Gestion informatique des données de séquençage

William Amory
M1 BI-IPFB Université de Paris

sous la responsabilité de Frédérick Gavory

Gestion informatique des données de séquençage

1. CEA - Genoscope
2. Contexte
3. Objectifs
4. Perspective

Section 1

CEA - Genoscope

CEA - Genoscope

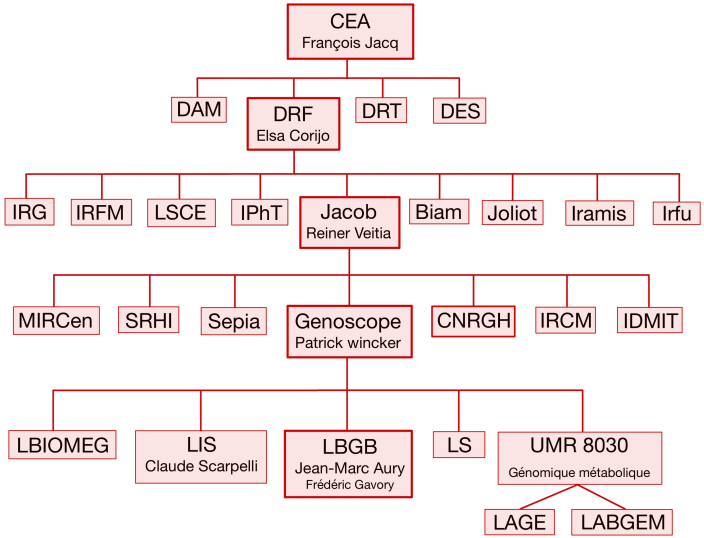
CEA (Commissariat à l'énergie atomique et aux énergies)

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

Genoscope (Centre National de Séquençage)

- Créé en 1996 - 250 salariés
  - Participation **projet Génome humain** (Séquençage du chromosome 14 humain)
  - Développer programmes de génomiques en France
  - Plus grand centre de séquençage français
  - **France génomique**
  - **Projet Tara Océans** - étude des écosystèmes marins planctoniques

Organigrame CEA - Genoscope - LBGB



Section 2

Contexte

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

missions

- Veille technologique
- Contrôle qualité des fichiers de séquences
- Assemblage des séquences et des génomes
- Annotation des séquences et des génomes
- 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 recherches et de productions
- Mise en place de pipelines automatisés
  - génération des fichiers de séquences
  - Contrôle qualité des fichiers de séquences
  - Analyses biologiques

LBGB - Workflow NGS

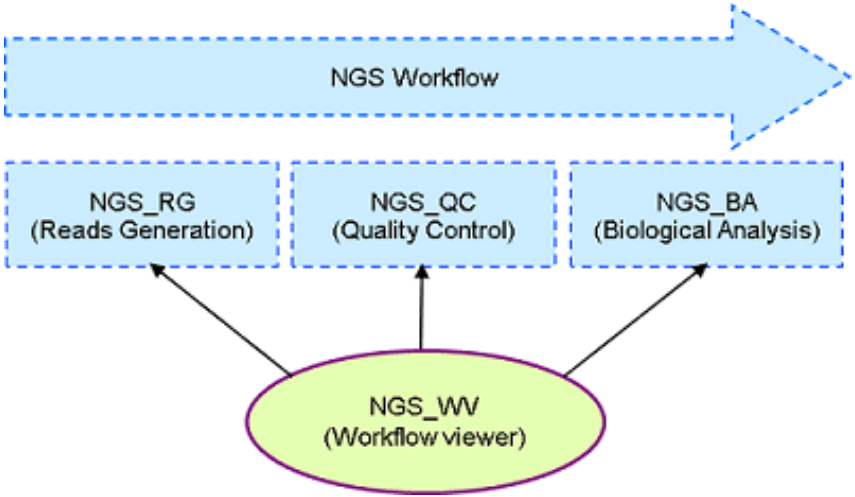


Figure 2: Workflow de génération, de controle qualité et d'analyse biologique des FASTQ

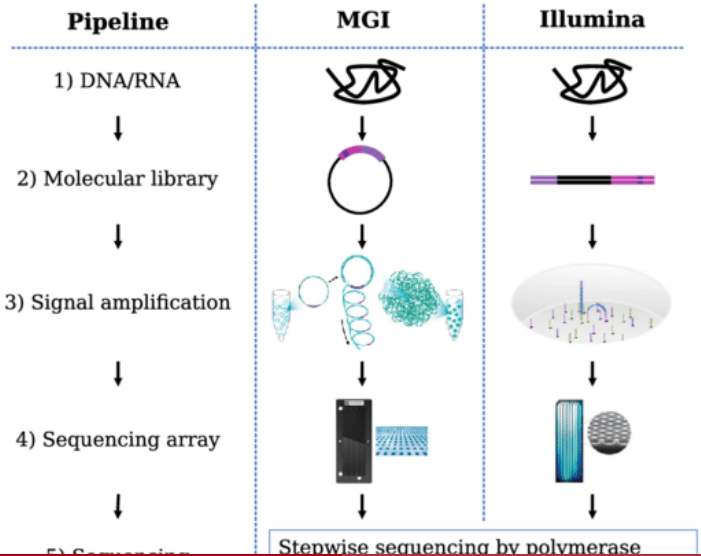
LBGB - MGI

Arrivée des séquenceurs MGI

- 2 DNBSEQ-G400
- 1 DNBSEQ-T7

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

La technologie 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 correspondantes
- 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 existants pour Illumina ?
  - utilisation de nouveaux outils et méthodes ?
- Ecriture du pipeline
  - déterminer l'ordre d'utilisation des outils et méthodes

Autres objectifs de la mission

Evaluation d'outils

- pour les pipelines :
  - Illumina
  - MGI
  - Oxford Nanopore
- Mise en place des outils pertinents
- Remplacement ou arrêt des outils non pertinents

codage d'outils

- Maintenir les pipelines
- Distributions des résultats d'analyses par projet, échantillon/run
- Mettre à jour la base de données (NGL)

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
  - taux de GC
  - moyene de la qualité de chaque read
  - ect ...
- Utilisation des modules utilisés dans le workflow d'illumina

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

permettent la génération des FASTQ et de réaliser le démultiplexage

Comparaison des performances

- Recherche des meilleurs paramètres pour bcl2fastq
  - Nombre de threads lecture/décompression *Bases Calls* (**r**)
  - Nombre de threads Conversion *Bases Calls* en FASTQ (**p**)
  - Nombre de threads écriture/compression FASTQ (**w**)
- Comparaison des performances entres les 2 soft
- Choix de changement de soft

bcl2fastq vs bcl-convert (Temps total)

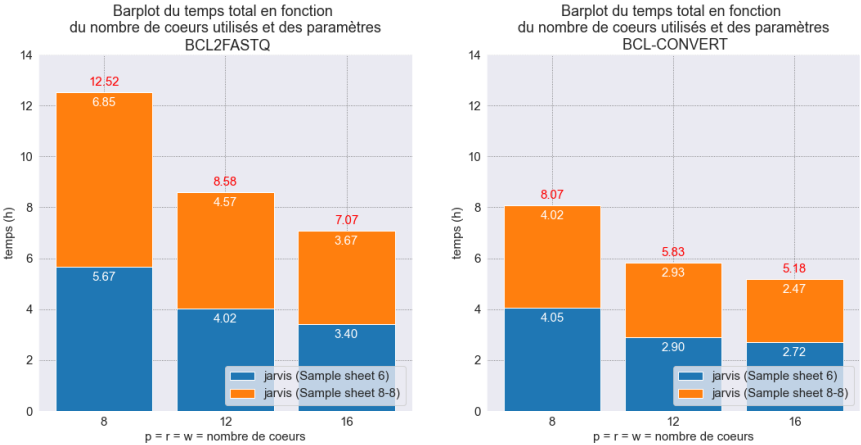


Figure 4: Temps total de génération des FASTQ pour bcl2fastqet bcl-convert

bcl2fastq vs bcl-convert (Temps cpu)

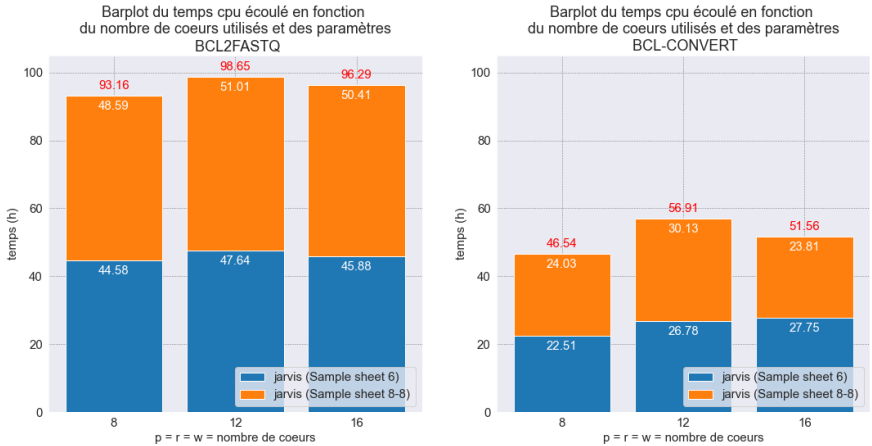


Figure 5: Temps cpu de génération des FASTQ pour bcl2fastqet bcl-convert

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

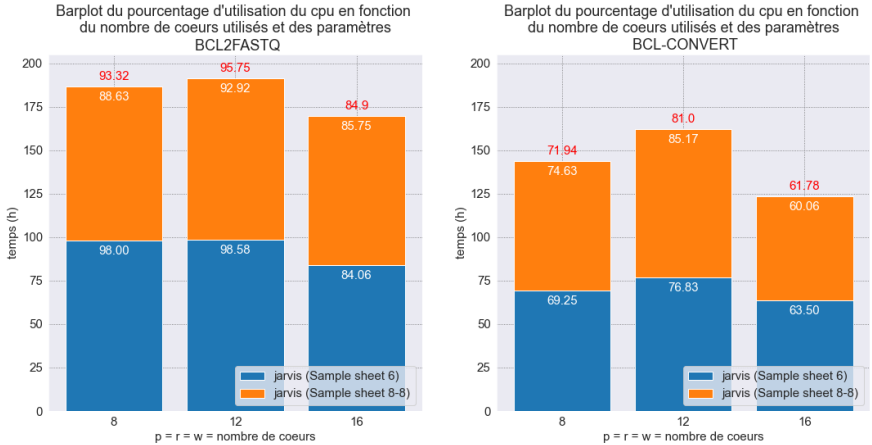


Figure 6: Pourcentage d'ulisation cpu pour la génération des FASTQ pour bcl2fastq et bcl-convert

Section 4

Perspective

Perspective

Détermination de la migration de bcl2fastq vers bcl-convert

- Discussion avec les équipes **LIS** et **LS**
- Mise à jour du pipeline de génération des FASTQ
- Prise en charge des sorties de bcl-convert pour les autres pipelines

Worflow MGI

- Ecriture du pipeline MGI
- Mise en service du pipeline MGI
- Automatisation total du workflow

Perspective

Evaluation d'autres outils

- outils d'assignation taxonomique
- outils de *filtering*, *trimming*
- intégration d'outils des autres groupes de travails dans les pipelines
  - outils de *mapping*, assemblage, *scafold* ...

Bibliographie

- Impact of sequencing depth and technology on de novo RNA-Seq assembly, Patterson. 2022-01-23, BMC Genomics. 10.1186/s12864-019-5965-x
- bcl2fastq2 Conversion Software v2.20 Software Guide (15051736). 2019, Illumina, Inc.
- BCL Convert Software Guide v3.7.5 (1000000163594). 2021, Illumina, Inc.
- perl - The Perl 5 language interpreter - Perldoc Browser. 2022-01-23, https://perldoc.perl.org/perl
- The Comprehensive Perl Archive Network. 2022-01-23, www.cpan.org

Merci de votre attention

Données supplémentaires - La technologie MGI

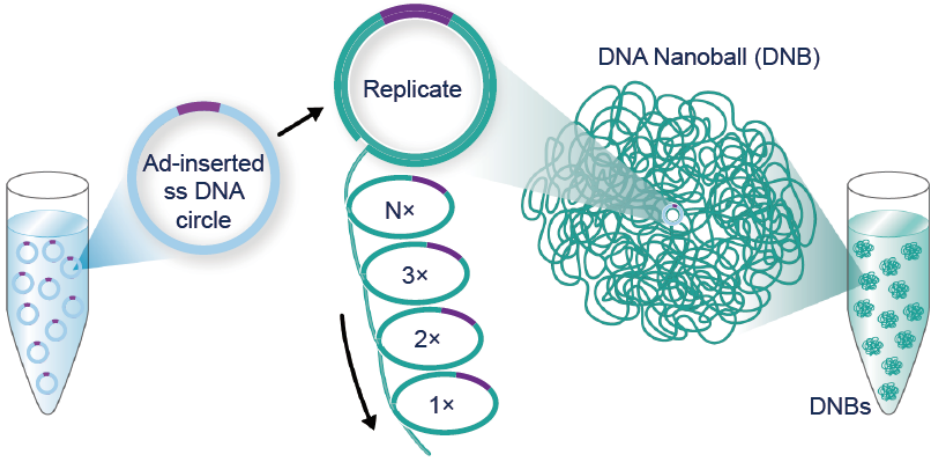


Figure 7: Schéma de la technologie des *DNA nanoballs* de MGI