Gestion informatique des données de séquençage

William Amory
M1 BI-IPFB Université de Paris

24/01/2022

Section 1

CEA - Genoscope

CEA - Genoscope

CEA (Commissariat à l'énergie atomique et aux énergies)

- créé le 18 octobre 1945 par Charles de Gaulle
- 20 000 Salariés
- 4 directions opérationnelles et 9 directions fonctionnelles

Genoscope (Centre National de Séquençage)

- 250 salariés
- Créé en 1996
  - Participation **projet Génome humain** (Séquençage du chromosome 14 humain)
  - Développer programmes de génomiques en France
  - Plus grand centre de séquençage français
  - ajouter

Organigrame CEA - Genoscope - LBGB

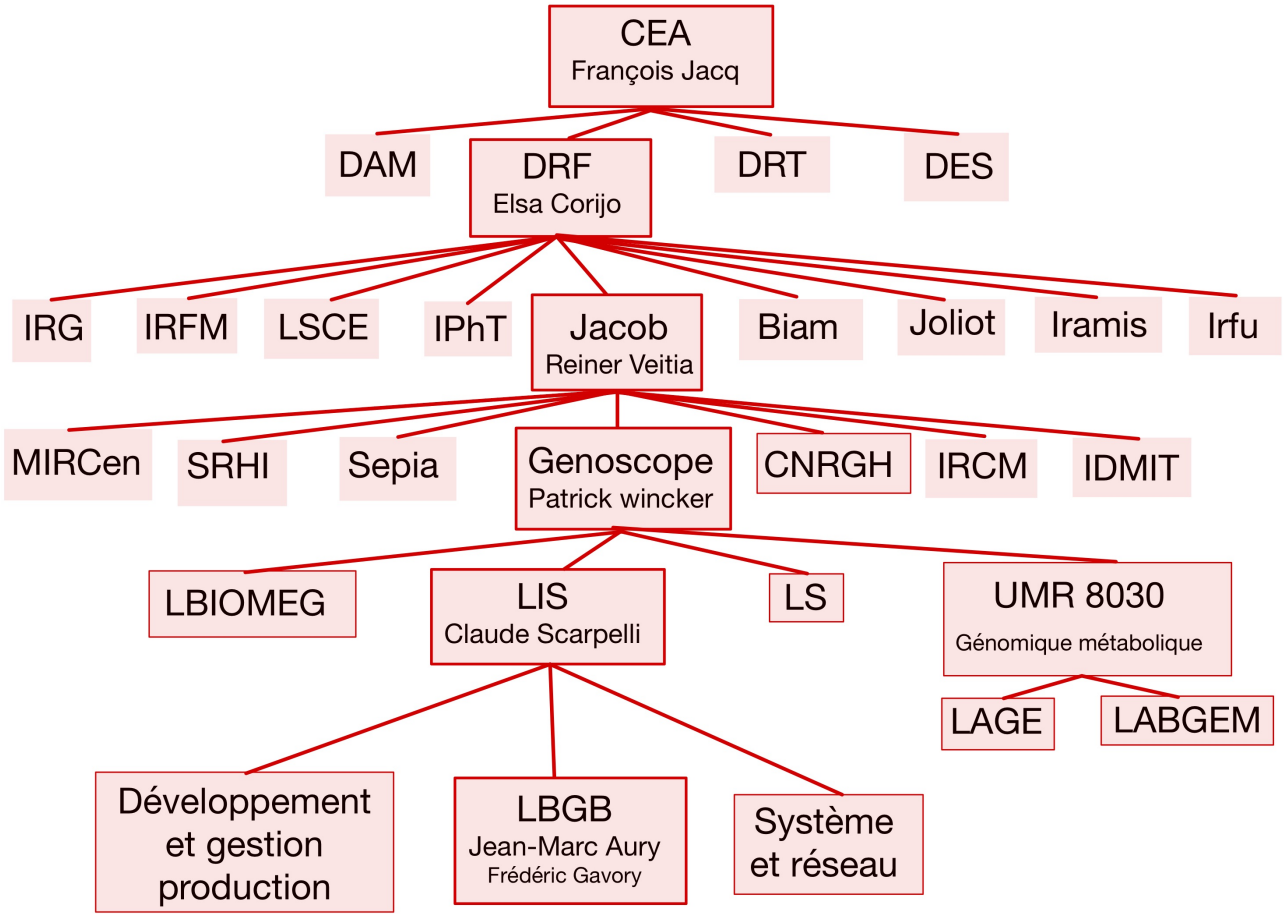

Figure 1: Organigramme situant l'équipe du *Laboratoire de Bioinformatique pour la Génomique et la Biodiversité (LBGB)* au sein du genoscope et du CEA

Section 2

Contexte

LBGB (Laboratoire de Bioinformatique pour la Génomique et la Biodiversité)

missions

- Veille technologique
- Contrôle qualité
- Assemblage
- Annotation
- Visualisation

Plusieurs groupes de travail

- Production
- Annotation
- Assemblage
- Evaluation des technologies de séquençage

LBGB (Production)

Missions

- Veille technologique
- Evaluation de nouveaux outils
- developper, tester et maintenir les codes
- Répondre aux besoins des équipes de recherche et de production
- Mise en place de pipeline automatisés
  - génération des fichiers de séquences
  - Contrôle qualité
  - Analyses biologiques

LBGB - Workflow NGS

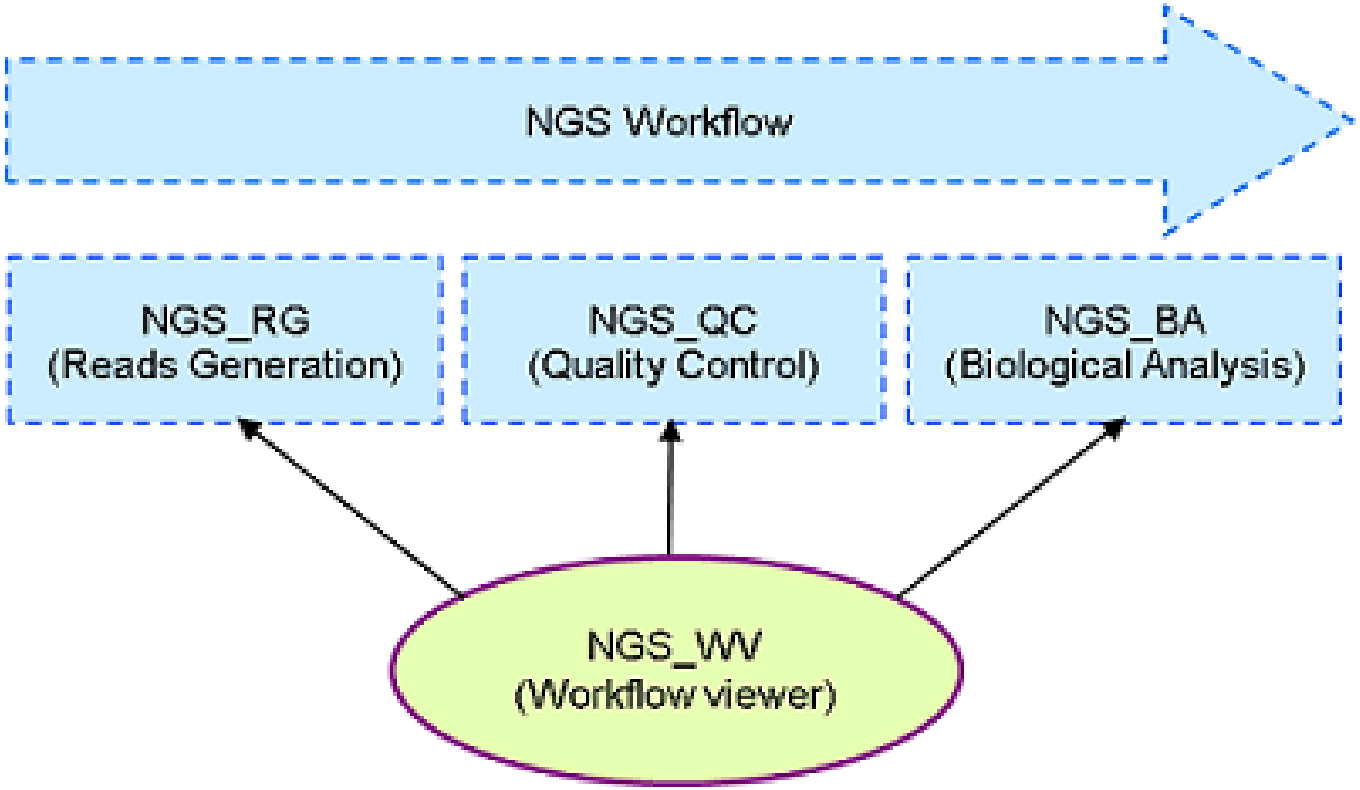

Figure 2: Workflow de génération, de controle qualité et d'analyse biologique des FASTQ

LBGB - MGI

Arrivé de séquenceurs MGI

- 2 DNBSEQ-G400
- 1 DNBSEQ-T7

https://en.mgi-tech.com/products/

La technologie MGI

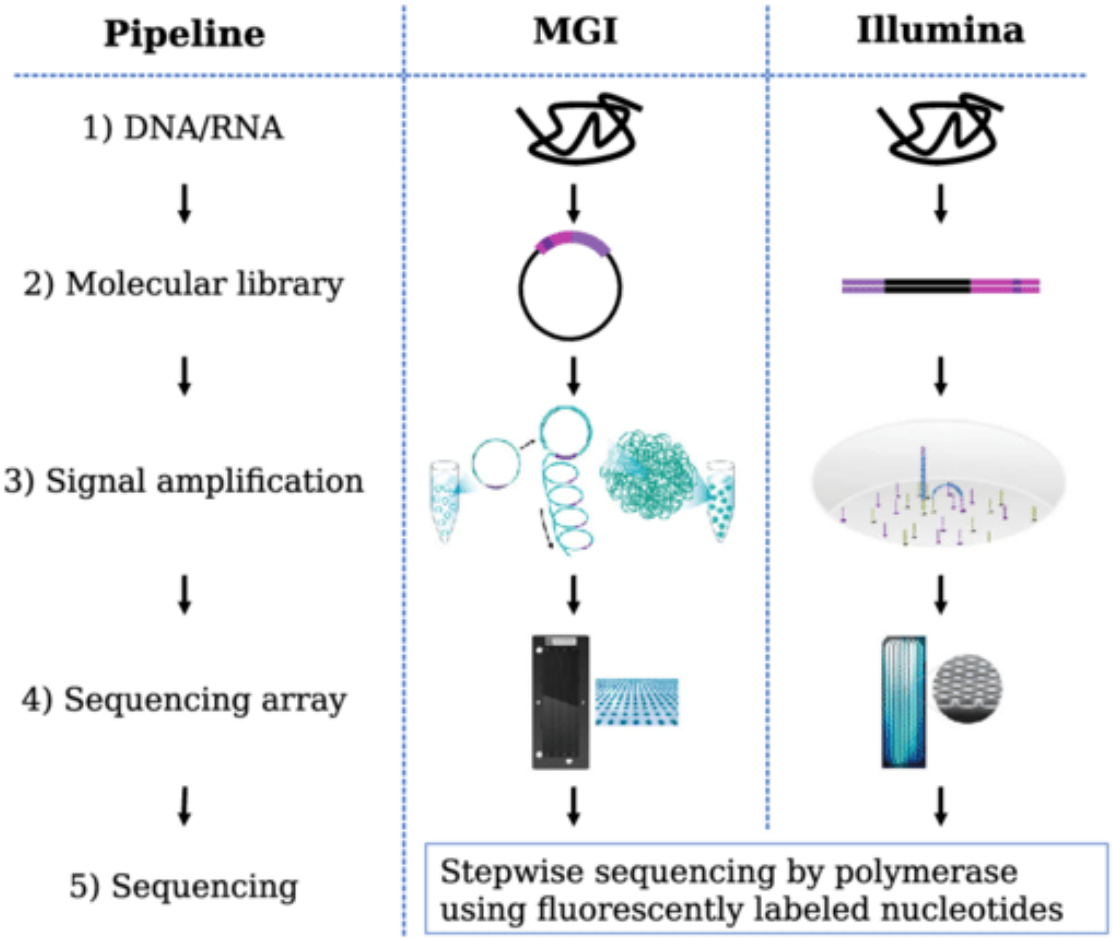

Figure 3: Différences entre Illumina et MGI de technologie NGL

La technologie MGI

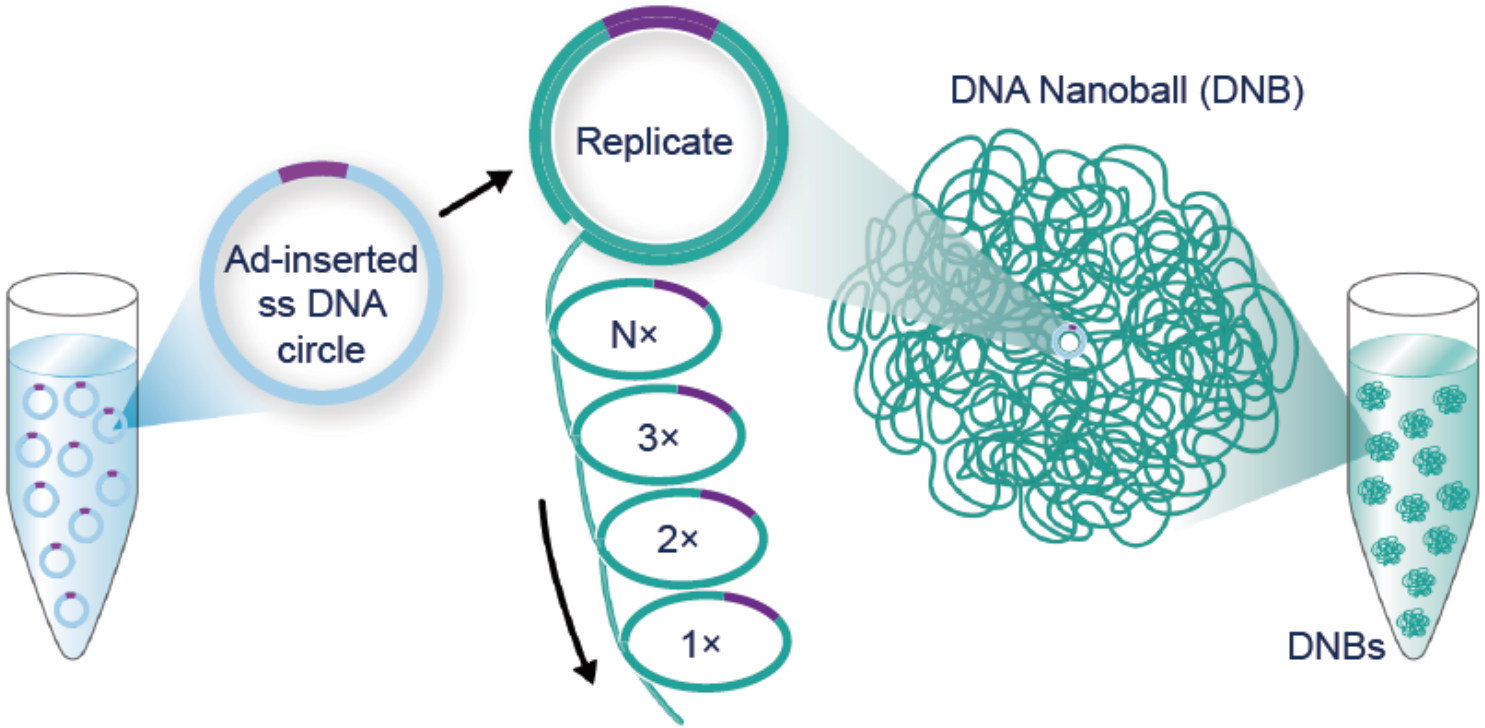

Figure 4: Schéma techno MGI

Section 3

Objectifs

Développement d'un pipeline automatique pour MGI

Objectifs du pipeline

- développement NGS-RG et NGS-QC pour MGI
- Distribution des FASTQ par projets
- Trie des FASTQ par echantillons, technologies et runs
- Mise à jour de la base de données de références (NGL)
  - création des entrées runs et readset
  - stockage des métriques et analyses correspondants
- Nettoyage des FASTQ générés
- Analyses des FASTQ générés

Développement d'un pipeline automatique pour MGI

Comment ?

- Déterminer les outils et methodes nécessaires
  - utisation d'outils et méthodes existant pour Illumina ?
  - utilisation de nouveaux outils et méthodes ?
- Ecriture du pipeline
  - déterminer de l'ordre d'utilisation des outils et méthodes

Evaluation et codage d'outils

ajouter les autres objectifs de ma mission

Apprentissage du Perl

Pourquoi ?

- Raison historique du laboratoire
- Toutes les librairies et modules utilisés sont en Perl
- Worflow d'Illumina écrit en Perl

Réalisation

- Programme effectuant des analyses statistiques élémentaires
  - compter le taux de GC
  - moyene de la qualité de chaque read
  - ect . . .
- Utilisation des modules utilisé dans le workflow d'illumina

Test de 2 software de génération de FASTQ (bcl2fastq et bcl-convert)

Permet la génération des FASTQ et de réaliser le démultiplexage
Comparaison des performances - Recherche

bcl2fastq vs bcl-convert (Temps total)

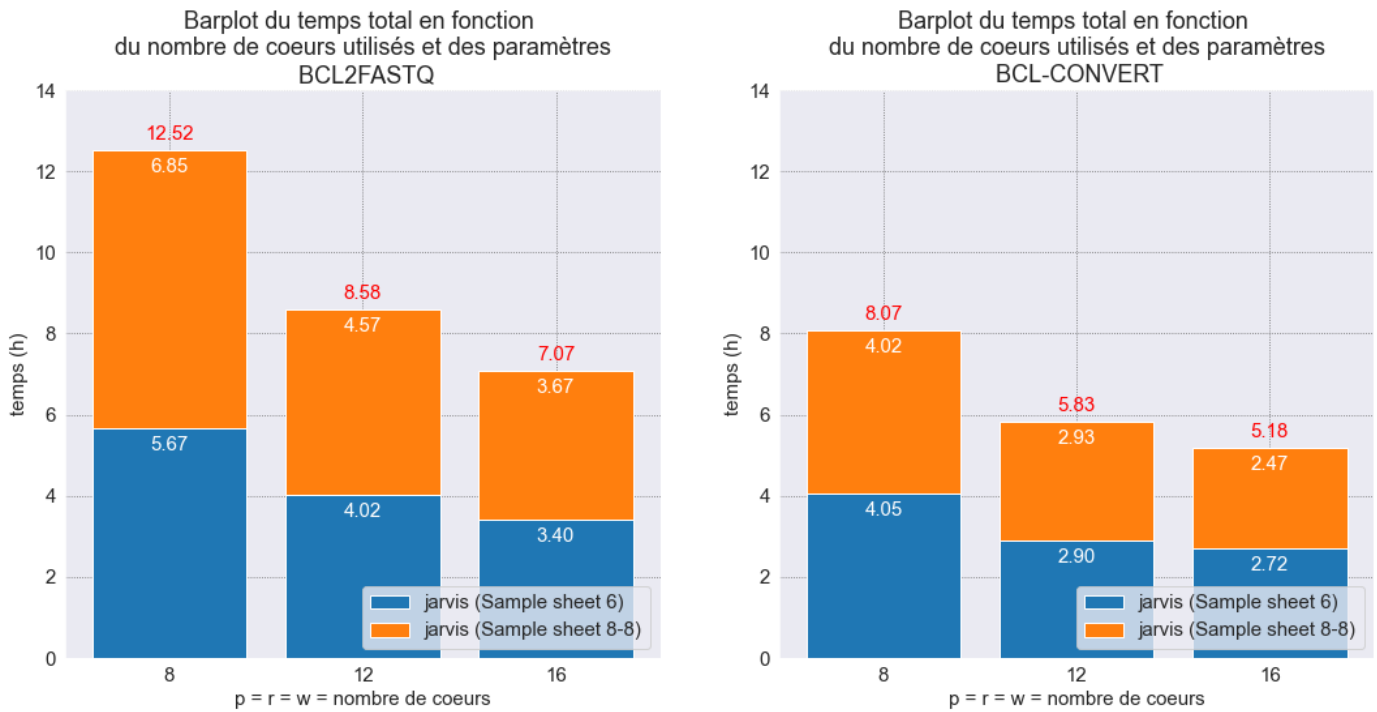

Figure 5: Temps total de génération des FASTQ pour bcl2fastq et bcl-convert

bcl2fastq vs bcl-convert (Temps cpu)

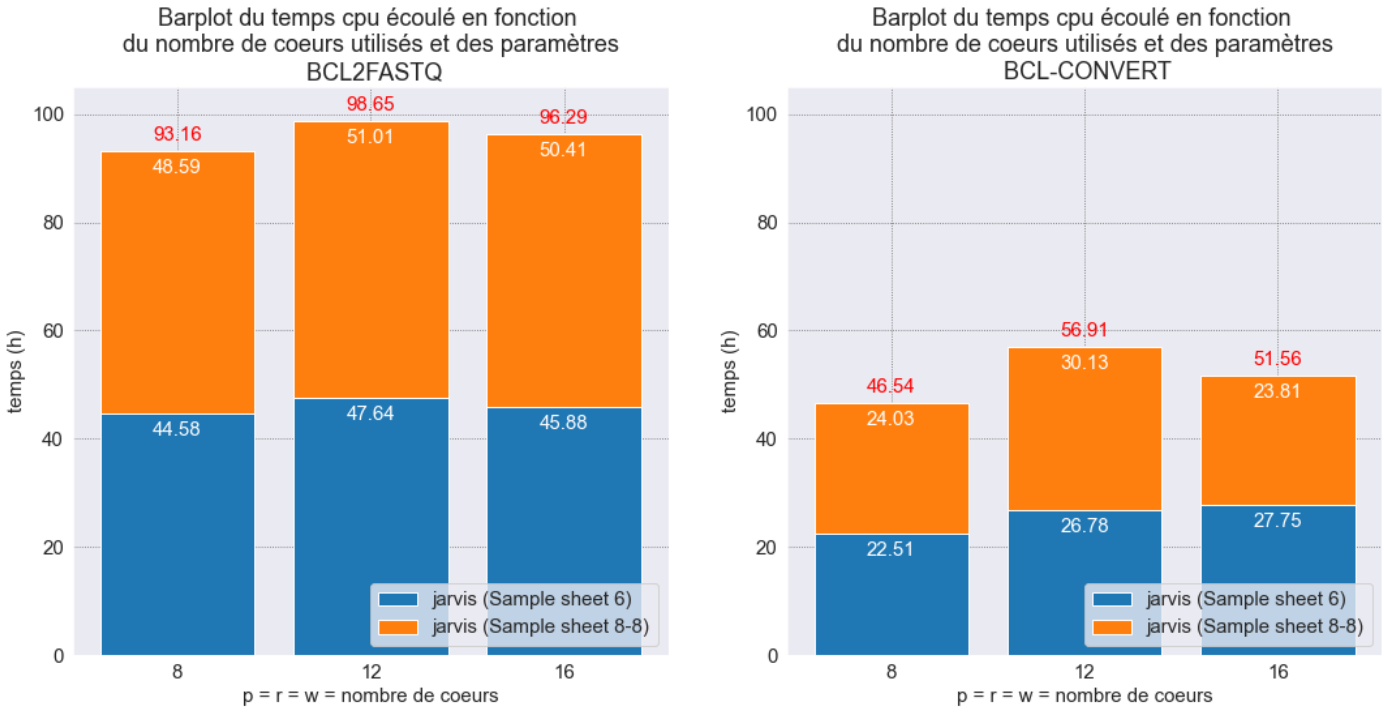

Figure 6: Temps cpu de génération des FASTQ pour bcl2fastq et bcl-convert

bcl2fastq vs bcl-convert (Pourcentage d'utilisation cpu)

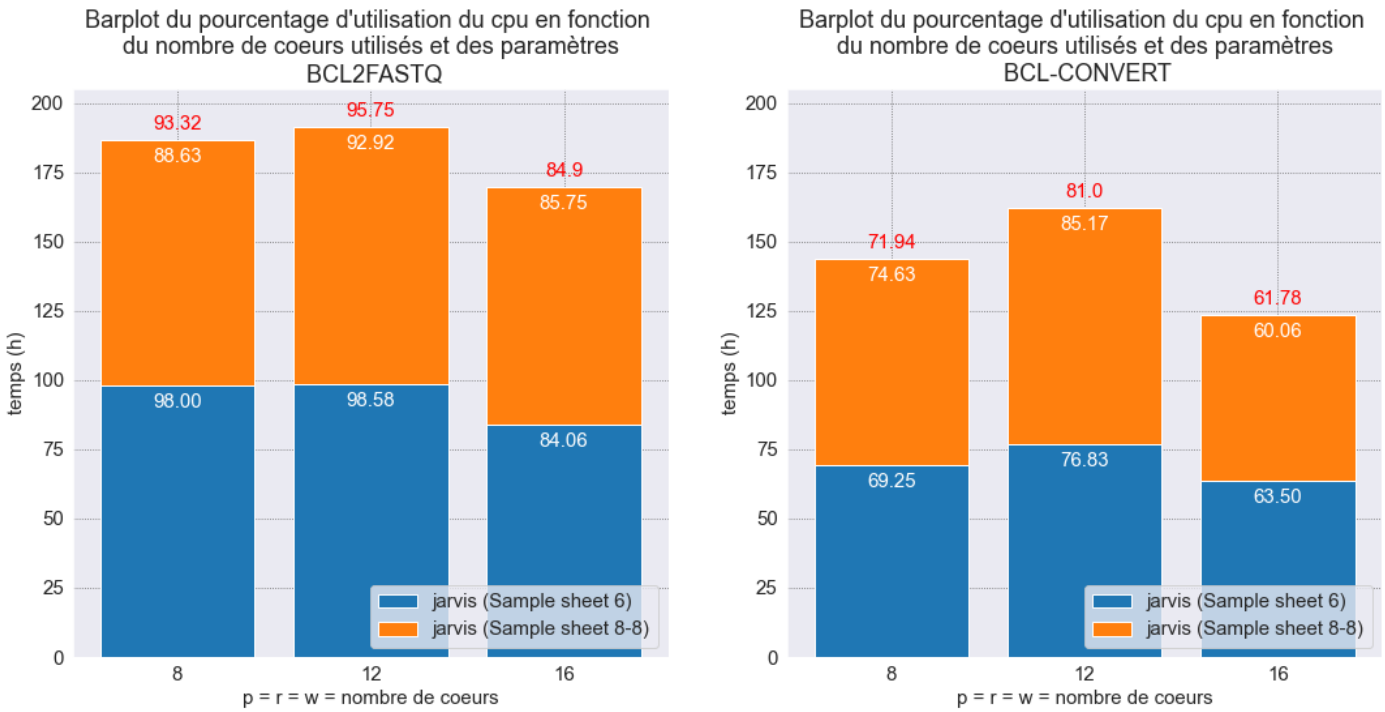

Figure 7: Pourcentage d'ulisation cpu pour la génération des FASTQ pour bcl2fastq et bcl-convert

Section 4

Perspective

Perspective

Détermination de la Migration de bcl2fastq vers bcl-convert

- Mise à jour du pipeline de génération des FASTQ
- Prise en charge des sorties de bcl-convert pour les autres pipelines

Worflow MGI

- Automatisation total du workflow

Ajouter les autres perspectives