Stratégies de reconstruction de génomes microbiens à partir de métagénomes

Kévin Gravouil [1,2,3], Corentin Hochart [2], Bérénice Batut [1], Clémence Defois [1], Cyrielle Gasc [1], Pierre Peyret [1], Didier Debroas [2], Marie Pailloux [3], Eric Peyretaillade [1]

[1] EA 4678 CIDAM ; [2] UMR CNRS 6023 LMGE ; [3] UMR CNRS 6158 LIMOS

Contacts : kevin.gravouil@udamail.fr ; didier.debroas@univ-bpclermont.fr ; pailloux@isima.fr ; eric.peyretaillade@udamail.fr

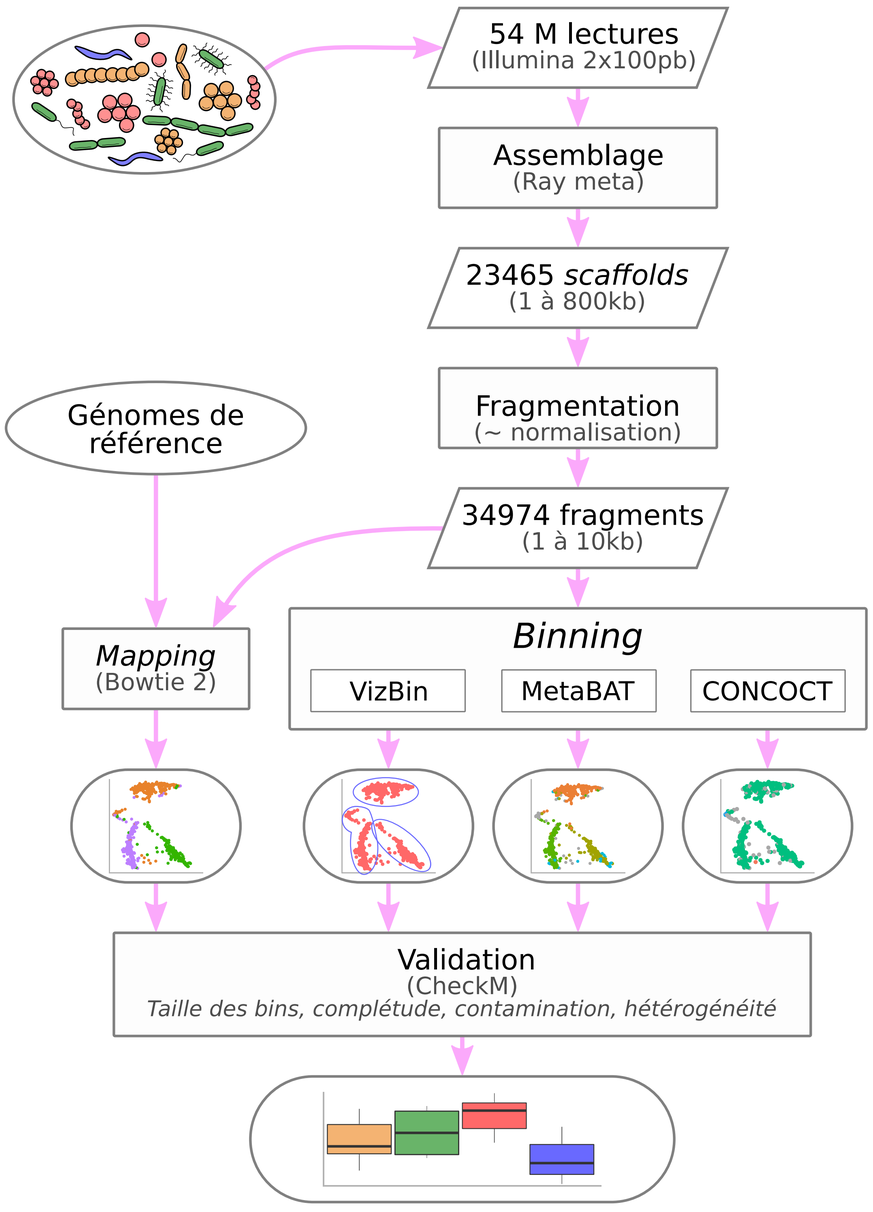
Introduction

Le séquençage à haut-débit permet d'accéder à l'extraordinaire diversité des micro-organismes notamment par des **approches métagénomiques**. Néanmoins, la compréhension globale d'un écosystème nécessite de relier efficacement **structure et fonctions**. De ce fait, la **reconstruction de génomes individuels** à partir de métagénomes devient une approche nécessaire pour remplir cet objectif.

La diversité génomique n'est pas connue *a priori* et dépend fortement de l'environnement étudié. Il convient donc d'employer des méthodes **non ciblées** et ***de novo*** pour explorer ces environnements.

Malgré la grande diversité de micro-organismes, le ***binning*** est une approche qui rend possible la reconstruction de génomes [1]. Cette approche repose sur le fait que deux **séquences similaires** en terme de composition appartiendraient à un **même génome**. Plusieurs stratégies ont depuis été proposées mais **aucun consensus** n'a pu être dégagé.

Matériel et méthodes



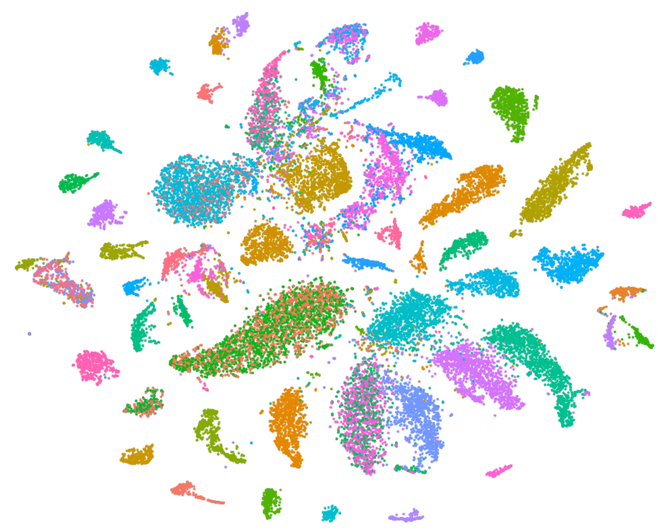
Afin d'évaluer la pertinence des méthodes de *binning* existantes, différents outils ont été testés sur un jeu de données **simulant un métagénome** composé de **64 micro-organismes** dont les génomes sont disponibles (SRR606249). Les bins de références sont obtenus par alignement (*mapping*) sur ces génomes.

Différents logiciels de *binning* ont été testés : (i) **VizBin** [2] qui exploite la composition nucléotidique ; (ii) **MetaBAT** [3] et (iii) **CONCOCT** [4] qui utilisent à la fois la composition et les différences de couverture des séquences. Ces trois approches diffèrent également par leurs méthodes de *clustering*. VizBin propose à l'utilisateur de définir les ***bins* manuellement** ; MetaBAT utilise un algorithme des **k-medoids** modifié ; CONCOCT utilise un **modèle de mélanges gaussiens** complétée d'une **approche bayesienne**.

La validation des *bins* avec **CheckM** [5] consiste à rechercher des **gènes-marqueurs uniques** au sein d'une lignée phylogénétique, évaluant ainsi : (i) la « **complétude** » (nombre de marqueurs au sein d'un *bin* par rapport à l'attendu) ; (ii) la **contamination** (nombre de marqueurs en plusieurs copies) et (iii) l'**hétérogénéité** de souche.

Etude comparative des outils de *binning*

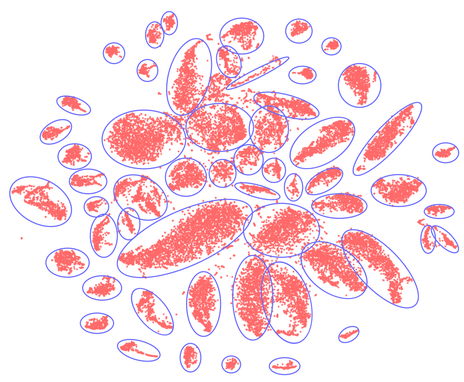
• ***Binning*** - Les 34974 fragments de séquences (de 1 à 10kb) issus de l'assemblage *de novo* des 54 M de lectures ont été alignés aux génomes de référence ou regroupés par *binning*.



Référence
- *Mapping* sur les génomes de référence
» **33507 séquences** alignées (95.8%)
» **64 organismes** connus

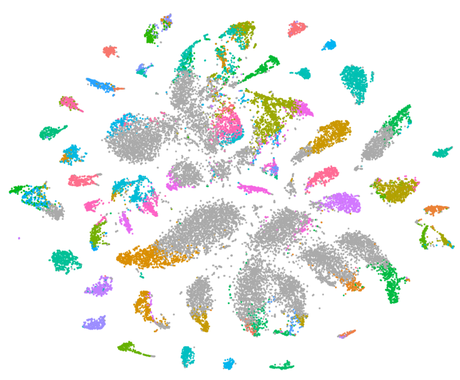
Légende
Chaque point représente un fragment de séquence.
Chaque couleur représente un *bin*.
Les points gris représentent les fragments non assignés.

VizBin



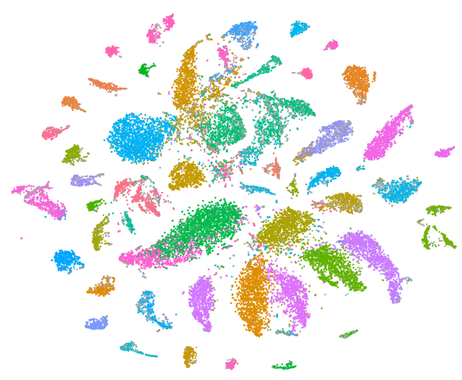
- Dénombrement des hexamers
- *Clustering* manuel
» **34559 séquences** *binnées* (98.8%)
» **55** *bins*

MetaBAT



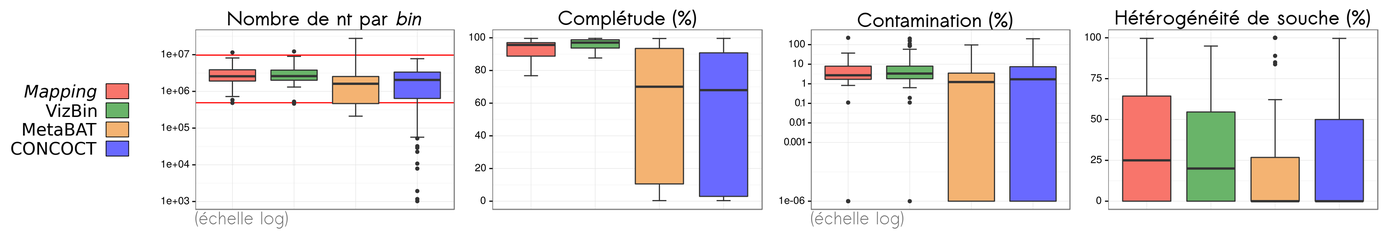
- Dénombrement des tetramers + couverture
- *Clustering* par k-medoids
» **19111 séquences** *binnées* (54.6%)
» **84** *bins*

CONCOCT

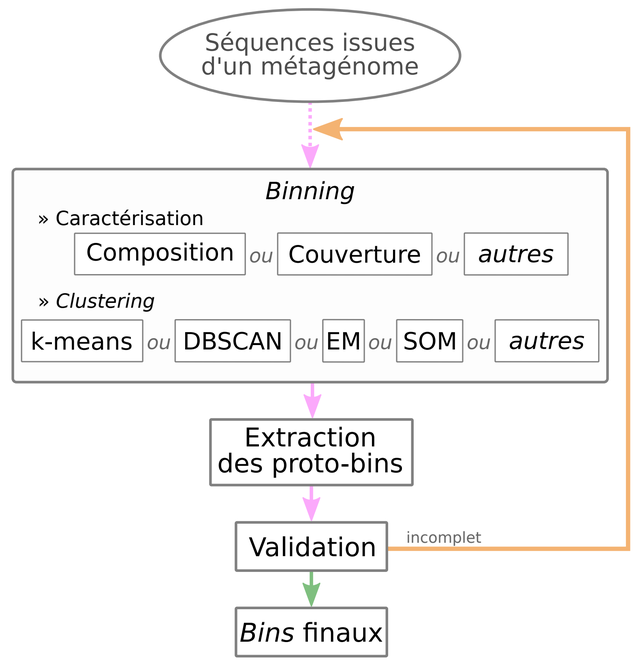


- Dénombrement des tetramers + couverture
- *Clustering* par mélange gaussien
» **31817 séquences** *binnées* (87.1%)
» **68** *bins*

• **Validation** - Des marqueurs sont recherchés pour chaque *bin* afin d'évaluer leur taux de complétude, de contamination et d'hétérogénéité de souche.



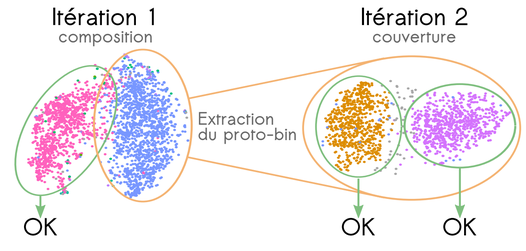
Stratégie alternative



Les méthodes de *binning* testées ne permettent pas toujours de dissocier correctement deux génomes d'espèces phylogénétiquement proches.

Pour pallier ce problème, il est possible de réaliser une **approche itérative**. Un premier *binning* permet d'**extraire des prototypes de *bins*** (ou « proto-bins »), et de les **valider individuellement**. Lorsque les critères de validation d'un proto-bin ne sont pas remplis (fig. 4, symbolisé par la flèche jaune), une autre méthode de binning est appliquée sur ce proto-bin.

• **Exemple** - Deux bins confondus deviennent séparables.



Conclusion et perspectives

» **Pas de consensus** en matière de reconstruction de génomes à partir de métagénomes
» Succession de **plusieurs méthodes** pour de meilleurs résultats
» Utilisation de **données de référence** pour la **validation** (si l'environnement étudié le permet)
» **Caractérisations alternatives** des séquences (par exemple, avec les *spaced-seed*)
» Utilisation de différentes méthodes de ***clustering***
» **Ré-analyse** des données existantes (e.g : microbiote humain)
» Utilisation de méthodes issues du **Big Data**

Références

[1] Sangwan *et al.*, Microbiome, 2016, DOI: 10.1186/s40168-016-0154-5
[2] Laczny *et al.*, Microbiome, 2015, DOI: 10.1186/s40168-014-0066-1
[3] Kang *et al.*, PeerJ, 2015, DOI: 10.7717/peerj.1165
[4] Alneberg *et al.*, Nature Methods, 2014, DOI: 10.1038/nmeth.3103
[5] Parks *et al.*, Genome Research, 2014, DOI: 10.1101/gr.186072.114

frama.link/kg-j16