DE LA RECHERCHE À L'INDUSTRIE

cea

Université Paris Cité

Centre National de Séquençage

GENOSCOPE

www.cea.fr

Gestion informatique des données de séquençage

William Amory
M1 BI-IPFB Université Paris Cité

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

Sous la responsabilité de Frédérick Gavory

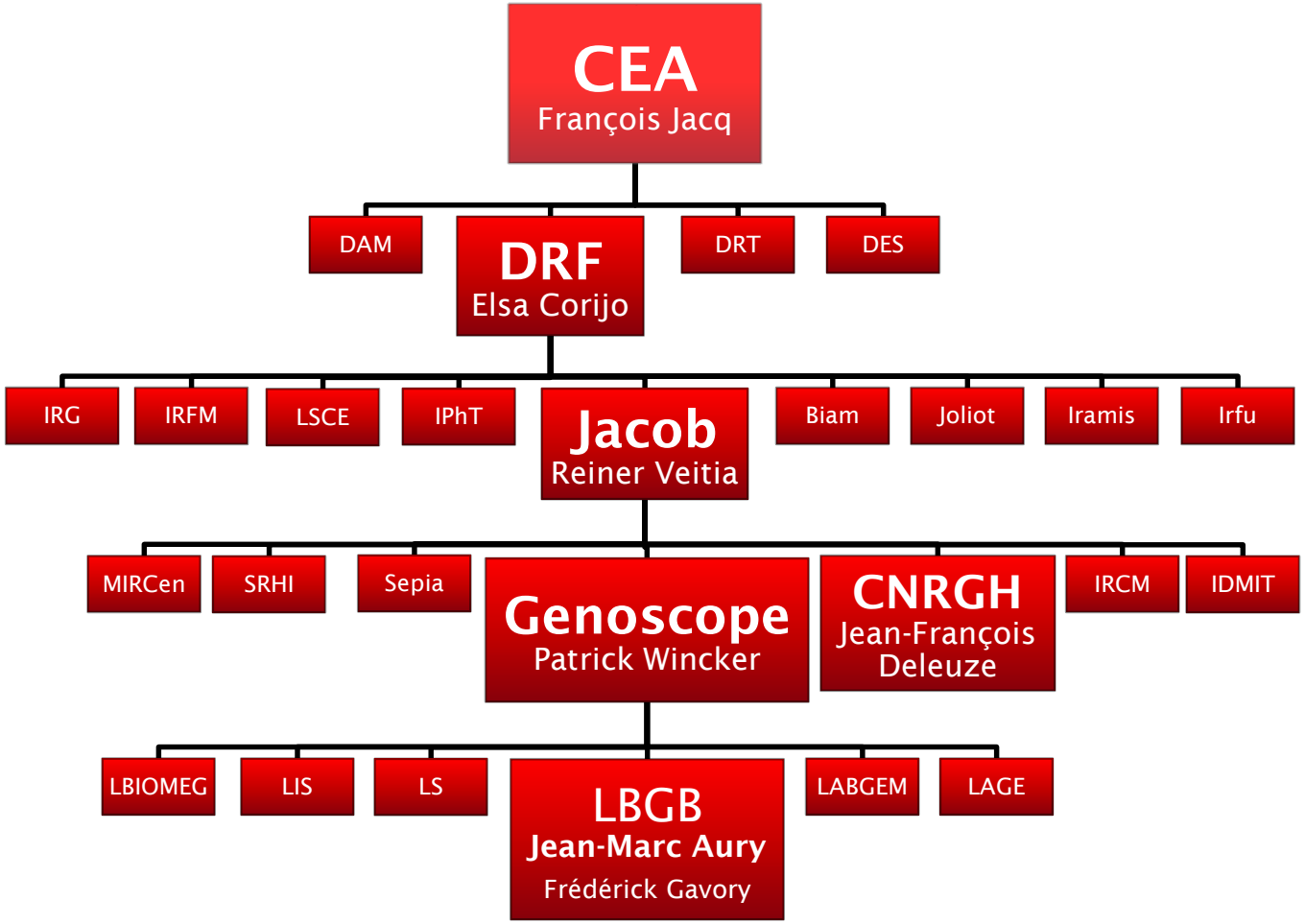


Figure 7 – Organigramme situant l'équipe du *Laboratoire de Bioinformatique pour la Génomique et la Biodiversité (LBGB)* au sein du Genoscope et du CEA (2022)

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

- 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 au **projet Génome humain** (Séquençage du chromosome 14)
- Développerment de programmes de génomiques en France
- Plus grand centre de séquençage français
- **France génomique** – unité mixe de service – regroupe les 4 principaux organismes de recherche (CEA, CNRS, TNRA, INSERM) – rassemblement de la majorité des plateforme de séquençage et de bioinformatique français
- **Projets Tara** (Pacific – Océans – Artic …) - étude des écosystèmes marins
- **Projets ERGA** (European Reference Genome Atlas) – création d'une base de données de références de haute qualité des génomes d'espèces européeene

Plusieurs groupes de travail

- Evaluation des techniques de séquençage
- **Production**
- Assemblage
- Annotation

Missions du groupe Production

- Répondre aux besoins des équipes de recherches et de productions
- Vielle technologique et évaluation de nouveaux outils
- Développer, tester et maintenir les librairies et scripts
- Mise en place et maintient de pipelines automatisant l'exécution de ces scripts pour le **Genoscope** (centre national de séquençage) et le **CNRGH** (centre national de recherche en génétique humaine)
  - Génération des fichiers de séquences
  - Contrôle qualité et nettoyage des fichiers de séquences
  - Analyses biologiques
- Mise à jour de la base de données de référence NGL (*Next Generation LIMS*)

Contexte et objectifs de la mission

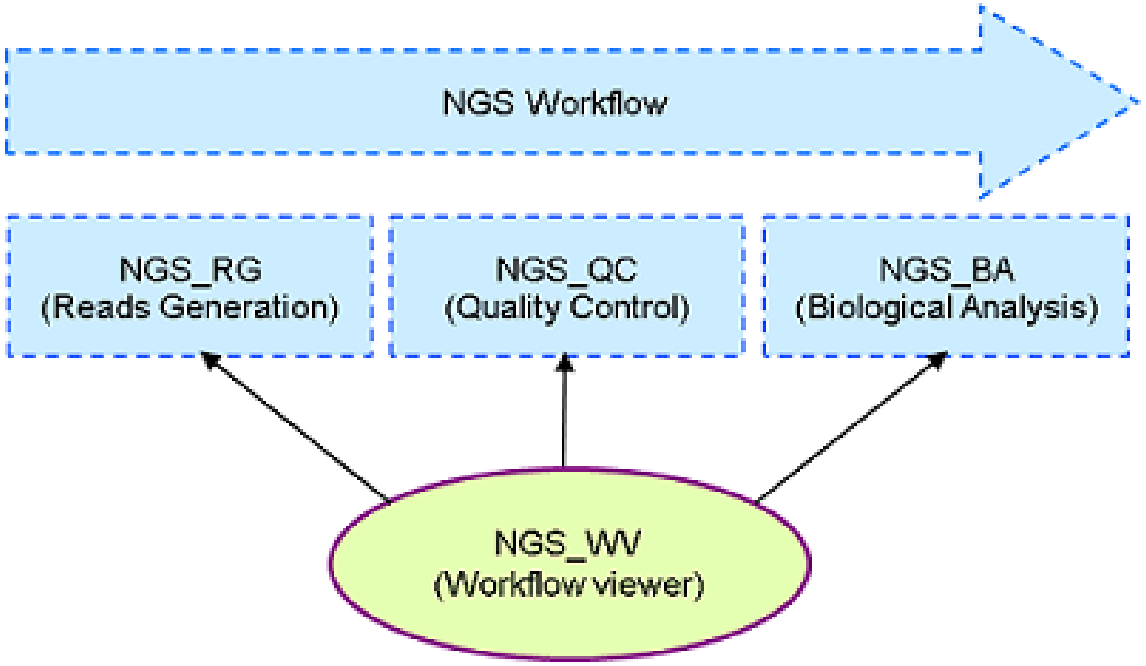


Figure 1 – Workflow de génération, de contrôle qualité et d'analyse biologique des FASTQ

https://www.genoscope.cns.fr/rdbioseq/ consulté le 21/06/2022

Arrivée des Séquenceurs MGI

2 DNBSEQ-G400

- 2 flowcell - 2/4 pistes
- 1.4 TB
- 5000 Millions de reads
- Taille max des reads :
  - 150pb PE
  - 400pb SE
- Temps moyen d'un run :
  - 24h ~ 30h

1 DNBSEQ-T7

- 4 flowcell - 1 piste
- 6 TB
- 1800 Millions de reads
- Taille max des reads :
  - 200pb PE
  - 400pb SE
- Temps moyen d'un run :
  - 14h ~ 109h

https://en.mgi-tech.com/products/ consulté le 21/06/2022

Contexte et objectifs de la mission

Développement d'un workflow NGS pour la technologie MGI

- Pipeline de génération de fichiers de séquences (NGS-RG MGI)
- Pipeline de contrôle qualité des séquences des fichiers de séquences (NGS-QC MGI)
- Pipeline d'analyses biologiques (NGS-BA MGI)

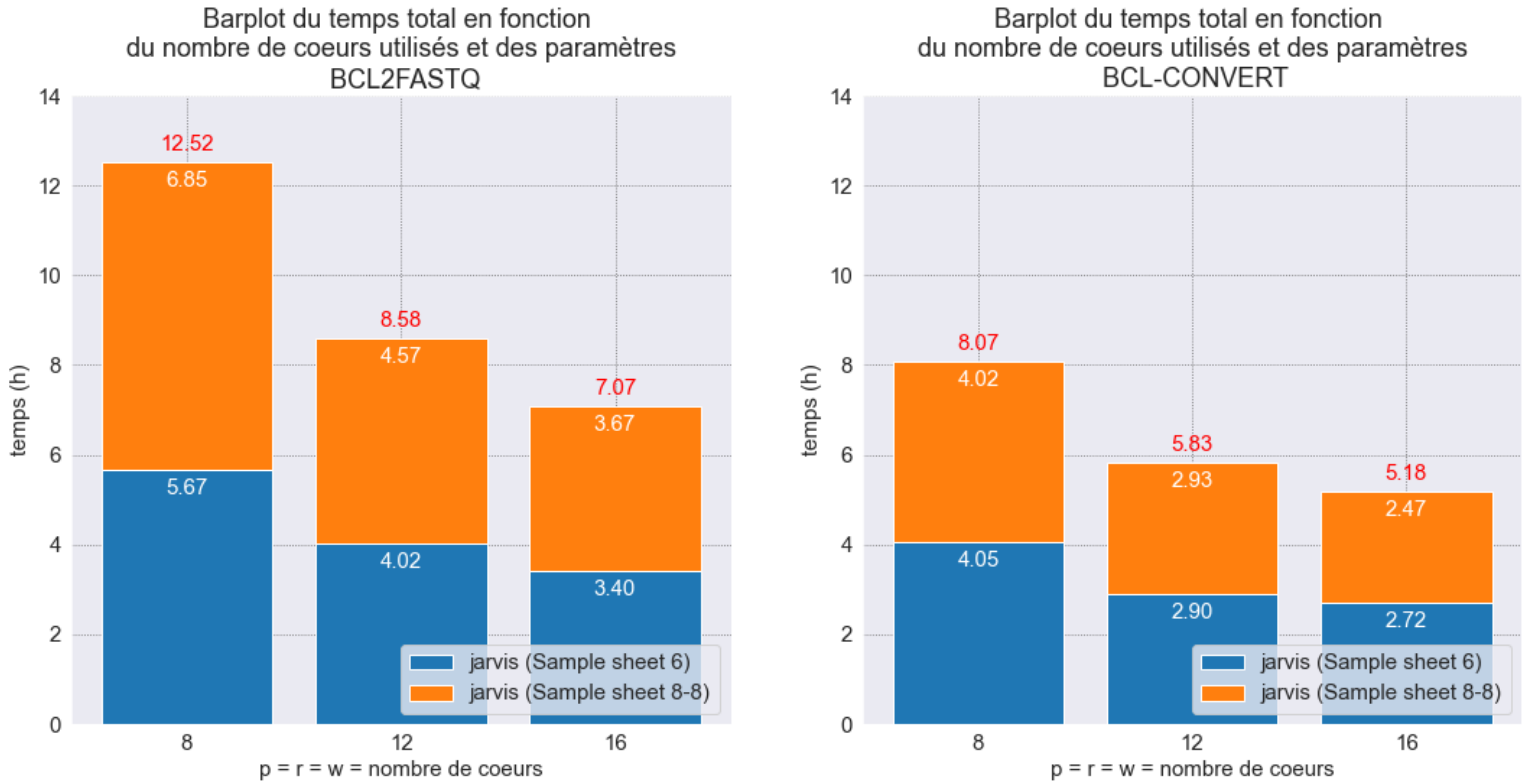
Autres missions

- Vielle technologique
- Evaluation de nouveaux outils pour les pipelines existants
- Développer, tester et maintenir les pipelines existants
- Répondre aux besoins des équipes de recherches et de séquençage

Logiciels permettant la génération des FASTQ et de réaliser le démultiplexage développé par Illumina

Comparaison des performances entres les 2 logiciels

- Temps total d'exécution
- Temps cpu
- Pourcentage d'utilisation des cpu



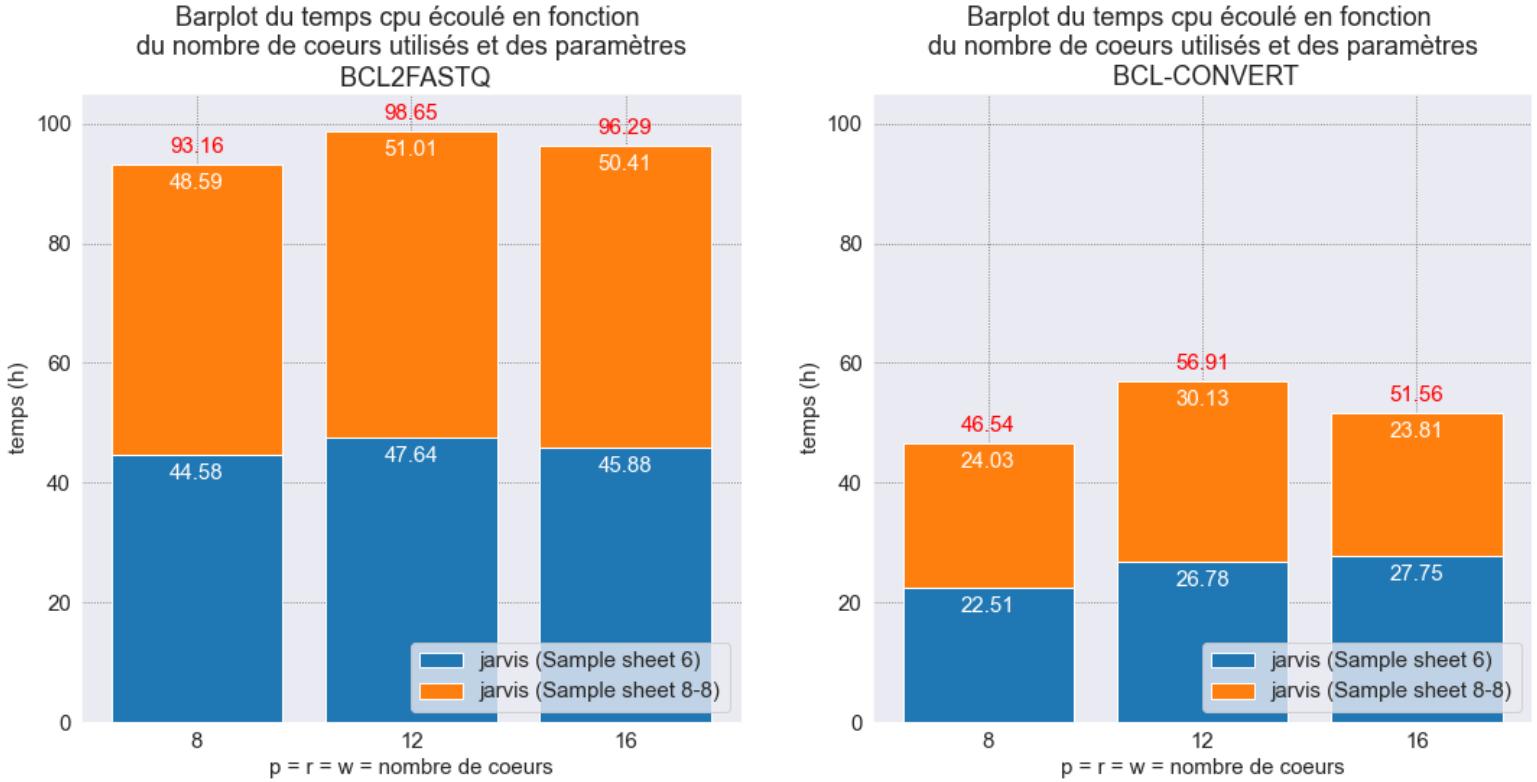
- p : nombre de cœurs pour la conversion des Base Calls en FASTQ et le démultiplexage
- r : nombre de cœurs pour la décompression et la lecture des Base Calls
- w : nombre de cœurs pour l'écriture et la compression des FASTQ

Figure 1 – Comparaison des performances en temps d'exécution des logiciels bcl2fastq et bcl-convert

Logiciels permettant la génération des FASTQ et de réaliser le démultiplexage développé par Illumina

Comparaison des performances entres les 2 logiciels

- Temps total d'exécution
- Temps cpu
- Pourcentage d'utilisation des cpu



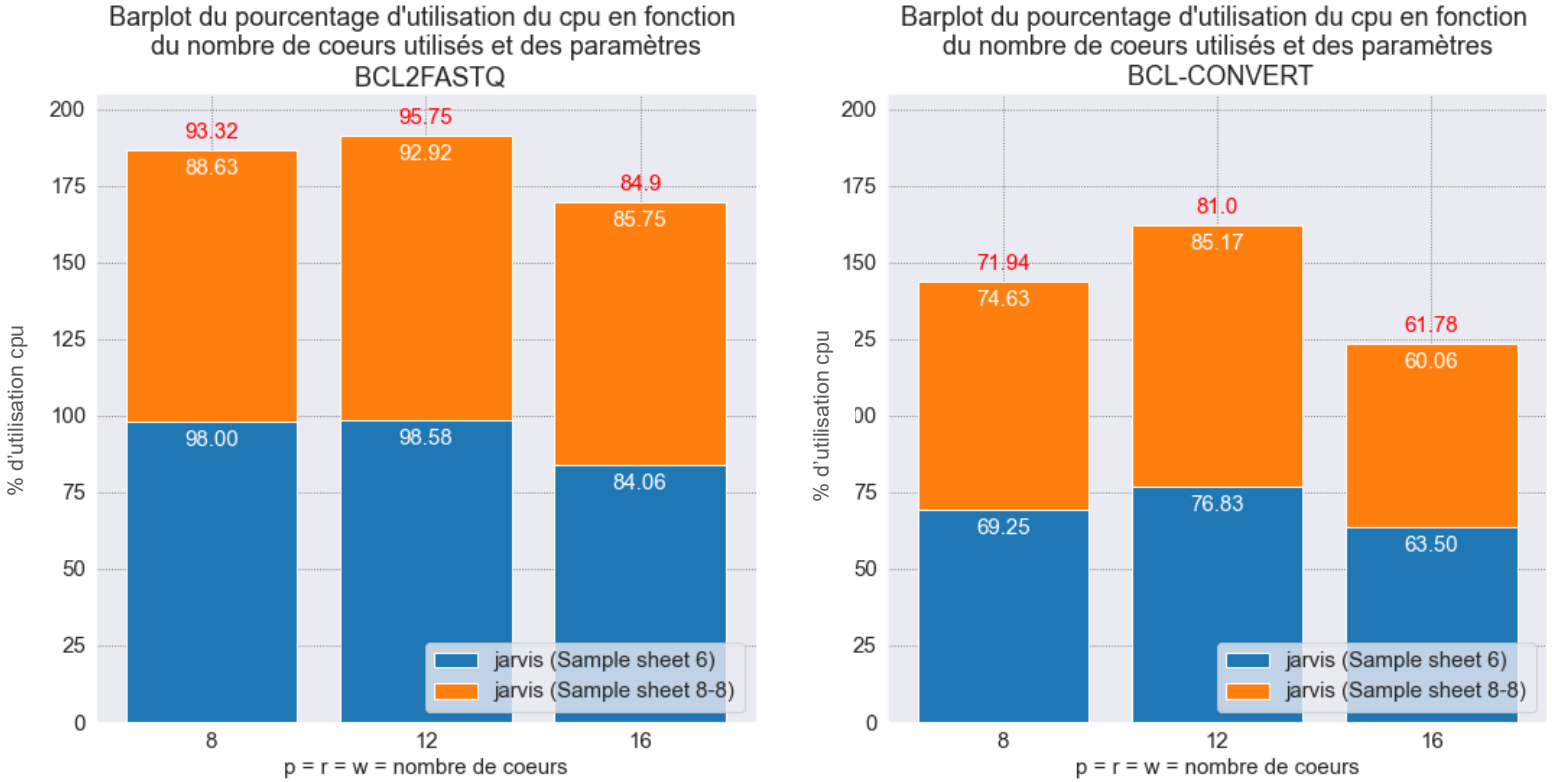
- p : nombre de cœurs pour la conversion des Base Calls en FASTQ et le démultiplexage
- r : nombre de cœurs pour la décompression et la lecture des Base Calls
- w : nombre de cœurs pour l'écriture et la compression des FASTQ

Figure 1 – Comparaison des performances en temps cpu des logiciels bcl2fastq et bcl-convert

Logiciels permettant la génération des FASTQ et de réaliser le démultiplexage développé par Illumina

Comparaison des performances entres les 2 logiciels

- Temps total d'exécution
- Temps cpu
- Pourcentage d'utilisation des cpu



- p : nombre de cœurs pour la conversion des Base Calls en FASTQ et le démultiplexage
- r : nombre de cœurs pour la décompression et la lecture des Base Calls
- w : nombre de cœurs pour l'écriture et la compression des FASTQ

Figure 1 – Comparaison des performances en pourcentage d'utilisation des cpu des logiciels bcl2fastq et bcl-convert

Préparation de la migration de bcl2fastq vers bcl-convert

Choix des paramètres à utiliser pour bcl-convert

- 16 cœurs sans spécifier les paramètres p, r et w car les nœuds de production font 16 cœurs

Cahier des charges de la migration de bcl2fastq vers bcl-convert

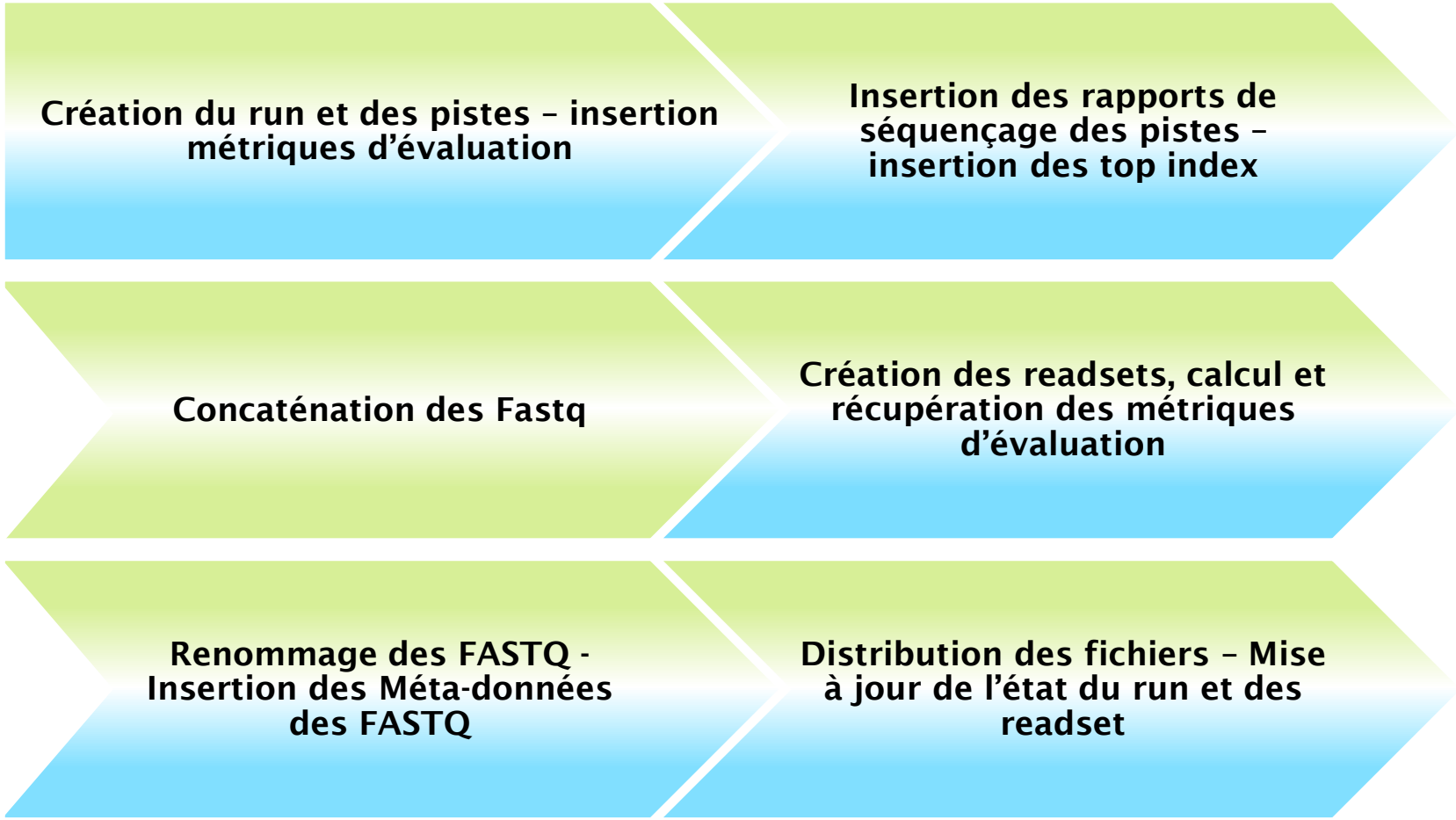
- Commande de bcl-convert à lancer pour la génération des FASTQ et le démultiplexage
- Les modules à charger dans l'environnement de travail
- Le chemin relatif des fichiers de sortie
- La description des fichiers de sortie (contenu, type de fichier …)
- Un exemple d'arborescence de fichier de sortie

Pipeline NGS-RG pour les séquenceurs MGI

1 script Perl

Qui fait appel à :

- 14 librairies de traitements de run MGI
- 3 librairies communes à tous les traitements de run MGI
- 11 librairies d'intéraction avec NGL pour les run MGI
- 1 librairie commune à tous les type de run



Création et insertion des métriques d'évaluation du run et des piste | Insertion des rapports de séquençage des pistes – insertion des top index | Concaténation des Fastq | Création des readsets, calcul et récupération des métriques d'évaluation | Renommage des FASTQ - Insertion des Méta-données des FASTQ | Distribution des fichiers – Mise à jour de l'état du run et des readset

Création et insertion des métriques d'évaluation du run et des pistes

Objectifs

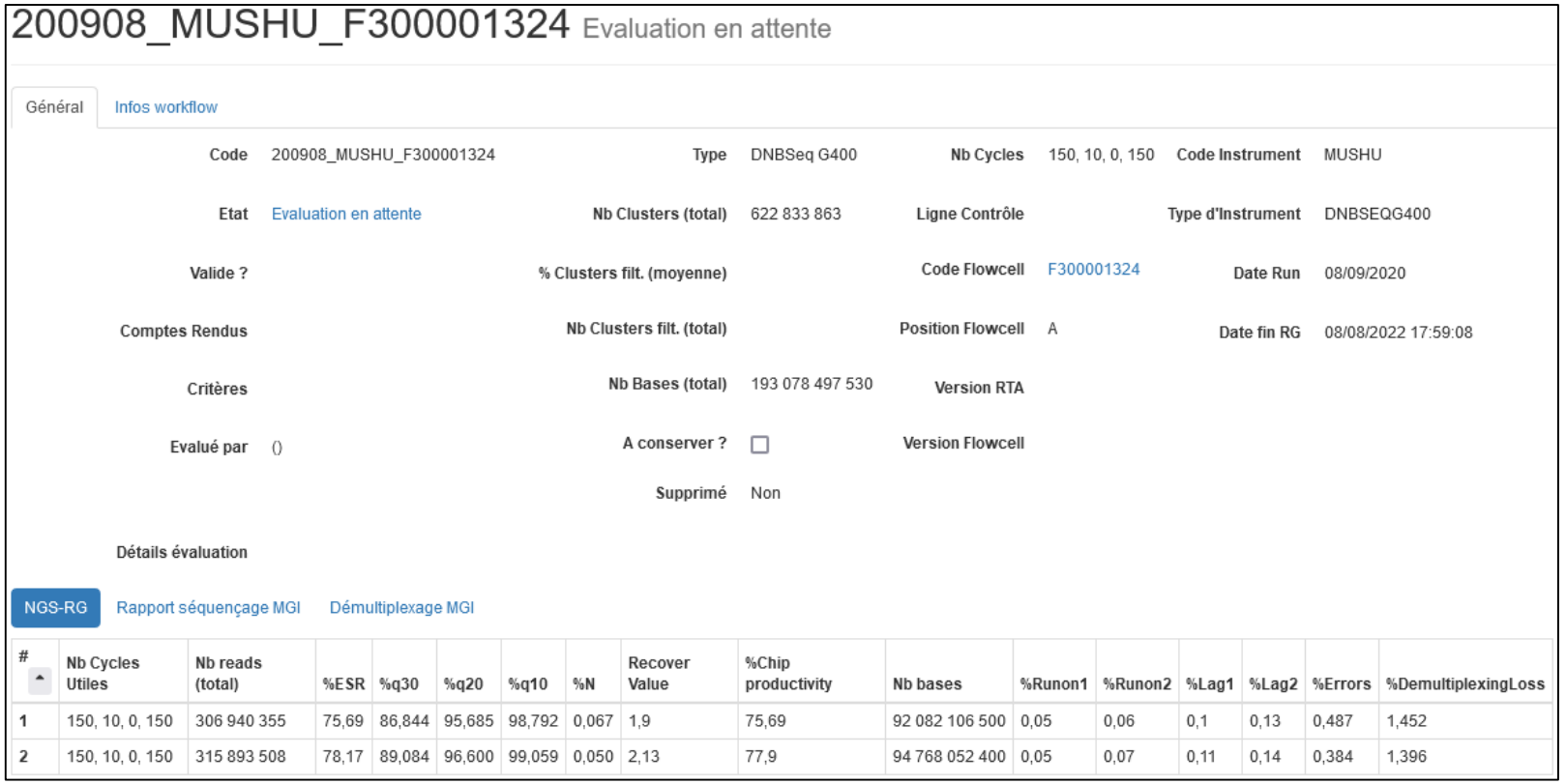
- Rendre disponible les informations à propos du run et des pistes, ainsi que l'état de traitement de ces derniers aux utilisateurs

Traitements

- Création d'un répertoire temporaire de traitement du run
- Récupération, calculs des métriques d'évaluation du run et des pistes

NGL

- Création du run et des pistes
- Insertion des métriques d'évaluation



200908_MUSHU_F300001324 Evaluation en attente

Général | Infos workflow

Code	200908_MUSHU_F300001324	Type	DNBSeq G400	Nb Cycles	150, 10, 0, 150	Code Instrument	MUSHU
Etat	Evaluation en attente	Nb Clusters (total)	622 833 863	Ligne Contrôle		Type d'Instrument	DNBSEQG400
Valide ?		% Clusters filt. (moyenne)		Code Flowcell	F300001324	Date Run	08/09/2020
Comptes Rendus		Nb Clusters filt. (total)		Position Flowcell	A	Date fin RG	08/08/2022 17:59:08
Critères		Nb Bases (total)	193 078 497 530	Version RTA			
Evalué par	()	A conserver ?	☐	Version Flowcell			
		Supprimé	Non				

Détails évaluation

NGS-RG | Rapport séquençage MGI | Démultiplexage MGI

#	Nb Cycles Utiles	Nb reads (total)	%ESR	%q30	%q20	%q10	%N	Recover Value	%Chip productivity	Nb bases	%Runon1	%Runon2	%Lag1	%Lag2	%Errors	%DemultiplexingLoss
1	150, 10, 0, 150	306 940 355	75,69	86,844	95,685	98,792	0,067	1,9	75,69	92 082 106 500	0,05	0,06	0,1	0,13	0,487	1,452
2	150, 10, 0, 150	315 893 508	78,17	89,084	96,600	99,059	0,050	2,13	77,9	94 768 052 400	0,05	0,07	0,11	0,14	0,384	1,396

Création et insertion des métriques d'évaluation du run et des piste | Insertion des rapports de séquençage des pistes – insertion des top index | Concaténation des Fastq | Création des readsets, calcul et récupération des métriques d'évaluation | Renommage des FASTQ - Insertion des Méta-données des FASTQ | Distribution des fichiers – Mise à jour de l'état du run et des readset

Insertion des rapports de séquençage des pistes – insertion des top index

Objectifs

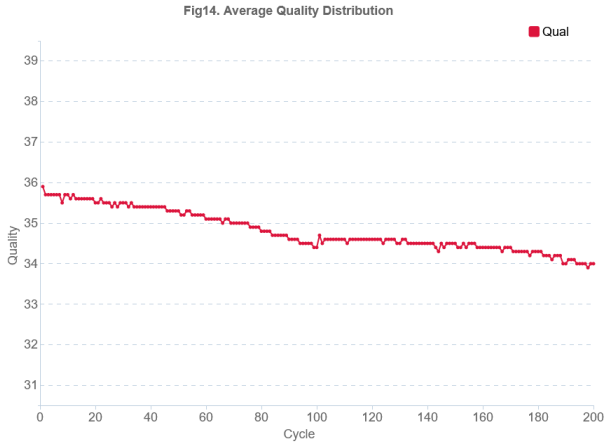
- Permettre l'évaluation des pistes

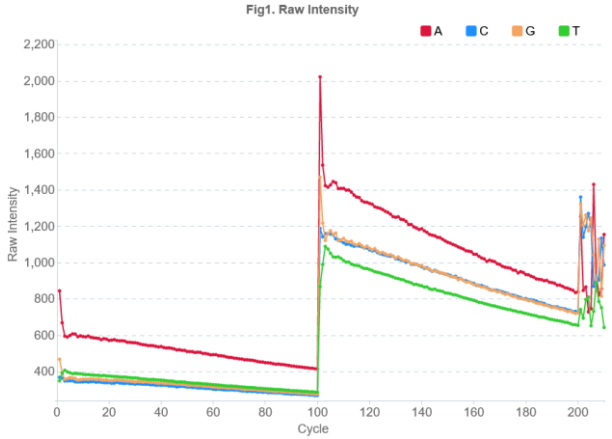
Traitements

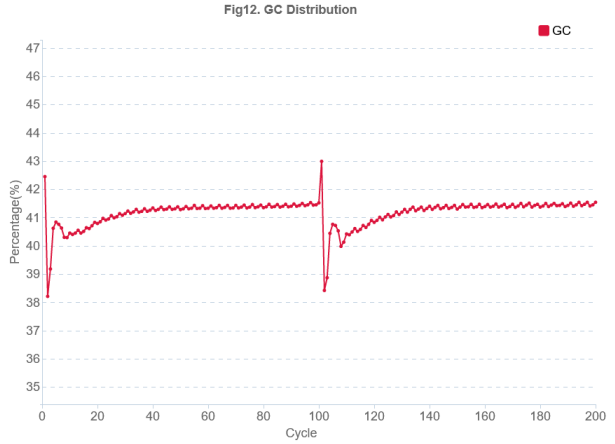
- Récupération des rapport de séquençage des piste
- Récupération des index représenté à plus de 0.01% de la pistes et des index attendus triés par ordre décroissant

NGL

- Insertion des rapport de séquençage des pistes
- Insertion top index par piste







Lane 1

barcode	count	percent
barcode2	89 597 340	29,190
barcode1	84 106 886	27,402
barcode3	74 172 719	24,165
barcode4	54 607 003	17,791
GATTCGTCCT	206 151	0,067
barcode29	181 509	0,059
GATCCGTCCT	156 796	0,051
barcode124	156 103	0,051

Concaténation des FASTQ

La technologie MGI requiert une homogénéité des bases pour chaque cycle des index

Le démultiplexage génère un FASTQ par index connu

Un échantillon peut être divisé en plusieurs fichiers

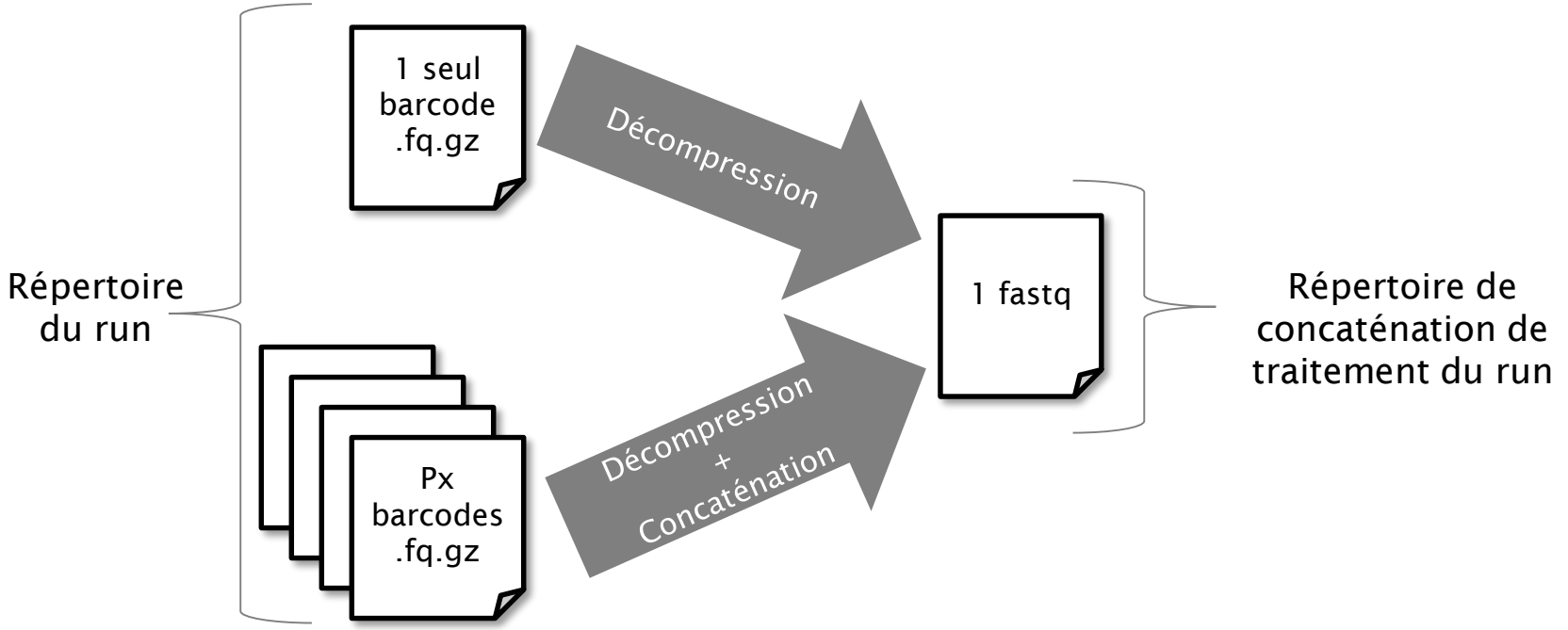
Objectifs

Obtenir un seul FASTQ par readset

Traitements

- Si un seul index :
  - Décompression et renommage du FASTQ
- Si plusieurs index :
  - Décompression, concaténation et renommage des FASTQ

La décompression et la concaténation est réalisé avec **unpigz** sur 2 threads



Création des readsets, calcul et récupération des métriques d'évaluation

Objectifs

Permettre l'évaluation des readsets

Traitements (3 traitements)

- NGSRG
  - Nombre de reads
  - Nombre de bases
  - Qualité moyenne
  - Etc.
- Global (sera mis à jour par NGS-QC)
  - Nombre de reads
  - Nombre de bases

NGL

- Création des readsets
- Insertion des métriques d'évaluation des readsets

Renommage et Insertion des méta-données des FASTQ

Objectifs

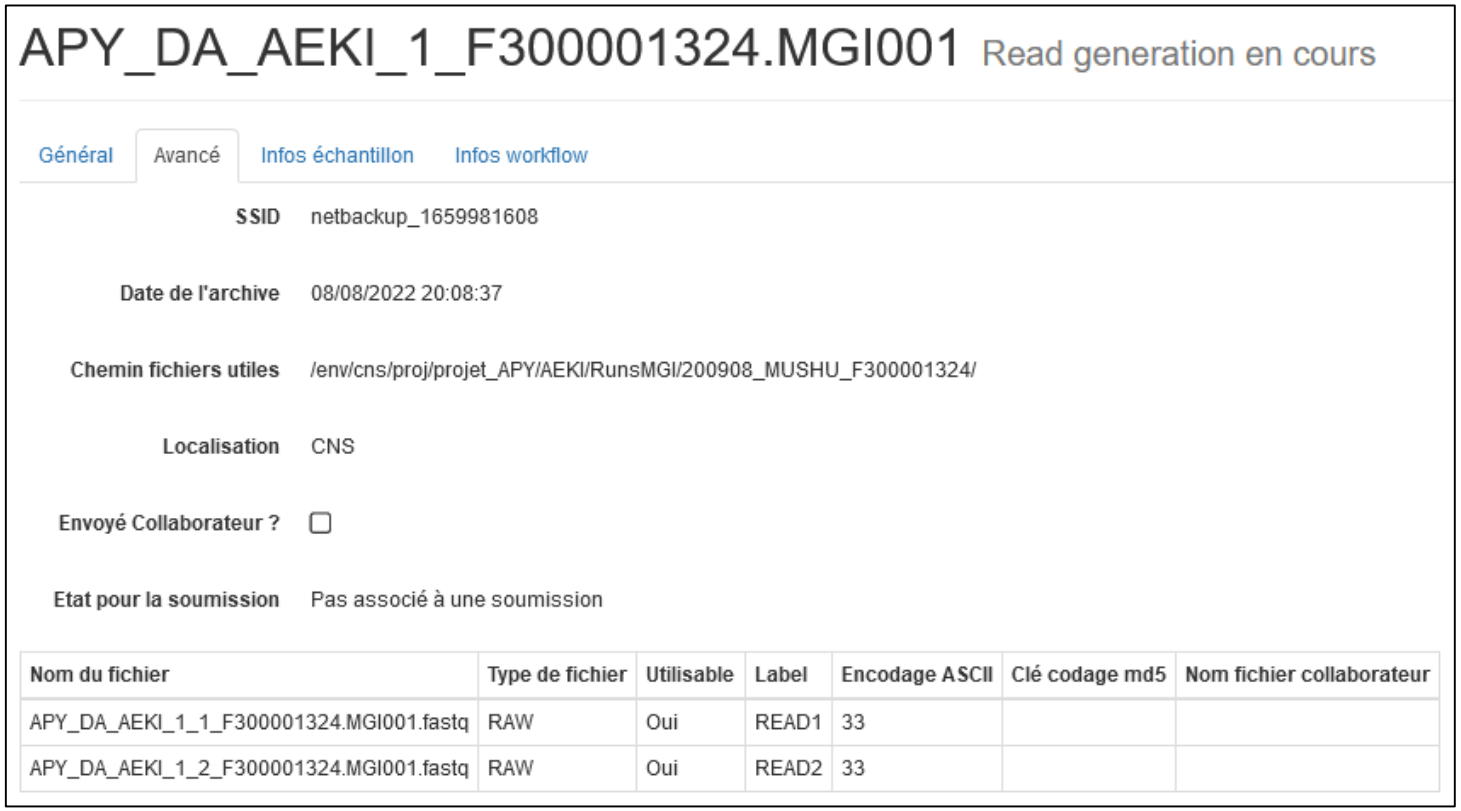
- Décrire les fichiers disponible pour un readset, ainsi que leurs emplacement dans le système de fichers
- Avoir des nom unique et « parlant »

Traitements

- Récupération de l'extension et du type d'encodage de la qualité
- Construction du chemin du répertoire des fichiers et du label
- Renommage des Fastq selon le format utilisé au Genoscope et CNRGH

NGL

- Insertion des méta-données des FASTQ de chaque readsets



APY_DA_AEKI_1_F300001324.MGI001 Read generation en cours

Général | Avancé | Infos échantillon | Infos workflow

SSID netbackup_1659981608

Date de l'archive 08/08/2022 20:08:37

Chemin fichiers utiles /env/cns/proj/projet_APY/AEKI/RunsMGI/200908_MUSHU_F300001324/

Localisation CNS

Envoyé Collaborateur ? ☐

Etat pour la soumission Pas associé à une soumission

Nom du fichier	Type de fichier	Utilisable	Label	Encodage ASCII	Clé codage md5	Nom fichier collaborateur
APY_DA_AEKI_1_1_F300001324.MGI001.fastq	RAW	Oui	READ1	33		
APY_DA_AEKI_1_2_F300001324.MGI001.fastq	RAW	Oui	READ2	33		

Création et insertion des métriques d'évaluation du run et des piste | Insertion des rapports de séquençage des pistes – insertion des top index | Concaténation des Fastq | Création des readsets, calcul et récupération des métriques d'évaluation | Renommage des FASTQ - Insertion des Méta-données des FASTQ | Distribution des fichiers – Mise à jour de l'état du run et des readset

Distribution des fichiers - Mise à jour de l'état du run et des readsets

Objectifs

- Rendre disponible les fichiers de séquences
- Conserver les fichiers de statistique du run
- Conserver les fichiers de séquences non-attendus
- Indiquer que l'évaluation du run peut être réalisé
- Indiquer au pipeline NGS-QC qu'il peut réaliser le contrôle qualité des readsets

Traitements

- Distribution des fichiers de séquences dans leurs répertoires dédiés en changeant les droits d'accès
- Archivage des fichiers de statistique par type (.html, .fq.stat …) avant de les distribuer dans leur répertoire dédié
- Renommage et archivage des fichiers de séquences non-attendus par pistes avant de les distribuer dans leurs répertoires dédiés

Mise à jour NGL

- Mise à jour du run et des readset en cascade à « Fin de génération de reads »
  - Mise à jour automatique du run à « Evaluation en attente »
  - Mise à jour automatique des readset à « Contrôle qualité en attente »

Développements terminés ou en cours

Pipeline NGS RG MGI

- ✓ Mise à jour pour les séquenceurs MGI du CNRGH
- ✓ Mise à jour pour le séquenceur MGI DNBT7 (séquenceur à très haut débit)
- ✓ Mise en production du pipeline pour le Genoscope et le CNRGH

Pipeline NGS QC MGI

- ❑ Développement des librairies pour les traitements du pipeline contrôle qualité
- ❑ Développement du pipeline de contrôle qualité
- ❑ Adaptation de certains scripts utilisés par les pipelines NGS QC pour la technologie MGI

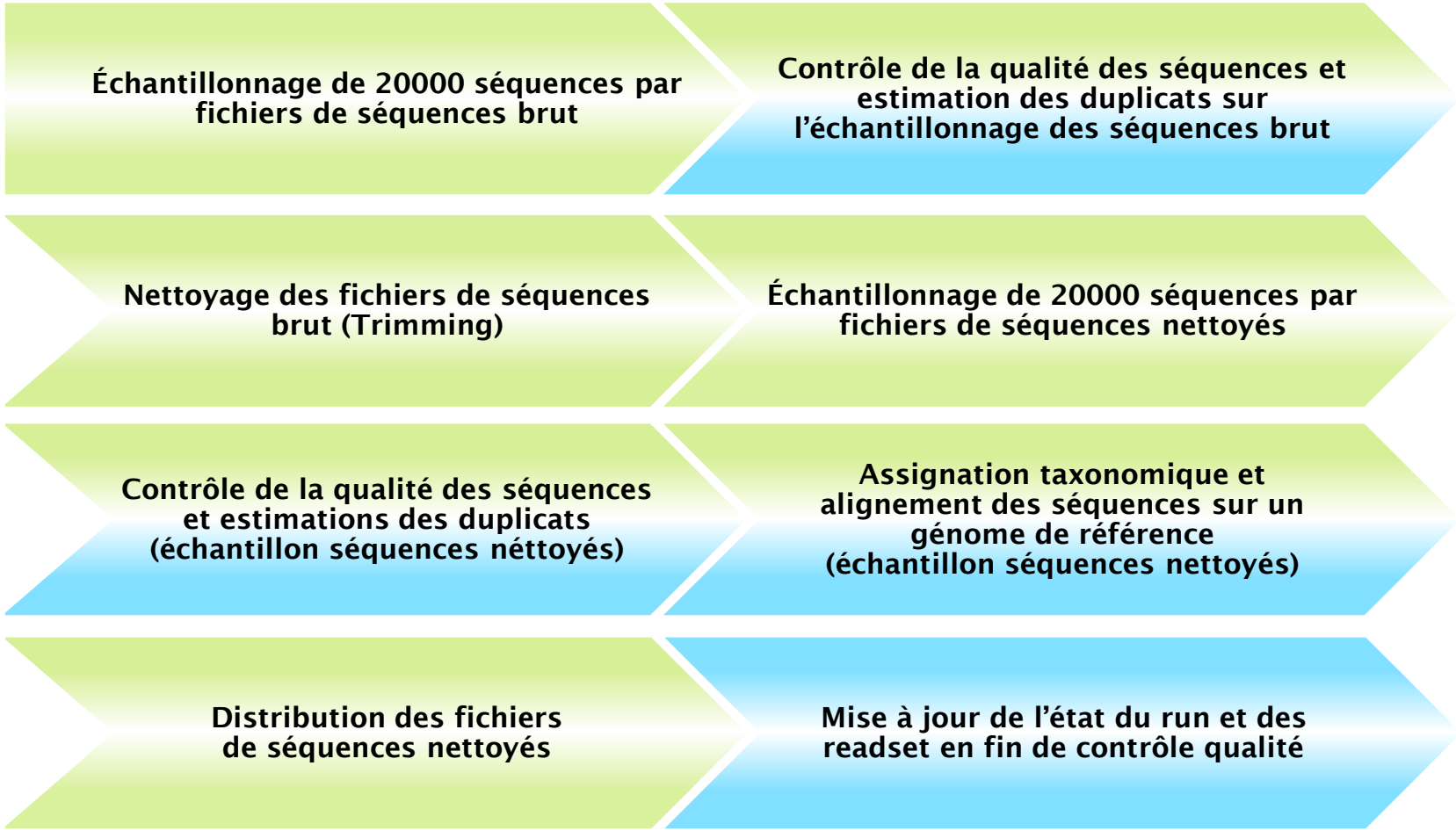
Développements terminés ou en cours

Le pipeline NGS QC MGI

1 script Perl

Qui fait appel à :

- 8 librairies de traitements de run MGI
- 3 librairies communes à tous les traitements de run MGI
- 6 librairies d'intéraction avec NGL pour les run MGI
- 3 librairie commune à tous les type de run



Perspectives

Workflow NGS MGI

Pipeline NGS-RG MGI

- Ajout du second démultiplexage (démidage) pour les run comportant des mids

Pipeline NGS-QC MGI

- Finir le développement du pipeline
- Mise en production du pipeline

Evaluation d'autres outils

- Outils d'assignation taxonomique par rapport à celui utilisé actuellement (Centrifuge)
- Outils de *trimming* par rapport à celui utilisé actuellement (fastx_clean de FASTX Toolkit)

Bibliographie

- Impact of sequencing depth and technology on de novo RNA-Seq assembly. Patterson. 2022-01-23, *BMC Genomics*. https://doi.org/10.1186/s12864-019-5965-x
- Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. Jeon, S.A., Park, J.L., Park, SJ. and al. *Genes Genom* **43,** 713–724 (2021). https://doi.org/10.1007/s13258-021-01096-x
- Best practices for the interpretation and reporting of clinical whole genome sequencing. Austin-Tse, C.A., Jobanputra, V., Perry, D.L. and al. *npj Genom. Med.* **7,** 27 (2022). https://doi.org/10.1038/s41525-022-00295-z
- Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. Hak-Min Kim and al. *GigaScience*, Volume 10, Issue 3, March 2021, giab014, https://doi.org/10.1093/gigascience/giab014
- Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms. Anslan S, Mikryukov V, and al. 2021. *PeerJ* 9:e12254 https://doi.org/10.7717/peerj.12254
- CoolMPS™: Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. Snezana Drmanac, Matthew Callow and al. *bioRxiv preprint*. https://doi.org/10.1101/2020.02.19.953307
- bcl2fastq2 Conversion Software v2.20 Software Guide (15051736). 2019, Illumina, Inc. 2022-01-23
- BCL Convert Software Guide v3.7.5 (1000000163594). 2021, Illumina, Inc. 2022-01-23
- 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

DE LA RECHERCHE À L'INDUSTRIE

cea

Université Paris Cité

Centre National de Séquençage

GENOSCOPE

www.cea.fr

Merci de votre attention

William Amory
M1 BI-IPFB Université Paris Cité

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

Sous la responsabilité de Frédérick Gavory

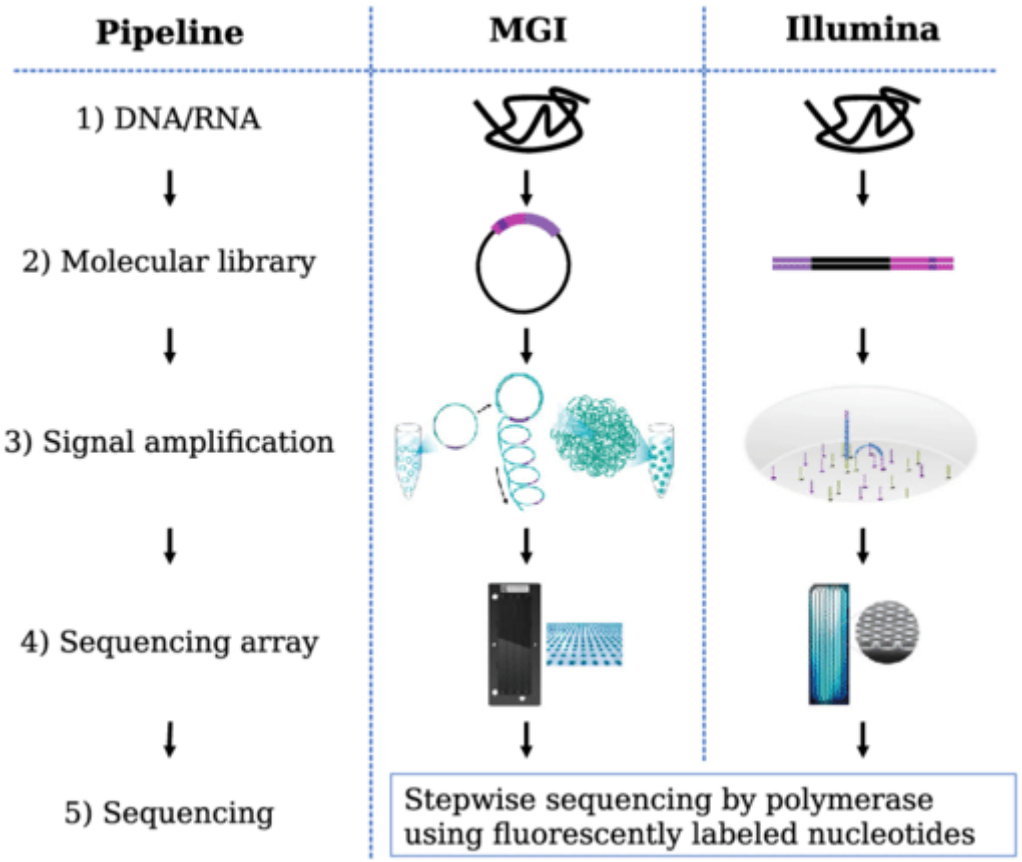


Figure 2 – Différences entre Illumina et MGI de technologie NGL

J. Patterson & all. (2019). Impact of sequencing depth and technology on de novo RNA-Seq assembly. BMC Genomics. 20. 10.1186/s12864-019-5965-x.

DNB (DNA Nanoball)

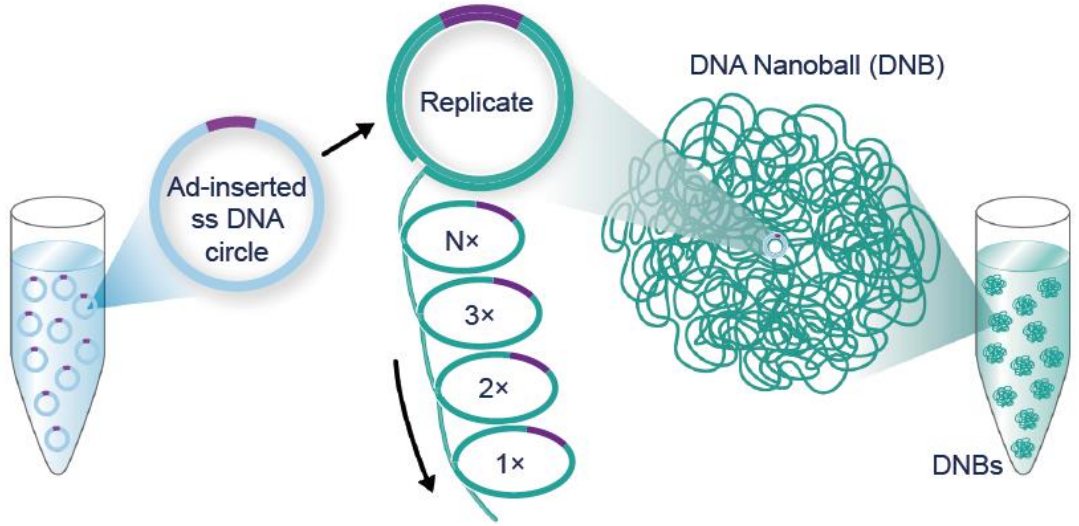


Figure 3 – Schéma de la technologie des *DNA nanoballs* de MGI

https://en.mgi-tech.com/products/ consulté le 21/06/2022

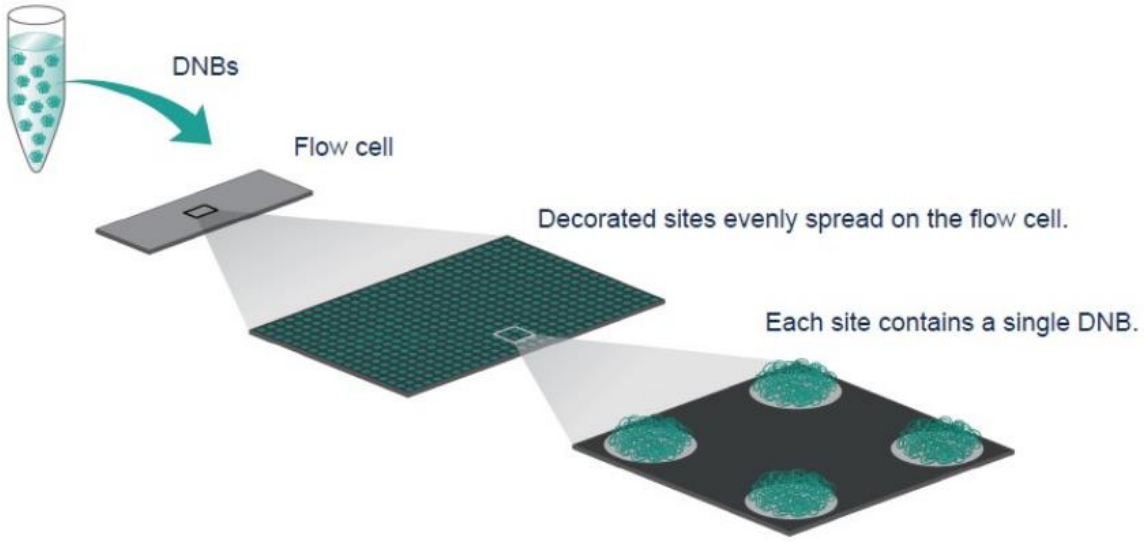


Figure 4 – Schéma d'une flowcell et des DNB dans les puits de la flowcell

https://en.mgi-tech.com/products/ consulté le 21/06/2022

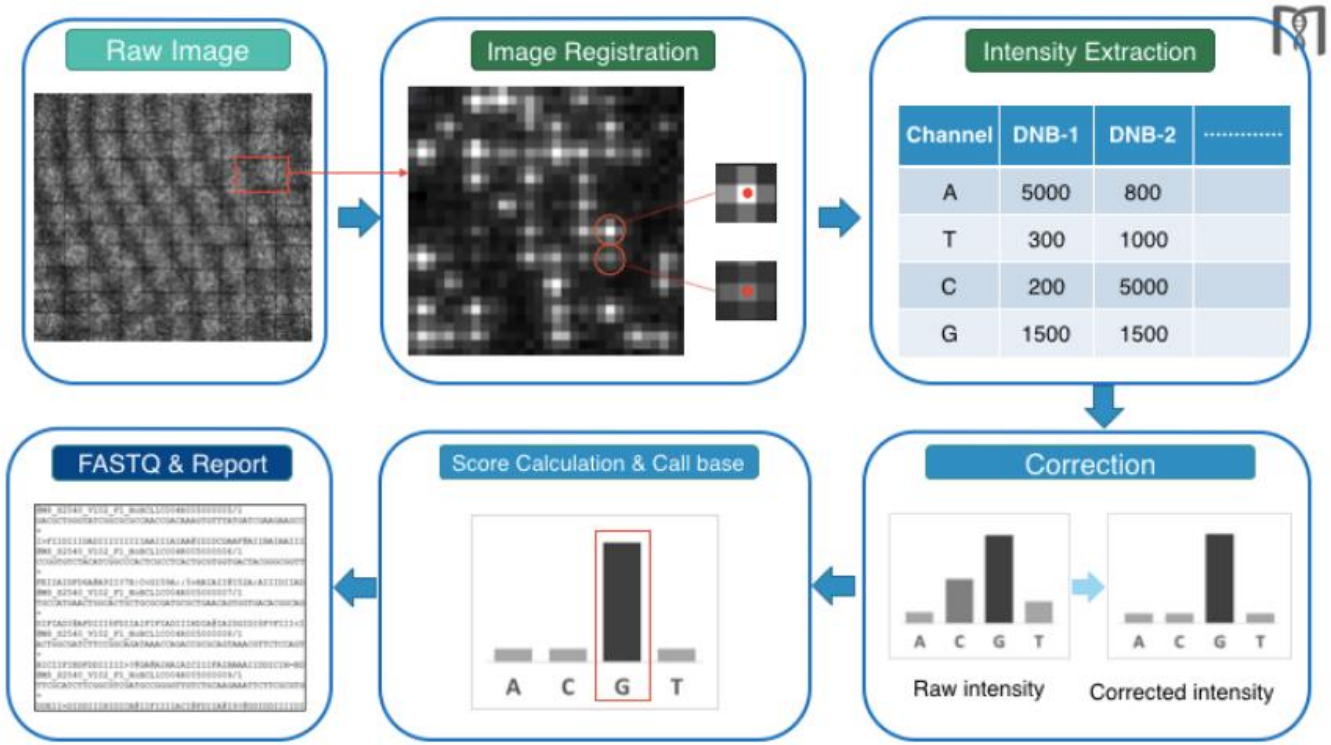


Figure 5 – Schéma de basecalling des séquenceurs MGI

https://en.mgi-tech.com/products/ consulté le 21/06/2022

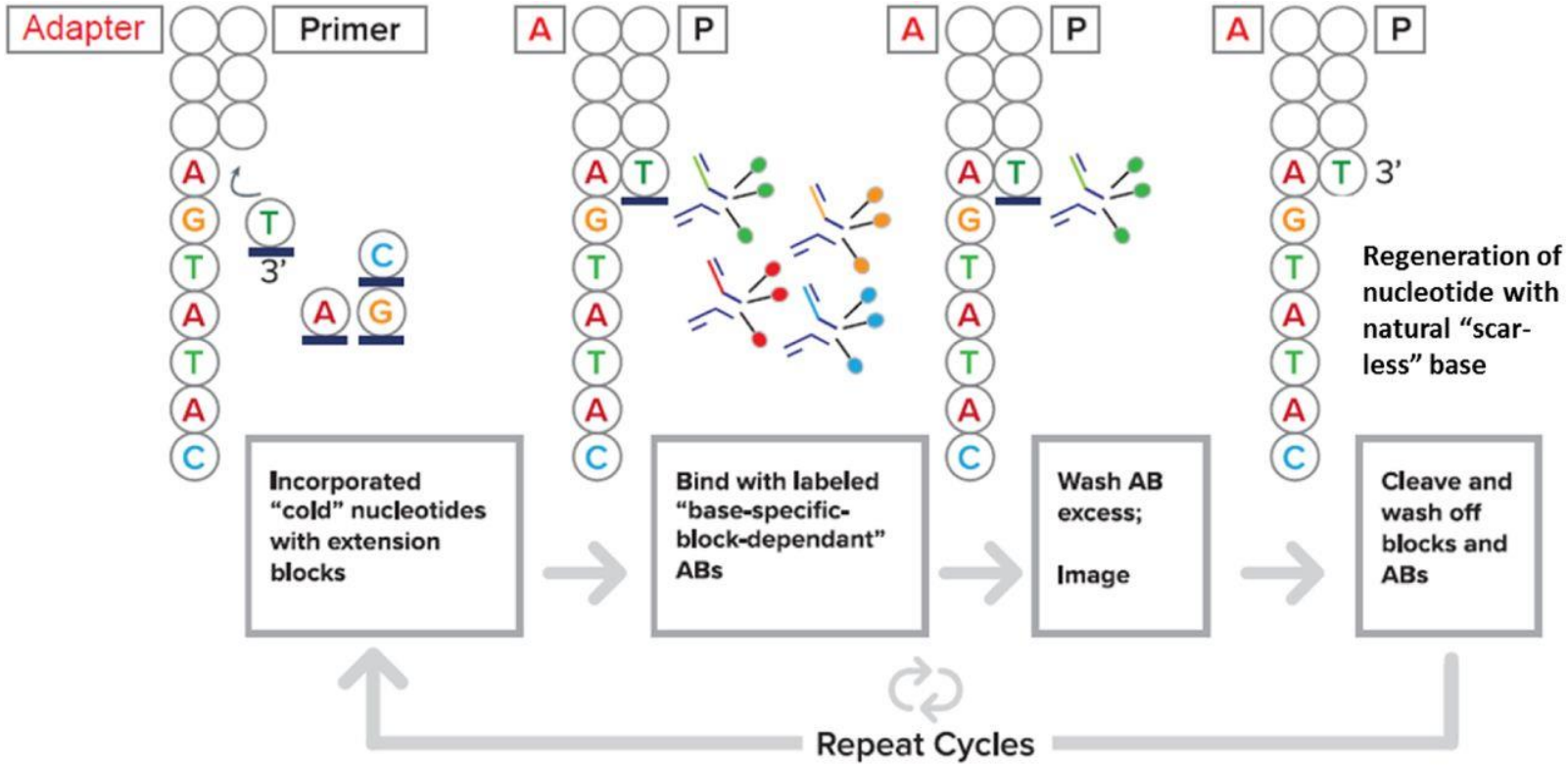
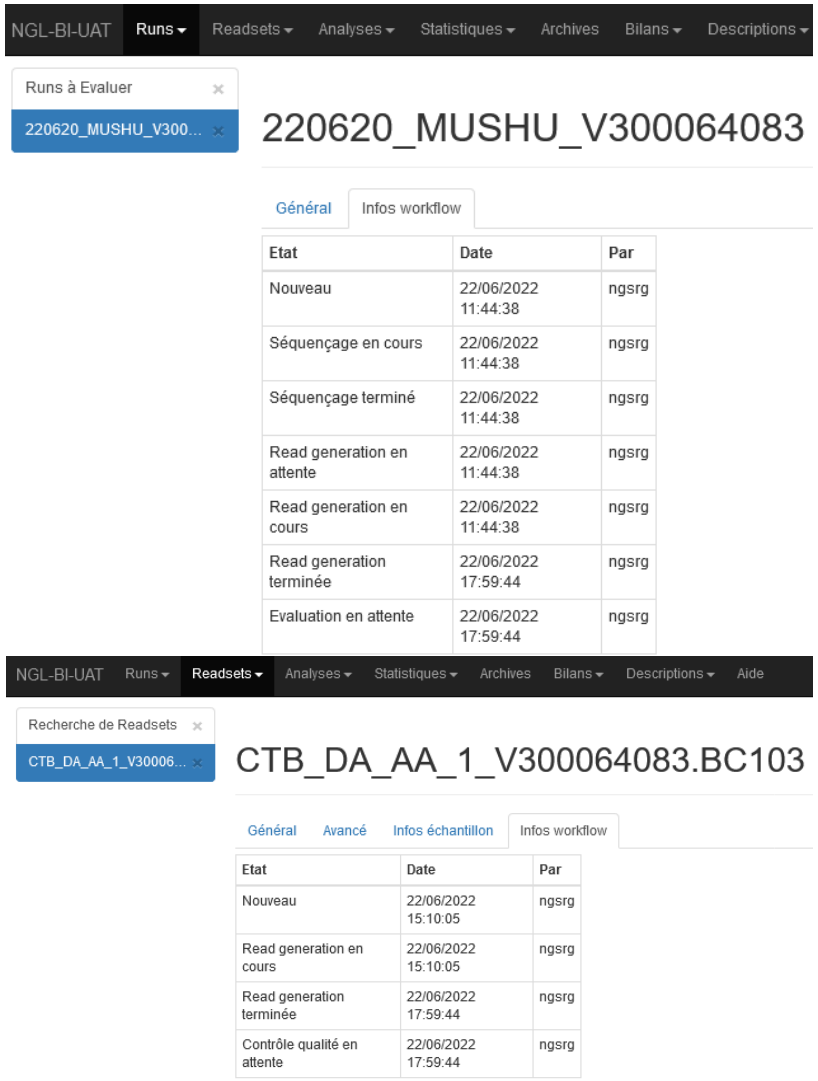


Figure 6 – Schéma de la nouvelle chimie MGI : CoolMPS

CoolMPS™: Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. Snezana Drmanac, Matthew Callow and al. bioRxiv preprint. https://doi.org/10.1101/2020.02.19.953307

Infos workflow du run et des readsets



Infos échantillon