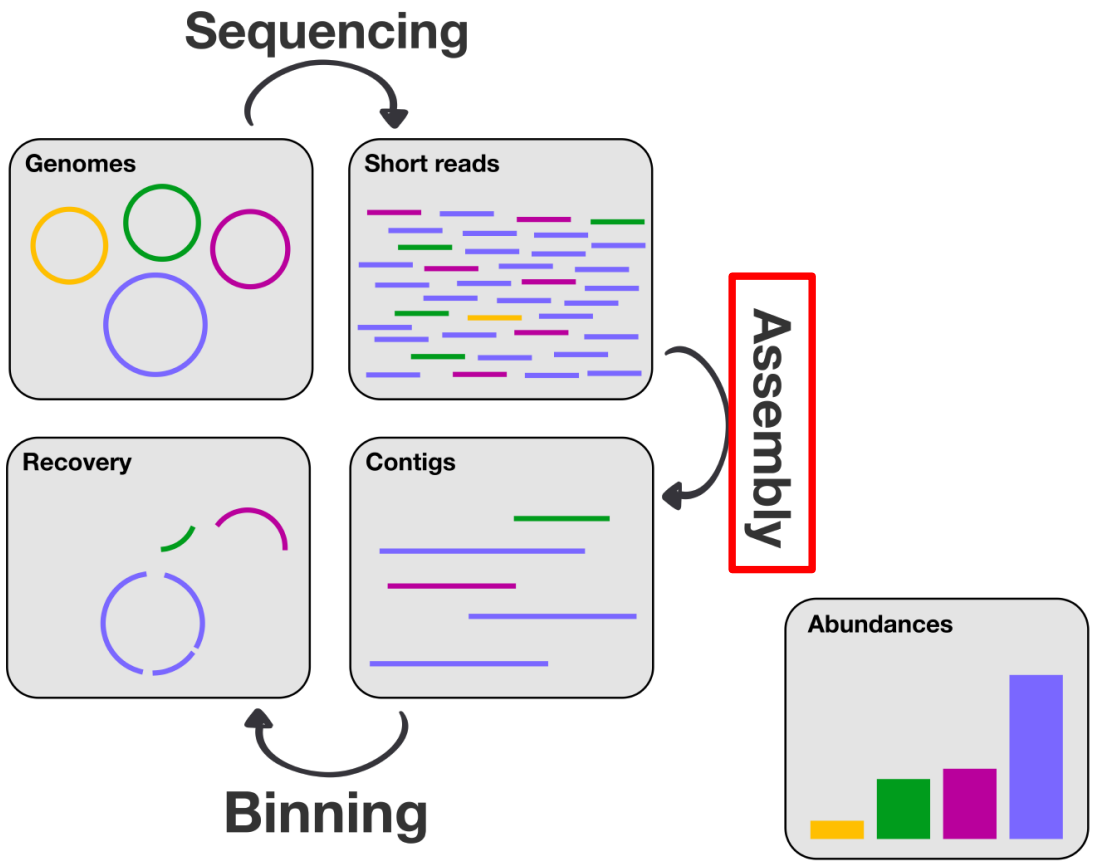
- Récupération de « draft genomes » de haute qualité pour les espèces les plus abondantes
- Pour un catalogue de génomes exhaustif:
  - Augmenter la profondeur de séquençage
  - Assemblage systématique de plusieurs échantillons

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

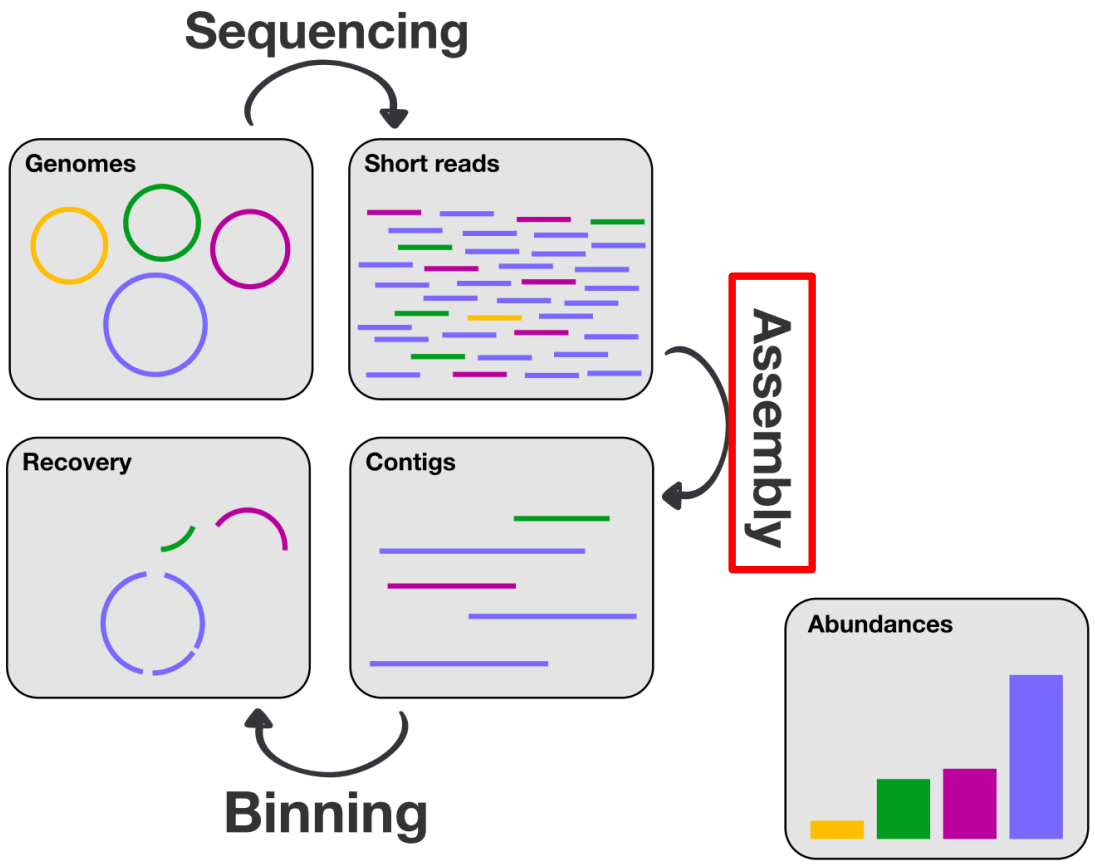
Reconstitution de génomes par assemblage métagénomique



d'après A. Murat Eren - Intro to metagenomic binning

- Conceptuellement similaire à l'assemblage génomique
  - Création d'un graphe de Bruijn
  - Recherche de chemin simples
  - Sortie: génomes incomplets fragmentés en contigs
- Difficultés propres à l'assemblage métagénomique
  - Coexistence d'espèces proches (surfragmentation + chimères)
  - Forte variabilité de la couverture de séquençage

Reconstitution de génomes par assemblage métagénomique



d'après A. Murat Eren - Intro to metagenomic binning

metaSPAdes: a new versatile metagenomic assembler

Sergey Nurk,[1,4] Dmitry Meleshko,[1,4] Anton Korobeynikov,[1,2] and Pavel A. Pevzner[1,3]

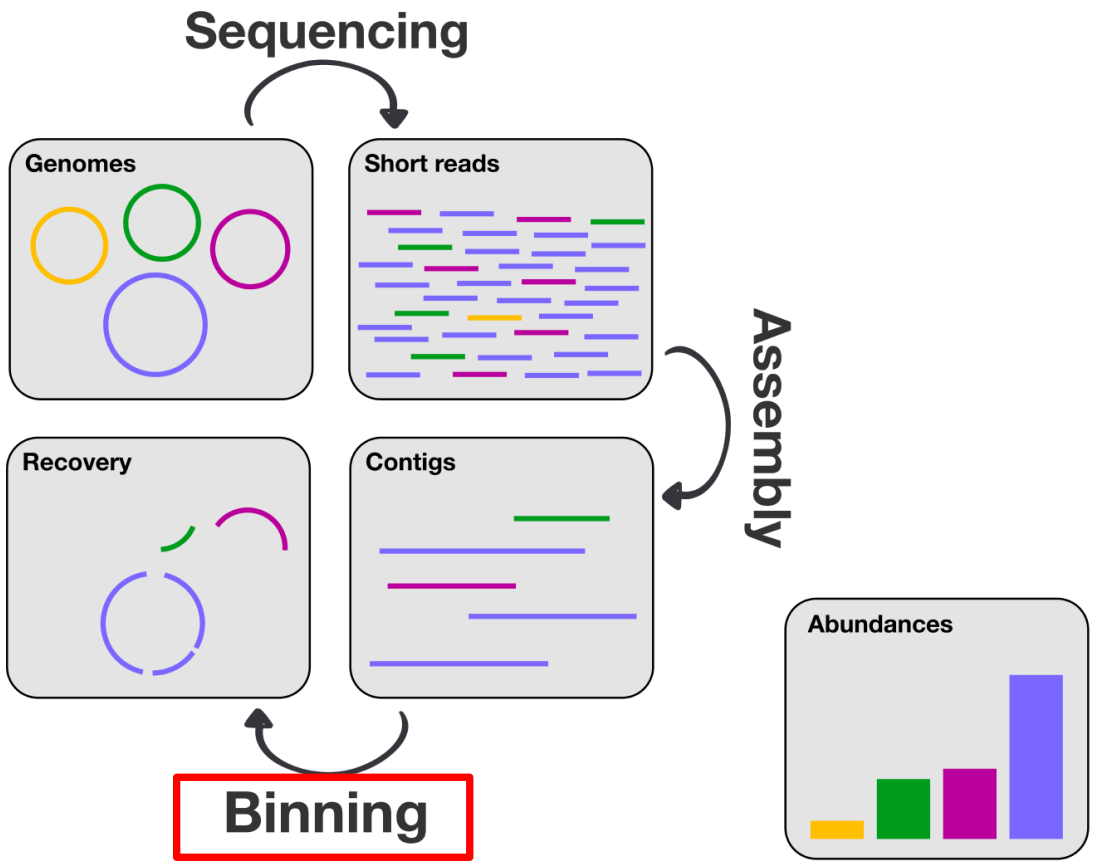
Génère les assemblages de meilleure qualité et les plus contigus

MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph FREE

Très rapide, consomme peu de mémoire
Accepte les reads non pairés

Voir van der Walt AJ et coll. BMC Genomics. 2017

Reconstitution de génomes par assemblage métagénomique



d'après A. Murat Eren - Intro to metagenomic binning

- Regroupement des contigs basé sur:
  - L'assignation taxonomique
  - La composition tétranucléotidique (TNF)
  - La couverture de séquençage

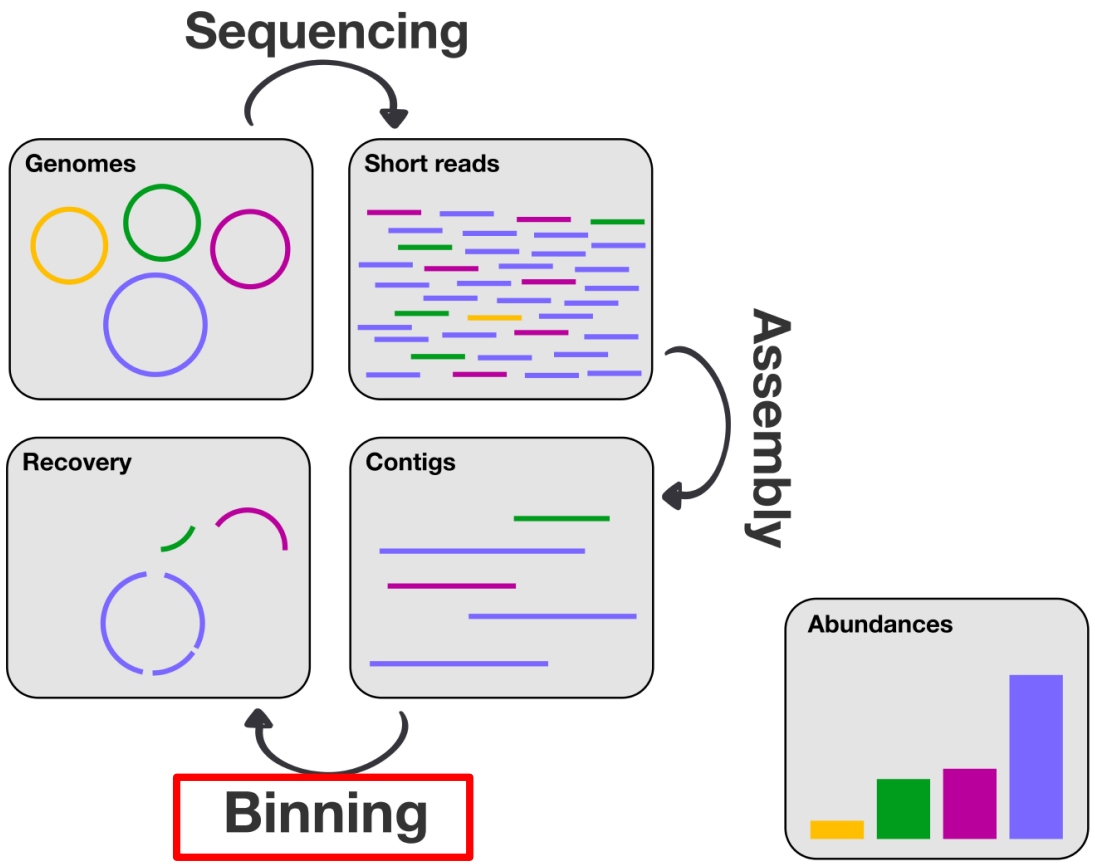
MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities

Dongwan D. Kang[1,2], Jeff Froula[1,2], Rob Egan[1,2] and Zhong Wang[1,2,3]

- Alternative: regroupement des gènes par co-abondance
  - Combine l'information de centaines d'échantillons
  - Plus sensible pour les espèces sous-dominantes

MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data

Reconstitution de génomes par assemblage métagénomique



d'après A. Murat Eren - Intro to metagenomic binning

- Validation des génomes basée sur le recensement de gènes marqueurs:
  - Complétion = tous les gènes détectés
  - Contamination = gènes dupliqués

CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks,[1] Michael Imelfort,[1] Connor T. Skennerton,[1] Philip Hugenholtz,[1,2] and Gene W. Tyson[1,3]

- Nettoyage manuel via une interface graphique

Anvi'o: an advanced analysis and visualization platform for 'omics data

Types d'analyses de données métagénomiques

- Profilage taxonomique
  Quels sont les microorganismes présents et quelle est leur abondance?

- Assemblage métagénomique

- Métagénomique au niveau souche

- Profilage fonctionnel
  Que peuvent t'il faire ?

- Métagénomique comparative
  Quelles sont les différences/similitudes entre plusieurs métagénomes?

Pourquoi une analyse niveau souche?

- Grande variabilité génétique et phénotypique intra-espèces :
  - Polymorphisme nucléotidique (SNPs) (Scholz et al., 2016)
  - Variabilité du nombre de copies des gènes (Greenblum et al., 2015)
  - Variabilité du contenu en gènes (Zhu et al., 2015)
    →notion de pangénome

- Résolution au-delà du niveau espèce nécessaire
  ex: épidémiologie, recherche clinique, génétique des populations

Epidemic Profile of Shiga-Toxin–Producing *Escherichia coli* O104:H4 Outbreak in Germany

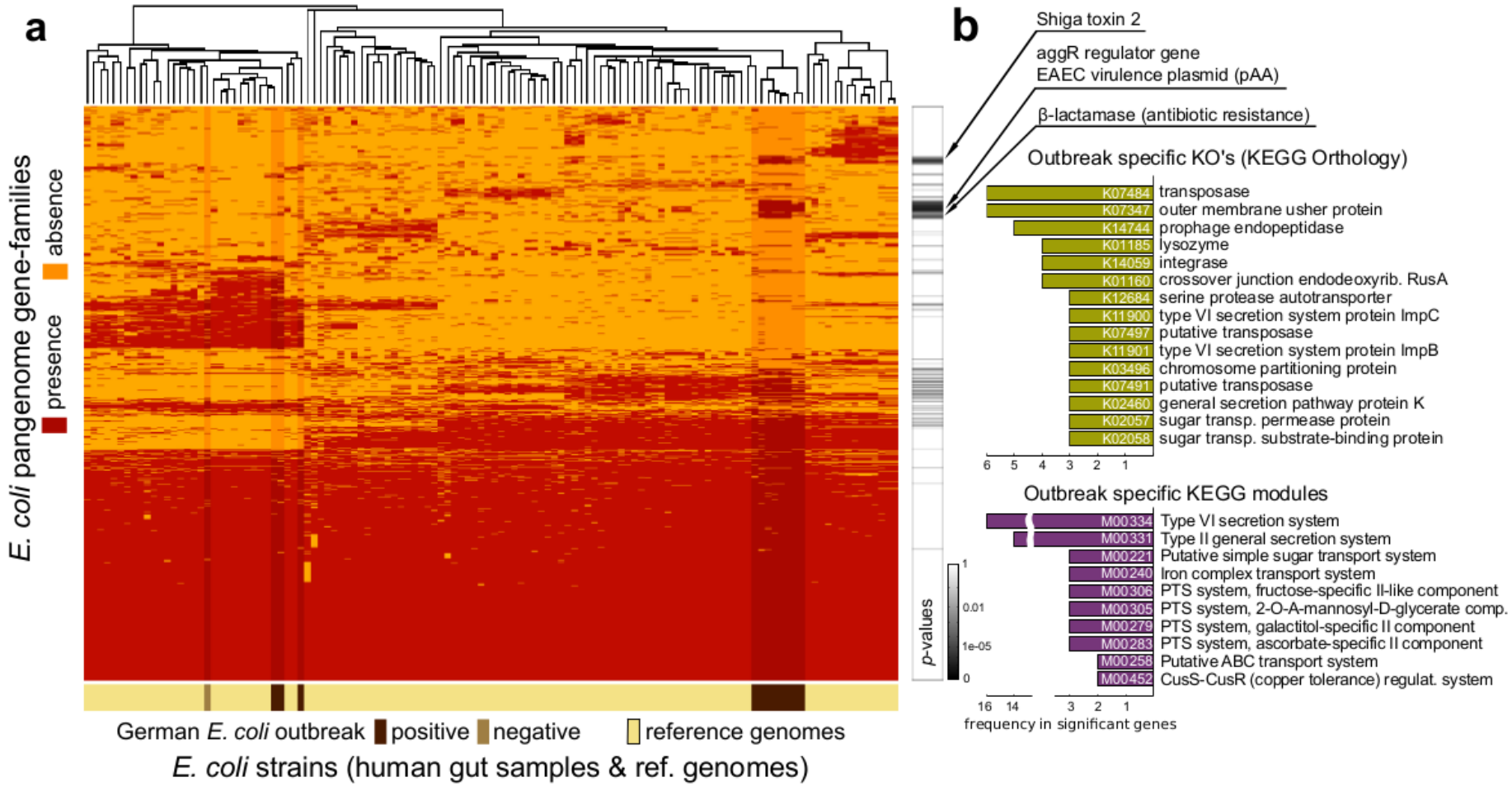
Distinct Genetic and Functional Traits of Human Intestinal *Prevotella copri* Strains Are Associated with Different Habitual Diets

Métagénomique au niveau souche
Approches par étude de la variabilité du contenu en gènes

- Caractérise les souches à partir de leurs gènes accessoires:
  - création de pangénomes à partir de génomes de référence
  - détermination des gènes présents/absents par alignement des reads
- Détermine le potentiel métabolique et pathogène de la souche considérée
- Limitations:
  - Couverture de 3x minimum
  - Considère uniquement la souche dominante d'un échantillon
  - Résultats dépendent du nombre de génomes disponibles

- PanPhlan (Scholz et al., 2016)
- MIDAS (Nayfach et al., 2016)

Métagénomique au niveau souche
Approches par étude de la variabilité du contenu en gènes



d'après Scholz et al., 2016

Métagénomique au niveau souche
Approches par étude de la variabilité nucléotidique

- Caractérisation par recensement des polymorphismes nucléotidiques
  - Sur gènes marqueurs espèce-spécifique: StrainPhlan (Truong *et al.*, 2017)
  - Sur l'intégralité d'un génome: MetaSVN (Costea *et al.*, 2017)
  - Inférence d'une signature par souche (haplotype)

- Construction d'un arbre phylogénétique
- Traçage longitudinal: vérifie la persistance temporelle d'une souche
- Traçage « vertical »: transfert d'un environnement à un autre

- Limitations:
  - Couverture de 4-5x minimum
  - Pas d'accès au potentiel fonctionnel de la souche

Types d'analyses de données métagénomiques

- Profilage taxonomique
  Quels sont les microorganismes présents et quelle est leur abondance?

- Assemblage métagénomique

- Métagénomique au niveau souche

- Profilage fonctionnel
  Que peuvent t'il faire ?

- Métagénomique comparative
  Quelles sont les différences/similitudes entre plusieurs métagénomes?

Estimation du potentiel fonctionnel d'une communauté

- Approche sans assemblage (HumaNN, MEGAN)
  - Traduction des reads dans les 6 phases de lectures
  - Alignement contre une base de protéines annotées (KEGG, UniProt, eggNOG etc.)
  - MEGAN (Huson et coll. 2016), HumaNN2 (Franzosa et coll., 2019)

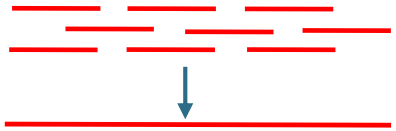
Fast and sensitive protein alignment using DIAMOND

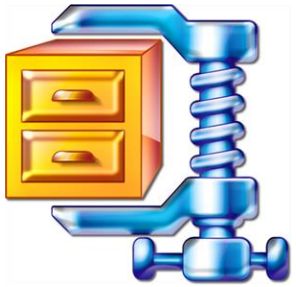
Benjamin Buchfink, Chao Xie & Daniel H Huson

20 000x plus rapide que blastx

Estimation du potentiel fonctionnel d'une communauté

- Approches par assemblage
  - Création d'un catalogue de gènes par assemblage métagénomique
  - Annotation du catalogue (e.g. eggNOG-mapper)
  - Alignement des reads sur le catalogue de gènes annotés





Séquençage → Assemblage → Prédiction des gènes → Suppression de la redondance → Annotation du catalogue

Estimation du potentiel fonctionnel d'une communauté

- Agrégation des comptages
  - Approche décloisonnée vs cloisonnée (espèce par espèce)

- Limitations:
  - Nombreuses familles de protéines non annotées (gènes accessoires)
  - Biais de représentativité: certaines clades mieux annotées
  - Annotations automatiques incorrectes
  - **Potentiel** fonctionnel → Quelles sont les fonctions réellement actives?

Références bibliographiques

- Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci*. 2014
- Nayfach S et Pollard KS. Toward Accurate and Quantitative Comparative Metagenomics. *Cell*. 2016.
- Quince C et coll. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*. 2017
- Plaza Oñate F. Reconstitution de pan-génomes microbiens par séquençage métagénomique aléatoire. *theses.fr*. 2018

Merci pour votre attention!