Assemblage de génomes bactériens séquencés par NGS - Comparaison d'outils et choix de paramètres

Fabien Melchiore, Cyprien Guerin Guérin, Pierre P. Nicolas, Valentin Loux

Assemblage de génomes bactériens séquencés par NGS
Comparaison d'outils et choix de paramètres

Fabien Melchiore, Cyprien Guérin, Pierre Nicolas, Valentin Loux
{ fabien.melchiore, cyprien.guerin, pierre.nicolas, valentin.loux } @ jouy.inra.fr

MIG – Mathématique, Informatique et Génome, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex

Introduction

- Les **Méthodes de Séquençage de Nouvelle Génération** (**NGS**) permettent de séquencer rapidement et à faible coût de nouvelles souches.
  → multiplication des outils d'assemblage de NGS (très nombreuses lectures courtes) + évolution rapide des techniques de séquençage.
  → nécessité de comparer les diverses stratégies d'assemblage sur un jeu de données commun.

- **Objectifs de l'étude :**
  → comparer deux stratégies d'assemblage des régions spécifiques.
  → évaluer diverses méthodes de nettoyage des lectures (l'assembleur de novo choisi ne prenant pas en compte la qualité des données).
  → déterminer la couverture minimale suffisante lors d'études de détection de nouveaux gènes.

Données et Logiciels

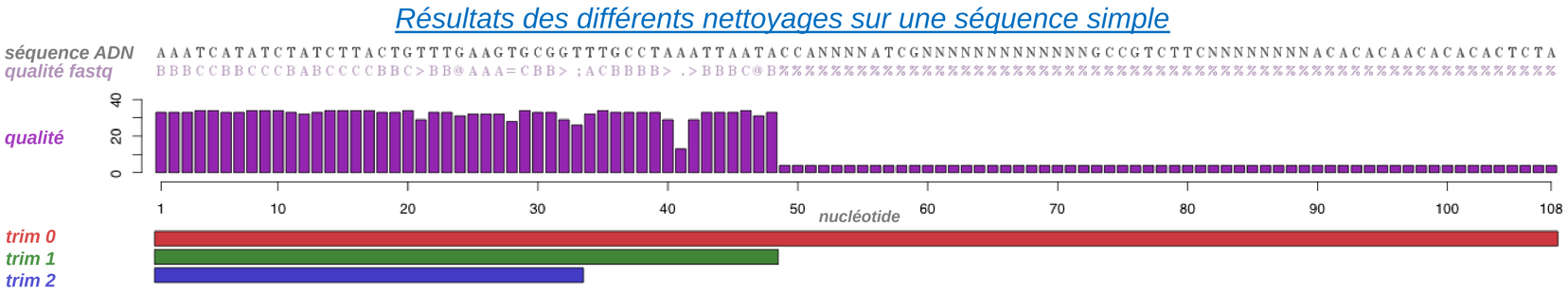
- **Séquençage Solexa/Illumina (single reads)**
  → six souches de la bactérie *Flavobacterium psychrophilum.*
- Chaque jeu de données = **10 millions** de lectures de **100 bp** (ANR FLAVOPHYLOGENOMICS, coordinateur E. Duchaud)
  → couverture maximale théorique > **200X**.
- Les **séquences complètes** et **annotées** de deux des six souches sont disponibles (souches THC et JIP).
- L'évaluation des divers assemblages est réalisée via une comparaison avec les gènes connus de ces deux souches.
- Seuls sont présentés ici les résultats obtenus pour la souche THC v. refTHC171109.

- **Assemblage de novo**
  → Velvet v. 0.7.58
  → VelvetOptimiser v. 2.1.4.+
- **Mapping**
  → MAQ v. 0.7.1
- **Similarité de séquences**
  → Yass v. 1.14
- **Détection de gènes**
  → Show v. 20061029

Nettoyage des données

- Afin de réduire les temps de calcul et d'améliorer la qualité de l'assemblage, il semble nécessaire de nettoyer les lectures avant de les assembler.
- Il est également essentiel de disposer d'un estimateur pour évaluer la qualité de l'assemblage.

Résultats des différents nettoyages sur une séquence simple

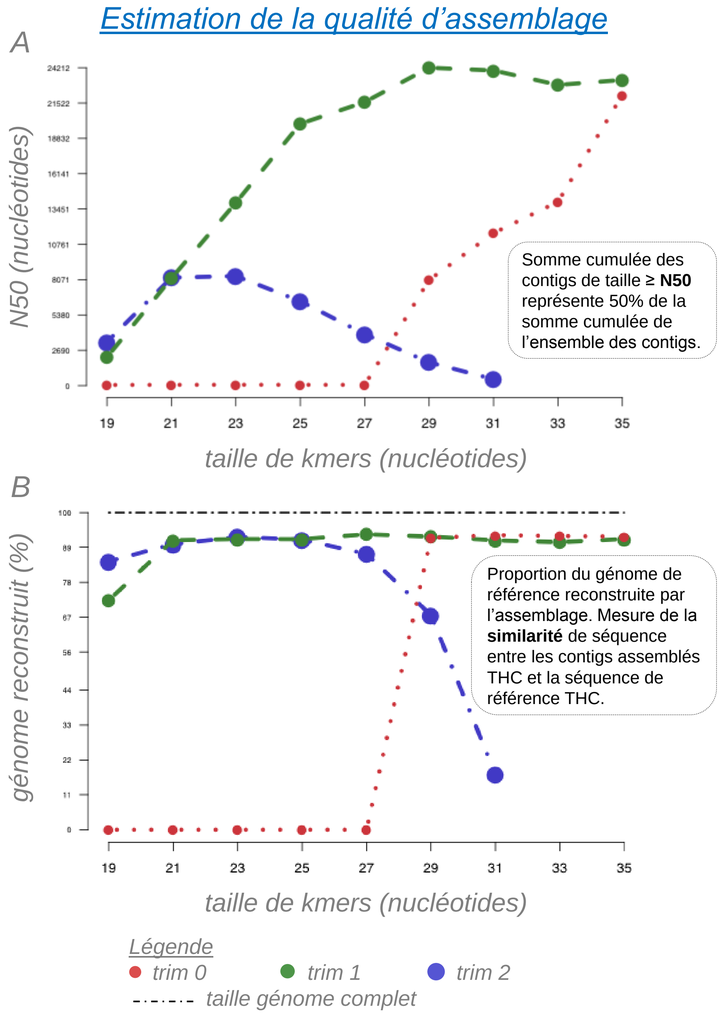


Préparation des jeux de données :

- ♦ ***trim 0*** : aucun nettoyage des lectures.
- ♦ ***trim 1*** : nettoyage 'adaptatif'.
  Pour chaque lecture, retrait première/dernière bases tant que qualité < seuil (10)
  Suppression lectures ayant au moins 1 'N' (base non déterminée)
  Suppression lectures dont longueur < seuil (20)
  Suppression lectures dont qualité moyenne < seuil (20)
- ♦ ***trim 2*** : nettoyage 'global'.
  Retrait dernière base de toutes les lectures tant que qualité moyenne < seuil (27)
  Suppression lectures ayant au moins 1 'N' (base non déterminée)
  Suppression lectures dont qualité moyenne < seuil (20)

	Nombre reads	Taille moyenne reads (nt)	Qualité moyenne reads	Nombre total 'N'	Nombre reads avec ≥ 1N	Taille fichiers graphes (MB)	Temps execution 19≤k≤35 (h)
Trim 0	11 166 909	108	17.29	4 068 026	1 217 536	De 509 à 8700	33 trim = 0
Trim 1	10 758 045	55	28.97	0	0	De 156 à 786	18.5 trim = 9.5
Trim 2	10 862 792	33	31.19	0	0	De 132 à 170	10.5 trim = 6.5

Estimation de la qualité d'assemblage



- ► Le nettoyage des lectures est nécessaire et efficace. Le trim 'adaptatif' fournit de meilleurs assemblages.
- ► Le N50 représente un bon estimateur de la qualité d'assemblage : plus il est élevé, plus la portion de génome reconstruit est grande.

Reconstruction des régions spécifiques

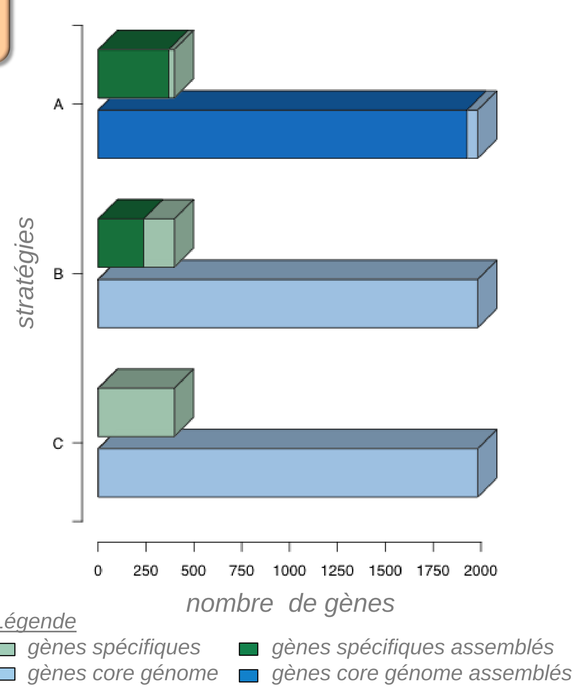
- Comparaison de stratégies d'assemblage des lectures après nettoyage de celles-ci.

Stratégies d'assemblage :

- ♦ *A* : assemblage de novo de l'ensemble des lectures THC → contigs représentant l'ensemble du génome THC séquencé (incluant les régions spécifiques THC).
- ♦ *B* : assemblage de novo uniquement des lectures THC rejetées lors du mapping sur référence JIP → contigs représentant les régions spécifiques THC d'intérêt.
- ♦ *C* : assemblage de novo uniquement des lectures THC rejetées lors du mapping sur référence THC → témoin négatif.

Le nombre de gènes reconstruits après assemblage est déterminé par mesure de la similarité de séquence entre les contigs assemblés THC et les gènes connus des références THC et JIP.

Nombre de gènes reconstruits en fonction de la stratégie choisie



- ► L'assemblage de novo de la totalité des lectures est la stratégie la plus efficace pour la reconstruction des régions spécifiques.

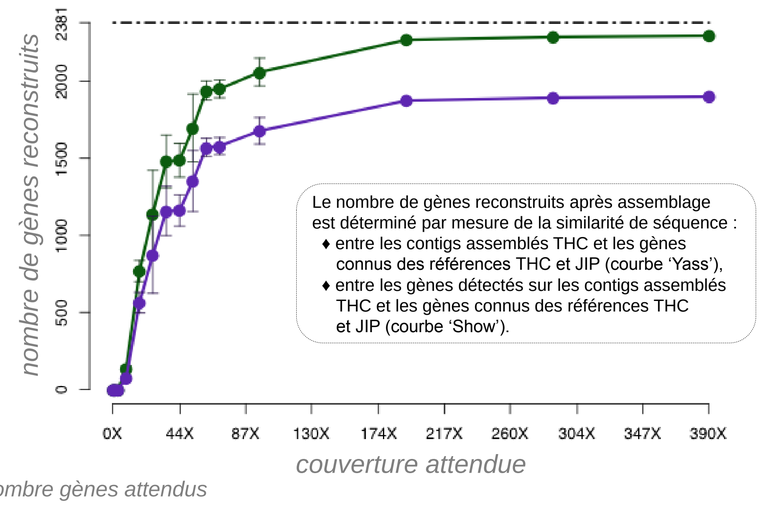
Reconstruction et Couverture

- Étude de l'impact du nombre de lectures (ou couverture du séquençage) sur la reconstruction des gènes.

Protocole :

- ♦ Pour chaque couverture,
  → extraction aléatoire de 10 sous-jeux de lectures.
- ♦ Pour chacun des jeux,
  → trim 'adaptatif' + assemblage de toutes les lectures.
  → recherche de gènes connus sur les contigs.
  → recherche de gènes connus parmi ceux détectés sur les contigs.

Nombre de gènes détectés dans l'assemblage en fonction de la couverture de séquençage



- ► Les résultats sont reproductibles.
- ► La détection des gènes est peu performante.
- ► Bonne reconstruction (≥85%) des gènes à partir d'une couverture d'environ 100X .

Conclusions et Perspectives

- ► L'assemblage de l'ensemble des lectures permet une meilleure reconstruction des régions spécifiques que l'assemblage des lectures 'spécifiques' uniquement.
- ► Le N50 s'avère être un bon estimateur pour l'assemblage de novo.
- ► Pour ce type de séquençage, la qualité de la reconstruction des gènes est suffisante (≥85%) à partir d'une couverture d'environ 100X.
- ► Des résultats similaires sont obtenus pour la souche JIP v. refJIP121009.

- Évaluation d'autres types de données (librairies et séquenceurs différents).
- Évaluation d'autres assembleurs (SOAP de novo, Mira, etc.).
- Évaluation de divers détecteurs de gènes.

Ce travail a été financé par le projet ANR BIOWIC.