A partir de données brutes de séquençage, le plateau technique propose l'assemblage *de novo* de génome bactérien en utilisant des données *long-read* + *short-read* ou *long-read* uniquement.

Prestation proposée

A partir des données brutes de séquençage des échantillons d'ADN d'intérêt, le plateau technique réalise les étapes de :

Production des fichiers fastq et démultiplexage des données

Le démultiplexage et la production des fichiers fastq pour les données Illumina sont réalisés grâce au logiciel Illumina bclconvert.

Le *base calling*, le démultiplexage et la production des fichiers fastq pour les données ONT sont réalisés grâce au logiciel d'Oxford Nanopore Technologies Dorado.

Contrôle qualité des données

Le contrôle qualité s'appuie sur plusieurs critères :

- validation du run en utilisant une série de critères associés (Illumina + ONT),
- distribution des scores de qualité à chaque cycle (Illumina),
- distribution des scores moyens de qualité par séquence (Illumina),
- pourcentage de bases "N" par cycle (Illumina),
- recherche de contaminants (Illumina),
- distribution de la qualité des reads (ONT),
- distribution de la taille des reads (ONT).

Cette étape est réalisée systématiquement. Les étapes suivantes sont réalisées sur demande.

Assemblage *de novo*

Suivant le type de données, différents outils sont utilisés. Dans tous les cas, les étapes suivantes sont réalisées :

- Assemblage des données
- *Polishing* de l'assemblage brut
- Correction de l'assemblage
- Contrôle qualité de l'assemblage
- Annotation de l'assemblage (contextualisation uniquement)

+ Assemblage hybride *short-read* / *long-read* (Illumina/ONT)

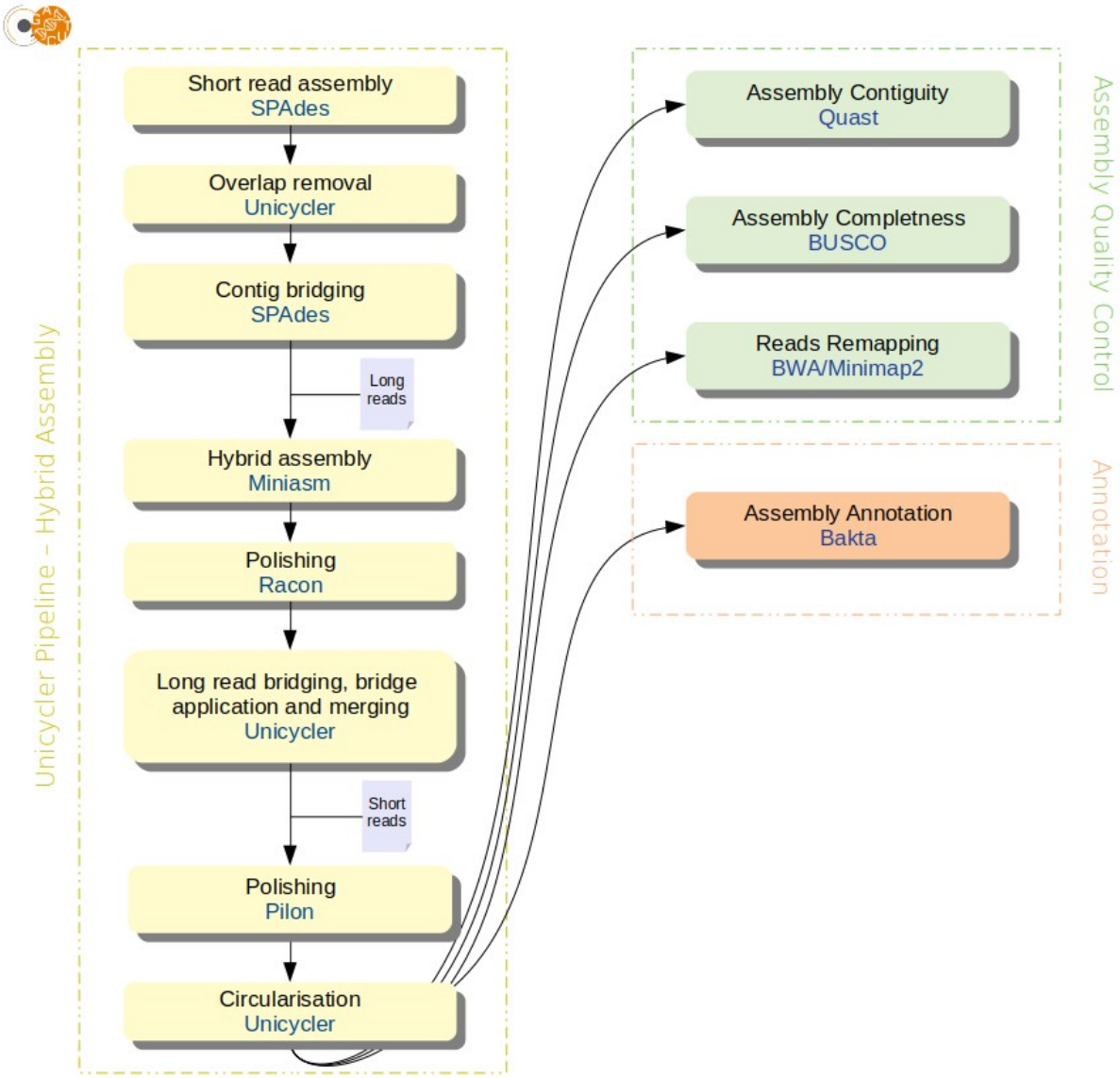

Figure 1 - Vue d'ensemble des étapes d'assemblage hybride

Unicycler [1] est un *pipeline* spécifiquement destiné à l'assemblage *de novo* de génomes bactériens. Il est capable d'assembler soit des *short-reads* issus de séquençages Illumina, soit des *long-reads* issus de séquençages PacBio ou Oxford Nanopore, soit de combiner les deux types de reads pour réaliser un assemblage hybride (le dernier cas étant celui recommandé par les développeurs du *pipeline*).

Pour l'assemblage hybride, Unicycler va dans un premier temps assembler les *short-reads* avec SPAdes [2], puis les contigs générés seront assemblés avec les *long-reads* à l'aide de Miniasm [3] + Racon [4]. Ensuite, les *long-reads* sont alignés sur le graphe d'assemblage pour créer des ponts entre paires de contigs. Pour terminer, l'assemblage est "polishé" avec l'outil Pilon [5] en utilisant les *short-reads,* puis circularisé.

[1] Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology 13(6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595
[2] Bankevich, Anton; Nurk, Sergey; Antipov, Dmitry; Gurevich, Alexey A.; Dvorkin, Mikhail; Kulikov, Alexander S.; Lesin, Valery M.; Nikolenko, Sergey I.; Pham, Son; Prjibelski, Andrey D. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021
[3] Heng Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics, Volume 32, Issue 14, 15 July 2016, Pages 2103–2110. https://doi.org/10.1093/bioinformatics/btw152
[4] Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737-746. doi:10.1101/gr.214270.116
[5] Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963. \href{doi:10.1371/journal.pone.0112963

+ Assemblage *long-read* (ONT)

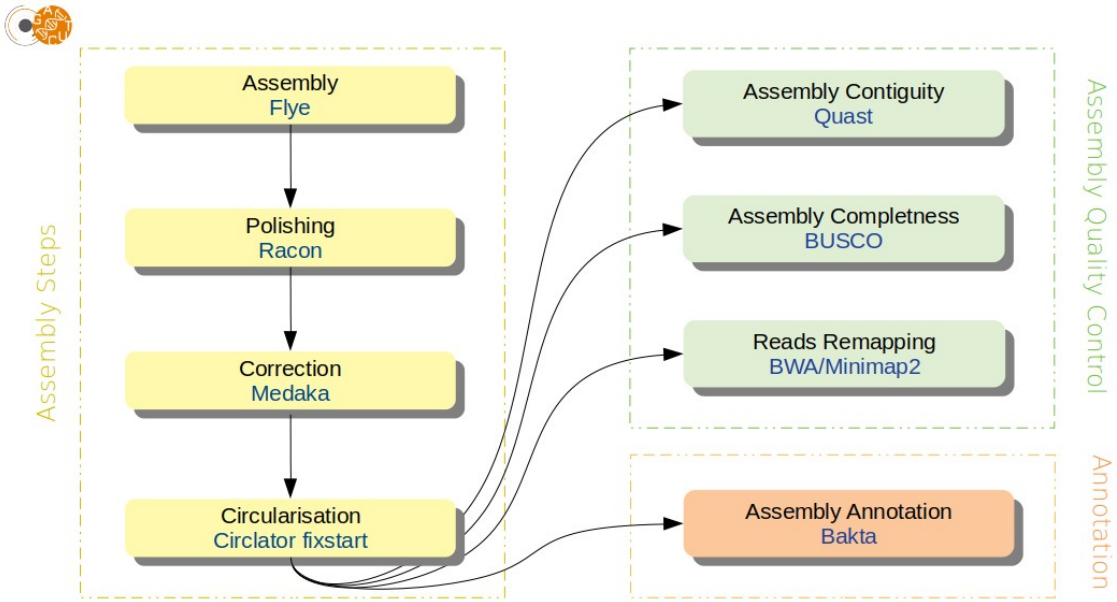

Figure 2 – Vue d'ensemble des étapes d'assemblage avec uniquement des *long-reads*

+ Assemblage avec Flye :

Flye [6] est un assembleur *de novo* dédié à l'assemblage de reads issus de séquençages ONT ou PacBio. Il est capable d'assembler autant des génomes de petite taille (bactériens) que des génomes de plus grande taille (mammifères, plantes, ...) ou des métagénomes.

La différence de cet outil par rapport aux autres assembleurs est que son algorithme se base sur un *repeat graph* au lieu des méthodes classiques du graphe de De Bruijn ou OLC (*Overlap Layout Consensus*). La particularité de ce graphe est qu'il est construit à partir de matches partiels entre kmers (les algorithmes basés sur le graphe de De Bruijn nécessitent des matches exacts), ce qui est particulièrement adapté aux *long-reads* qui sont de moindre qualité par rapport aux *short-reads*.

+ *Polishing* avec Racon et correction avec Medaka :

Racon [7] est un outil qui permet de faire un *polishing* d'assemblages réalisés, entre autre, à partir de *long-reads* non corrigés. Pour cela, il a besoin uniquement des reads, de l'assemblage et des *overlaps* entre les reads et les contigs assemblés. Le but est de générer un consensus de qualité similaire ou supérieure à celle des contigs bruts.

Medaka est un outil de correction développé par ONT. Il utilise pour cela un algorithme basé sur les réseaux de neurones qu'il applique à l'alignement des *long-reads* sur l'assemblage.

+ Correction de la position de début de circularisation avec Circlator *fixstart* :

Circlator est un outil permettant de circulariser des assemblages. En sachant que Flye arrive en général à générer des molécules circulaires pour les génomes bactériens, seuls le module fixstart est utilisé. Ce module a pour but de fixer le début d'un contig de manière à ce que ce soit la séquence du gène DnaA, s'il est retrouvé.

[6] Yu Lin, Jeffrey Yuan, Mikhail Kolmogorov, Max W Shen, Mark Chaisson and Pavel Pevzner, "Assembly of Long Error-Prone Reads Using de Bruijn Graphs", PNAS, 2016 doi:10.1073/pnas.1604560113
[7] Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017 May;27(5):737-746. doi: 10.1101/gr.214270.116. Epub 2017 Jan 18. PMID: 28100585; PMCID: PMC5411768.

Contrôle qualité de l'assemblage généré

Le contrôle qualité s'appuie sur les métriques générées à l'aide des outils suivants :

+ QUAST :

QUAST est un outil qui permet d'évaluer un assemblage en générant différentes métriques telles que :

- la taille totale de l'assemblage,
- le nombre de contigs,
- la N50,
- la NG50 (si un génome de référence est disponible).

+ BUSCO :

BUSCO (*Benchmarking Universal Single-Copy Orthologs*) est un outil permettant d'évaluer l'intégrité d'un assemblage, d'un ensemble de gènes ou d'un transcriptome. Il se base sur le fait que des gènes orthologues unie-copies devraient en principe êtres hautement conservés entre espèces d'un même groupe taxonomique. Pour cela, BUSCO dispose de plusieurs *datasets* contenant un ensemble de gènes issus de la base de donnée OrthoDB, qui regroupe des orthologues présents en une seule copie dans plus de 90% des espèces.

+ Réalignement des reads sur l'assemblage :

Réaligner les reads sur un assemblage est un moyen de détecter les *misassemblies*. En effet si une zone de l'assemblage n'est pas couverte par les reads, il est très probable que cette zone soit incorrectement assemblée.

Le réalignement des reads se fait avec minimap2 pour les *long-reads* et bwa-mem pour les *short-reads*.

Prestations complémentaires

Des analyses complémentaires peuvent être effectuées sur demande (bioinfo niveau 3).

Annotation de l'assemblage

Bakta est un outil d'annotation "généraliste" de génomes bactériens, c'est à dire qu'il annote un génome indépendamment de sa taxonomie. Il utilise différents outils de prédiction de gènes et autres *features* tels que Prodigal, tRNAscan-SE, Aragorn, Infernal, Diamond ou encore Blast+. Il utilise également une base de données *custom* contenant un ensemble d'acides aminés et séquences d'ADN issues de la base de données UniProt (clusters de protéines UniRef100 et UniRef90).

Les éléments suivants sont annotés :

- CDS
- rRNA
- tRNA
- tmRNA
- ncRNA (gènes)
- ncRNA (régions cis-régulatoires)
- CRISPR *arrays*
- ncRNA cis-regulatory regions
- oriC/oriV/oriT
- sORF (*small proteins/short open reading frames*)

Matériel initial

Les données nécessaires à l'analyse bioinformatique sont directement issues du séquenceur du plateau technique.

Restitution des résultats

A l'issue des analyses, plusieurs fichiers sont disponibles :

– 1 rapport d'analyse au format PDF disponible depuis le logiciel de gestion de projet redmine

Pour chaque échantillon (*) :

– *.fastq : fichier texte contenant les séquences nucléotidiques ainsi que les scores de qualité correspondants
– 1 fichier fasta contenant le génome assemblé

Pour chaque échantillon (*), dans le cas avec annotation :

- *.fastq : fichier texte contenant les séquences nucléotidiques ainsi que les scores de qualité correspondants
- 1 fichier *.fasta contenant le génome assemblé
- 1 archive *.tar.gz contenant les résultats de l'annotation

L'ensemble de ces fichiers est disponible sur le serveur SFTP du plateau, à partir de la mise en ligne du rapport sur le gestionnaire de projet. Ce serveur est accessible par login et mot de passe, fournis avec le rapport d'analyse.

Durée de conservation des données

Les login et mot de passe pour accéder au gestionnaire de projet et les documents qui y sont mis en ligne n'ont pour l'instant aucune limite de validité.

En revanche, les fichiers qui sont mis en ligne sur le serveur sftp n'y sont hébergés que pour une durée de 10 jours à compter de l'édition du rapport de résultat.

Il est conseillé de bien sauvegarder les données sur un disque dur externe ; nous ne nous engageons pas à conserver ces données.

Le dépôt des données brutes dans une base de données peut être demandé avant publication.