



LABORATOIRE DE BIOINFORMATIQUE POUR LA GÉNOMIQUE ET LA BIODIVERSITÉ

Master de bioinformatique - ingénierie de plate-forme en biologie
UNIVERSITÉ DE PARIS

Rapport d'alternance

Gestion informatique des données de séquençage

29 janvier 2022

William Amory
sous la responsabilité de Frédérick Gavory



Table des matières

1	Introduction	2
1.1	LBGB au sein du Genoscope et du CEA	2
1.2	Contexte et missions du LBGB	2
1.3	Présentation du workflow NGS	3
1.4	La technologie MGI	3
2	Ojectifs	4
3	Matériels et Méthodes	4
4	Résultats des évaluations de bcl2fastq et bcl-convert	5
4.1	Détermination des meilleurs paramètres pour bcl2fastq	5
4.2	Comparaison entre bcl2fastq et bcl-convert	5
5	Perspectives	6
5.1	diagramme de gantt	7
	Notes	7
	Références	8

1 Introduction

1.1 LBGB au sein du Genoscope et du CEA

Le Genoscope (CNS¹) a été créé en 1996 pour participer au projet mondial de séquençage du génome humain (*Human Genome Project*) qui a débuté en 1990 et c'est terminé en 2003. Il a notamment participé au séquençage du chromosome 14 humain. Le Genoscope participe également à développer des programmes de génomiques en France dans le cadre du projet France génomique. Aujourd'hui un des projets phare du Genoscope est le projet **Tara**, qui a pour objectifs l'étude des écosystèmes marins.

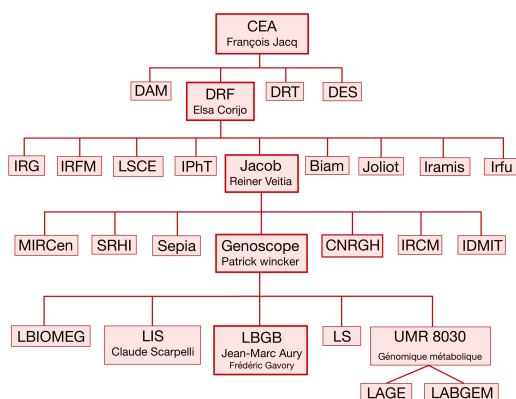


FIGURE 1 – Organigramme situant l'équipe du LBGB au sein du Genoscope et du CEA

Le Laboratoire de Bioinformatique pour la Génomique et la Biodiversité (**LBGB**) dirigé par Jean-Marc Aury, fait partie du Genoscope qui est une composante de l'institut François Jacob (**Jacob**) de la direction de la recherche fondamentale (**DRF**) du Commissariat à l'Énergie Atomique et aux Énergies Alternatives (**CEA**), qui a été fondé le 18 octobre 1945 par Charles de Gaulle. L'intégration du genoscope au CEA a été réalisée en 2007, et en 2017 il devient une direction de l'institut François Jacob.

1.2 Contexte et missions du LBGB

Les missions qui sont confiées au LBGB sont de réaliser le contrôle qualité des données de séquences issues des différents séquenceurs, d'effectuer l'assemblage² des séquences et l'annotation³ des génomes. Tout en permettant la visualisation de chacune des missions (visualisation des annotations, de la qualité des reads⁴, ect.). Il a également la mission de faire de la veille technologique d'outils et méthodes permettant de réaliser ses autres missions. Le laboratoire est divisé en plusieurs groupes de travail. Le groupe *production* (dont je fais parti), le groupe *assemblage*, le groupe *annotation* et le groupe *évaluation des technologies de séquençage*.

Les missions du groupe de *production* sont de développer, tester et maintenir les scripts dans l'objectif de répondre aux besoins des équipes de recherche et de séquençage en automatisant au maximum les processus. Notamment dans la mise en place et au maintien de pipelines automatiques pour la génération des fichiers de séquences, le contrôle qualité et les analyses biologiques de ces derniers. Le groupe a également la mission de faire de la veille technologique et d'évaluer de nouveaux outils et méthodes pour chacune de ses autres missions.

1.3 Présentation du workflow NGS

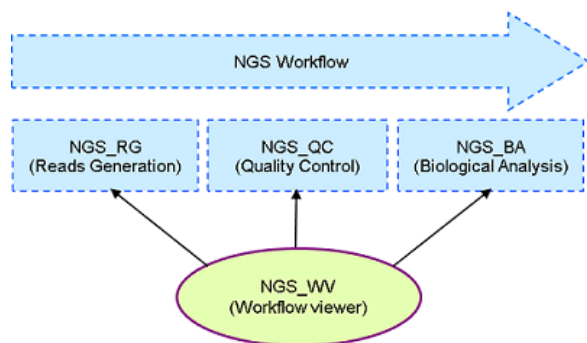


FIGURE 2 – Workflow de génération, de contrôle qualité et d'analyse biologique des fastq

Le workflow ngs est composé de trois pipelines pour la technologie Illumina. Le premier (ngs_rg), permet la génération des reads. Le second (ngs_qc), permet de réaliser le contrôle qualité. Le dernier (ngs_ba), permet de faire les analyses biologiques. Ces trois pipelines sont automatisés dans le workflow et permettent de réaliser la distribution des données de séquences par projet, de les trier par échantillons, runs et technologies de séquençage. Ils réalisent aussi le nettoyage, l'analyse de ces fichiers et mettent à jour la base de données de référence ngl.

1.4 La technologie MGI

Le genoscope et le CNRGH⁵ ont récemment fait l'acquisition de séquenceurs MGI⁶ (2 DNBSEQ-G400 et 1 DNBSEQ-T7).

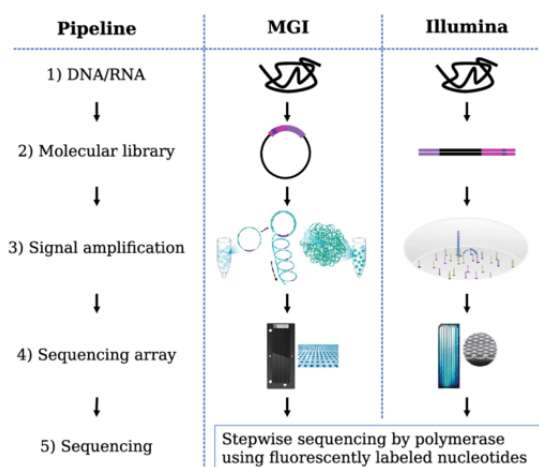


FIGURE 3 – Différences entre Illumina et MGI de technologie NGS

Il s'agit de séquenceurs à haut débit équivalents à un HiSeq 4000 (Illumina) pour le DNBSEQ-G400 et à un NovaSeq 6000 (Illumina) pour le DNBSEQ-T7. Les principales différences entre MGI et Illumina sont dans la création des librairies et dans la méthode d'amplification d'ADN. Les librairies sont double brins circulaire pour MGI, alors que pour Illumina elle est double brins linéaire. L'amplification ADN est réalisée en solution pour MGI puis déposée sur la Flowcell⁷, alors que pour Illumina elle est réalisée après immobilisation sur les Flowcell.

	DNBSEQ-G400	DNBSEQ-T7	HiSeq 4000	NovaSeq 6000
Max Number of Flow Cells	2	4	2	2
Max Lane/Flow Cell	4	1	4	4
Run Time	~ 14-37 h	~ 20-30 h	~ 24-84 h	~ 13-44 h
Max Reads Per Run	1.8 billion	5 billion	10 billion	20 billion
Max Read Length	2 × 200 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp

TABLE 1 – Spécification des séquenceurs

2 Ojectifs

L'objectif principale de ma mission est la mise en place d'un workflow pour les séquenceurs de MGI. Plus précissément il s'agira de créer un pipeline de génération de reads (`ngs_rg_mgi`) et un pour le contrôle qualité (`ngs_qc_mgi`). Le workflow devra créer et mettre à jour l'état des run et des *lanes* dans `ngl`, réaliser le démultiplexage⁸ et faire un contrôle qualité des fastq à chaque étapes. De plus, il renommera les fichiers fastq et les déplacera dans le projet associé dans `ngl`. Il devra aussi déplacer les fichiers de statistiques et les fichiers fastq correspondant aux index non attendus dans leurs répertoire dédiés dans `ngl`.

Les autres objectifs de ma mission sont de réaliser des évaluations de nouveaux outils pour les différents pipelines mis en place pour les différentes technologies de séquençage. Tel que l'évaluation d'outils de génération de fichier fastq et de démultiplexage en vu du remplacement de `bcl2fastq` par `bcl-convert`. Ainsi que de maintenir les pipelines des différentes technologies en ajoutant, remplaçant certains outils suite à une évaluation de ces derniers ; notamment pour le workflow d'Illumina et celui de MGI une fois ce dernier créé.

3 Matériels et Méthodes

La Genoscope possède 7 clusters de calucls pour la *production* sur le serveur *inti*, ces derniers disposent de 16 coeurs et de 257 Go d'espace de stockage. L'accès à l'utilisation des clusters est réalisé par le logiciel Slurm⁹. Il dispose également de sa propre base de données de référence (`ngl`). La gestion et le suivi du développememnt informatique est réalisé par le système Jira¹⁰. L'écriture du workflow des pipelines pour les séquenceurs MGI sera réalisée dans le langage de programmation Perl. L'utilisation de ce langage est rendu necessaire pour des raisons historique du laboratoire, puisque de nombreuses librairies et modules qui seront à utiliser dans l'écriture des pipelines sont écrits en Perl. Le workflow pour MGI s'appuiera sur le workflow d'Illumina qui est totalement implémenté en Perl.

C'est pour toutes ces raisons qu'il m'a été nécessaire d'apprendre à coder en Perl en réalisant un programme permettant de faire des analyses statistiques élémentaires sur des fichiers fastq. Tel que le taux de GC, la moyenne du score de la qualité, ainsi que plusieurs autres métriques. Le programme est capable de gérer les fichiers fastq issue de séquençage *single end*¹¹ et *paired end*¹². Cela m'a permis de prendre en main les modules utilisés pour les différents pipelines déjà en place, notamment pour le pipelines d'llumina. De plus cela m'a permi de prendre en main de l'environement de travail, l'utilisation du lancement de job sur les noeuds de calculs et l'utilisation des modules utilisés pour le workflow d'llumina

Une première évaluation d'outils a également été effectuée en vu du remplacement de `bcl2fastq` par `bcl-convert`. Il s'agit de deux logiciels de génération de fichiers fastq et de démultiplexage qui sont développés et commercialisés par Illumina.

Dans un premier temps, il a été nécessaire de déterminer les conditions optimales (temps total rapide, pourcentage d'utilisation cpu¹³ maximum) en fonction des ressource disponible sur

les clusters de la *production* (16 coeurs maximum) du nouveau serveur (*inti*), pour bcl2fastq dans l'objectif de pouvoir comparer les performances des deux logiciels dans les mêmes conditions. Les conditions optimales sont déterminées en fonction des paramètres suivant de bcl2fastq :

- **r** : nombre de *threads*¹⁴ accordé pour la décompression et la lecture des *Bases Calls*¹⁵
- **p** : conversion des *Bases Calls* en fastq
- **w** : écriture et compression des fichier fastq

Tous ces tests ont été réalisés sur le même noeud de calcul, dans l'objectif de minimiser les biais. La comparaison est effectuée sur le temps total de génération des fastq et le démultiplexage, ainsi que le temps cpu et le pourcentage d'utilisation des cpu.

4 Résultats des évaluations de bcl2fastq et bcl-convert

4.1 Détermination des meilleurs paramètres pour bcl2fastq

Après avoir effectué différentes combinaisons des paramètres, nous avons mis en évidence que la variation du paramètre **r** et **w** en fixant le paramètre **p**, n'apportait pas de différences significatives pour le temps total d'exécution, le temps cpu ou le pourcentage d'utilisation cpu, comme on peut l'observer sur la figure 4.

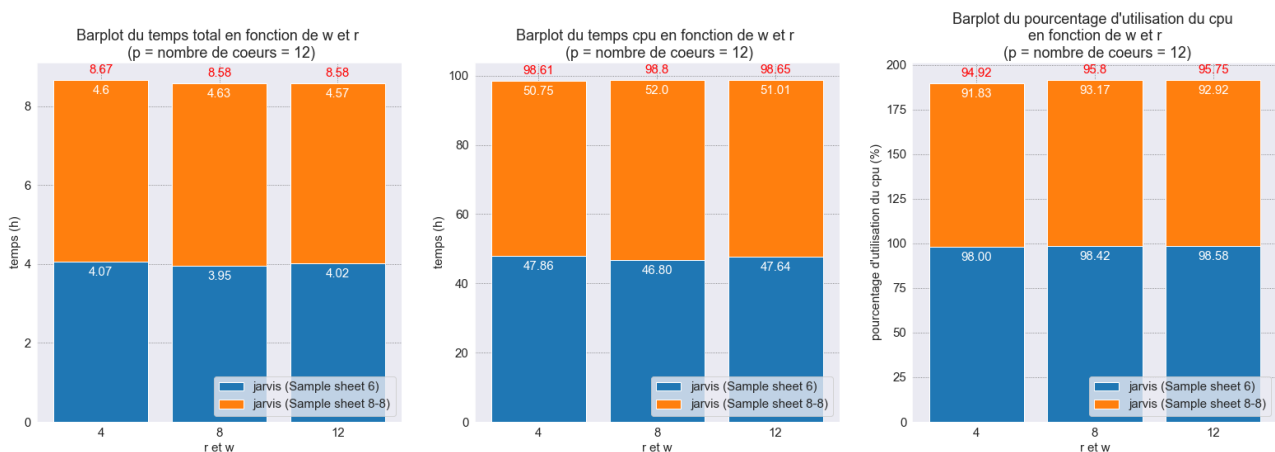


FIGURE 4 – Digrammes en bâtons du temps total d'exécution (à gauche), temps cpu (au milieu) et du pourcentage d'utilisation des cpu (à droite)

Il y a deux *sample sheet*, car le nombre de bases considérés des *reads index* entre les *lanes* est différent, obligeant à réaliser deux appels différents au logiciel pour générer les fastq et le démultiplexage.

4.2 Comparaison entre bcl2fastq et bcl-convert

Nous avons donc fait varier les paramètres **p**, **r** et **w** de manière à ce que chacun des paramètre soit égale au nombre de coeurs accordés aux deux logiciels. On observe bien, sur la figure 5, que plus on augmente le nombre de coeurs pour chacun des logiciels (et donc le nombre de *threads*

pour p , r et w) plus la génération des fastq et le démultiplexage est rapide. De plus on remarque que bcl-convert permet de réduire le temps d'environ 1/3 par rapport à bcl2fastq.

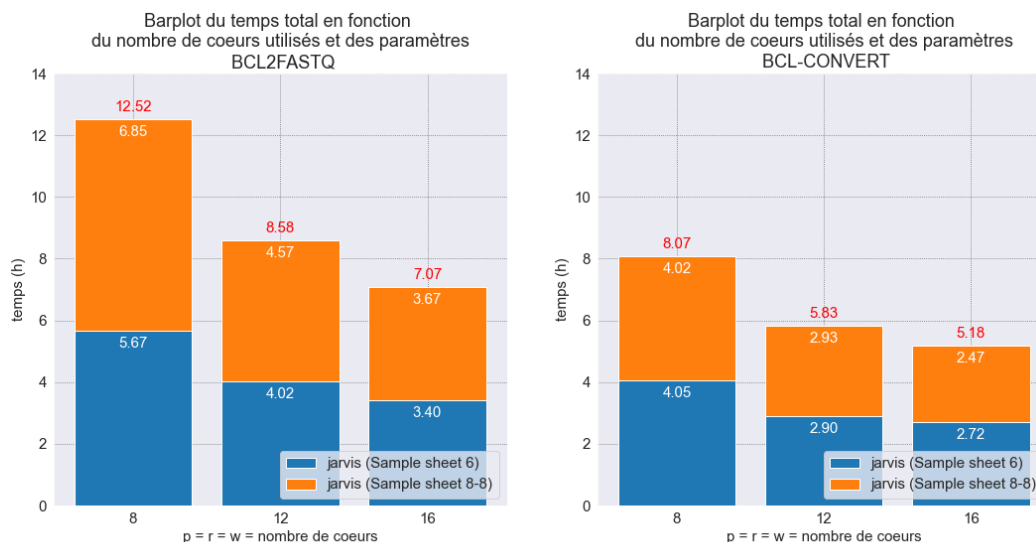


FIGURE 5 – Temps total de génération des fastq pour bcl2fastq et bcl-convert

5 Perspectives

Concernant les perspectives de bcl-convert il reste à réaliser un cahier des charges référençant tous les changements à effectuer dans les différents pipelines pour la mise à jour de bcl2fastq vers bcl-convert. Ce cahier des charges prendra en compte le changement d'arborescence des fichiers de sortie entre les deux logiciels, ainsi toutes les modifications à effectuer dans les différents pipelines pour permettre le bon fonctionnement des workflow. Dû à la pression actuelle autour de la technologie MGI, c'est un autre développeur qui se chargera de suivre ce cahier des charges et de réaliser les modifications.

Concernant le workflow de MGI, il nous faut dans un premier temps déterminer les outils et méthodes nécessaires (utilisation de ceux du workflow d'Illumina ou de nouveaux). Une fois ceci déterminé il restera à écrire les deux pipelines, celui de génération de reads (ngs_rg_mgi) et celui de contrôle qualité (ngs_rg_mgi). L'objectif sur le long terme est d'arriver à un workflow totalement automatisé, comme celui d'Illumina.

Il y a aussi l'évaluation d'autres outils utiles pour les pipelines, comme l'évaluation d'outils de *trimming*, *filtering*, d'assignation taxonomique, ect.

5.1 diagramme de gantt

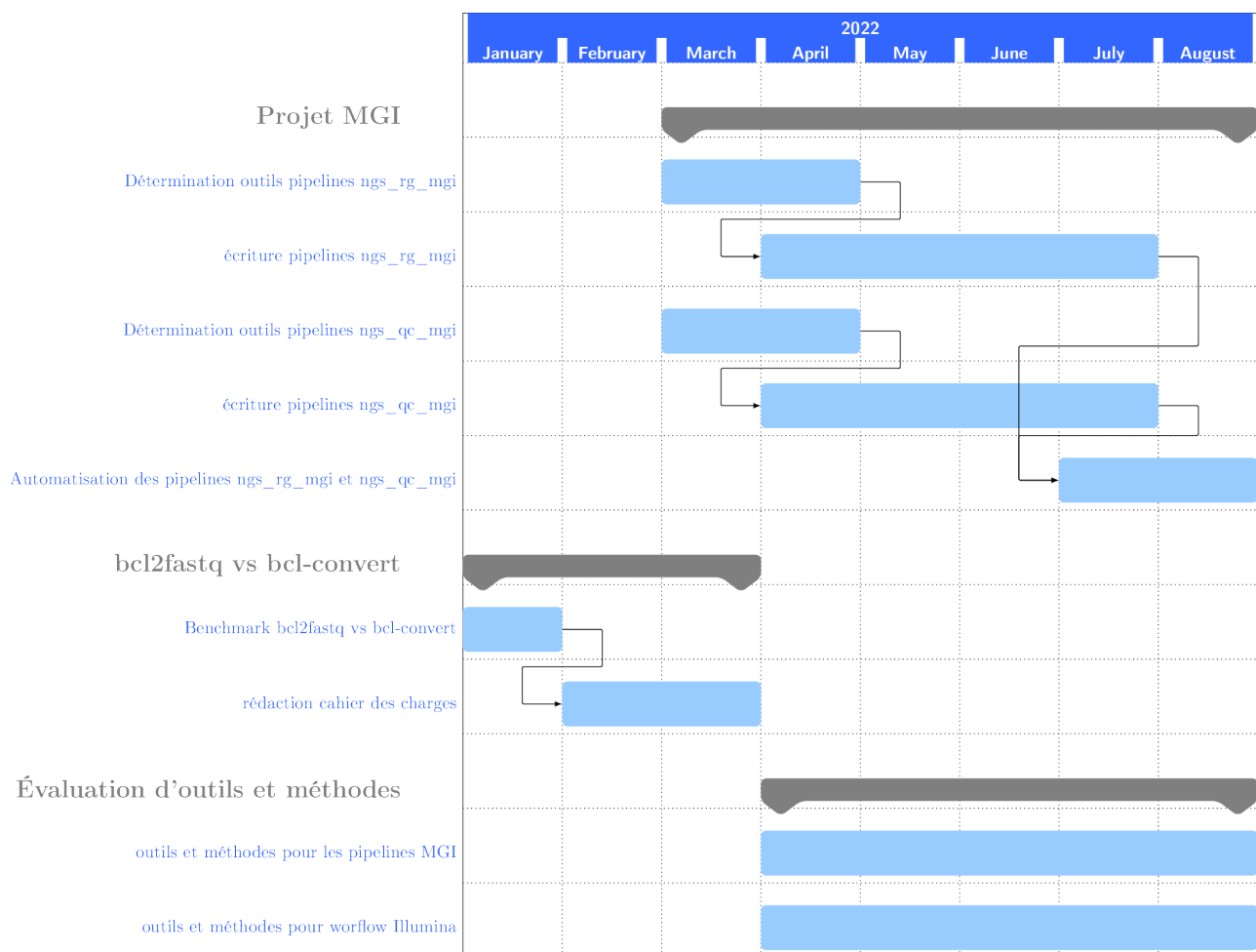


FIGURE 6 – Diagramme de gantt des différents projets et missions

Notes

¹Centre National de Séquençage

²Reconstruction d'un génome à partir fragment de ce dernier

³Documenter le plus exhaustivement possible les informations de l'assemblage permettant de prédire la fonction d'une molécule

⁴Lecture d'une séquence par un séquenceur d'ADN d'un fragments d'ADN.

⁵Centre National de Recherche en Génétique Humaine

⁶Membre du groupe BGI dont les missions sont : R&D, production et vente d'instruments de séquençage d'ADN, de réactifs et de produits connexes

⁷Lame d'absorption des fragments d'ADN et cuve réacteur du séquençage

⁸Séparation des différents reads d'une *lane* en fonction de l'index d'échantillon

⁹Logiciel open source d'ordonnancement des tâches informatiques

¹⁰Logiciel de gestion de projet, d'incidents et de suivi de bugs

¹¹Lecture dans un seul sens des reads par le séquenceur

¹²Lecture dans les deux sens des reads par le séquenceur

¹³Central Processing Unit (unité centrale de traitement, en français)

¹⁴Processus : instructions du langage machine d'un processeur.

¹⁵Fichier d'attribution des bases nucléiques en fonction des pics du chromatogramme lors du séquençage

Références

- [1] BCL Convert.
- [2] BCL Convert Software Guide v3.7.5 (1000000163594). page 22.
- [3] bcl2fastq2 Conversion Software v2.20 Software Guide (15051736). page 27.
- [4] The Comprehensive Perl Archive Network - www.cpan.org.
- [5] perl - The Perl 5 language interpreter - Perldoc Browser.
- [6] J. Patterson, E. J. Carpenter, Z. Zhu, D. An, X. Liang, C. Geng, R. Drmanac, and G. K.-S. Wong. Impact of sequencing depth and technology on de novo RNA-Seq assembly. *BMC Genomics*, 20(1) :604, July 2019.