La technique d'analyse BRB-seq permet de réaliser une étude qualitative et quantitative des différents transcrits d'un ensemble d'échantillons. Elle présente un avantage sur le RNA-seq en terme de coût, puisqu'une seule banque est produite après ajout de barcodes aux échantillons d'ARN d'intérêt.

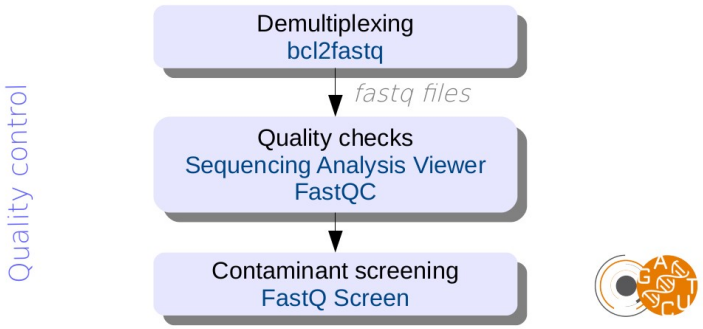
Prestation proposée

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

Production des fichiers fastq du pool de données

Le démultiplexage et la production des fichiers fastq est réalisée grâce au logiciel Illumina BCL-Convert. Il permet d'obtenir un couple de fichiers fastq par pool BRB-seq, le Read 1 correspondant au séquençage du barcode suivi de l'UMI, et le Read 2 correspondant au séquençage du cDNA.

Contrôle qualité des données



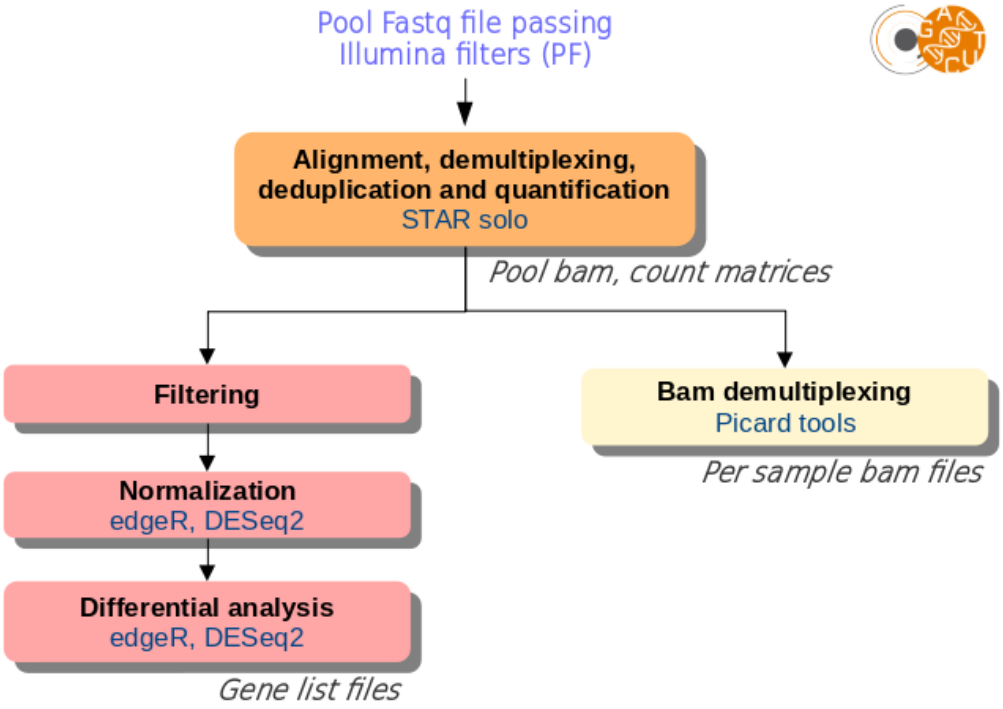
Le contrôle qualité s'appuie sur plusieurs critères :

- validation du run en utilisant une série de critères associés,
- distribution des scores de qualité à chaque cycle,
- distribution des scores moyens de qualité par séquence,
- pourcentage de bases "N" par cycle,
- distribution des bases par cycle,
- proportion d'adaptateurs séquencés,
- recherche de contaminants.

Dans le cas où le pool de banque est construit par le plateau technique, l'étape d'alignement et de démultiplexage par échantillon est réalisée afin de vérifier l'équilibre des différents échantillons au sein du pool (voir section suivante).

Le contrôle qualité des données est réalisé systématiquement (bioinfo niveau 1). Les étapes suivantes sont réalisées sur demande (bioinfo niveau 2).

Alignement, comptages et analyses statistiques



Alignement, comptage et déduplication des *reads*

Il est nécessaire que le génome soit disponible (format fasta), ainsi que son annotation (format gtf / gff) pour pouvoir procéder à cette étape.
Les étapes d'alignement, de démultiplexage, de déduplication et de comptage sont réalisées simultanément à l'aide de STARsolo (inclus dans le logiciel STAR).
STARsolo a été conçu pour traiter les données de *single-cell* RNA-seq, qui ont une structure similaire au BRB-seq : un Read 1 portant un barcode suivi d'un UMI, et un Read 2 contenant la séquence du cDNA. Ainsi, il permet de réaliser simultanément les étapes :

- d'alignement (grâce à la méthode d'alignement classique du logiciel STAR),
- de démultiplexage selon le barcode de l'échantillon,
- de déduplication à l'aide des UMI,
- de quantification des gènes (avant ou après déduplication), en comptant les reads / UMI alignés sur les exons des gènes, et en tenant compte de l'orientation des reads.

On obtient en sortie un fichier au format bam contenant les reads alignés pour l'ensemble du *pool* d'échantillons, ainsi que les matrices de comptages, par read brut (sans déduplication) ou par UMI (résultat de la déduplication).

Le format bam permet de visualiser l'alignement RNA-Seq avec un logiciel dédié comme IGV.

Démultiplexage du fichier bam

Afin d'obtenir un fichier d'alignement par échantillon, l'outil FilterSamReads de la suite Picard tools est utilisé.

Dans le fichier bam produit par STARsolo, un tag CR, qui contient la séquence du barcode de l'échantillon telle qu'elle a été lue lors du séquençage (sans correction), est associé à chaque alignement. L'outil Picard FilterSamReads permet de séparer les alignements selon ce tag et produit ainsi un fichier bam par échantillon.

Il est ensuite possible d'obtenir un fichier fastq par échantillon à l'aide de l'outil bedtools bam2fastq. Ces fichiers peuvent être demandés lors du dépôt sur des bases de données publiques comme Gene Expression Omnibus (GEO).

Analyse différentielle

Ces analyses visent à rechercher les gènes différentiellement exprimés, c'est à dire qui ont un niveau d'expression significativement différent dans une condition par rapport à l'autre. Plusieurs packages R sont utilisés pour les analyses statistiques :

- edgeR (Robinson *et al*., 2010) : est basé sur un modèle linéaire généralisé (GLM). Une correction pour tests multiples (Benjamini-Hochberg) est réalisée pour contrôler le taux de faux positifs (FDR).
- DESeq2 (Love *et al*., 2014) : test basé sur un modèle linéaire généralisé (GLM). Une correction pour tests multiples (Benjamini-Hochberg) est réalisée pour contrôler le taux de faux positifs (FDR). DESeq2 utilise une estimation de la variance qui la rend moins permissive pour les grandes variabilités entre conditions. A l'opposé, quand la variabilité intra-condition est plus faible, DESeq2 semble être plus « confiante » et sélectionner des gènes qui ont un *fold-change* plus faible que ceux sélectionnés par edgeR.

Pour chaque package, un fichier Excel est généré avec la liste des gènes différentiellement exprimés. Ce fichier contient pour chaque gène : un *fold-change* (log du rapport des deux moyennes observées entre les deux conditions), une *p-value* et une *p-value* ajustée après correction des tests multiples.

Prestations complémentaires

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

Analyse Gene Ontology

L'annotation fonctionnelle des gènes différentiellement exprimés peut être réalisée grâce à une analyse statistique sur les termes Gene Ontology.

Analyse GSEA (Gene set enrichment analysis)

L'analyse GSEA (Subramanian *et al*., 2005) peut-être réalisée pour déterminer si un ou plusieurs ensembles de gènes défini a priori (ou signatures) sont significativement sur ou sous-représentés dans une des deux conditions biologiques étudiées. L'analyse peut être réalisée sur :

- une ou plusieurs bases de signatures choisies par l'utilisateur sur MSigDB (Molecular Signature Database),
- l'ensemble des 8 bases de signatures principales disponibles sur MSigDB,
- un fichier de signatures fourni par le client au format gmx ou gmt.

Une analyse de '*leading edge*' peut également être réalisée à partir des résultats de l'analyse GSEA afin d'identifier des gènes particulièrement intéressants parce qu'ils participent à l'enrichissement de plusieurs signatures entre les deux conditions biologiques.

Analyse de recherche de différence dans l'usage des exons

Le package R DEXSeq (Anders *et al.*, 2012) implémente une méthode permettant de tester s'il existe des différences d'usage des exons entre plusieurs conditions expérimentales. On appelle « différence d'usage d'exons » les modifications dans l'utilisation relative des exons qui sont causées par les conditions expérimentales, c'est-à-dire l'augmentation ou la diminution de l'expression d'un exon par rapport à l'ensemble des exons d'un gène donné. Cette modification d'usage des exons est généralement due à un changement dans le taux d'utilisation des exons lors de l'épissage alternatif.

La méthode de DEXSeq consiste, pour chaque exon de chaque gène, à compter le nombre de reads qui s'alignent sur cet exon, et le nombre de reads qui s'alignent sur tous les autres exons de ce gène. C'est le ratio entre ces deux comptages qui permet d'estimer l'usage des exons.

Matériel initial

Les données nécessaires à l'analyse bioinformatique sont directement issues d'un des séquenceurs du plateau technique.

En revanche, ce type d'analyse n'est envisageable que sur des espèces pour lesquelles il existe un génome de référence annoté. L'analyse Gene Ontology n'est réalisable que pour les organismes dont les gènes possèdent une annotation Gene Ontology. En ce qui concerne l'analyse GSEA, l'utilisation de MSigDB n'est possible que pour l'homme, la souris et le rat. L'analyse GSEA pour les autres espèces est possible uniquement si un fichier de signatures est fourni par le client.

Restitution des résultats

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

- 1 rapport d'analyse au format PDF disponible depuis le depuis le gestionnaire de projet (https://projet.mgx.cnrs.fr)

Pour chaque pool d'échantillon et chaque échantillon (*) :

- *.fastq : fichier texte contenant les séquences nucléotidiques ainsi que les scores de qualité correspondants

Pour chaque pool d'échantillon et chaque échantillon (*), en cas d'alignement et de comptage :

- *.bam : le format bam est la version binarisée (compressée) du format sam. Ce fichier contient les résultats de l'alignement, et permet notamment leur visualisation avec le logiciel IGV
- *.bam.bai : fichiers d'index allant de paire avec les fichiers *.bam. Cette paire de fichier vous permet de visualiser les alignements sur IGV
- *RawCounts*.txt : comptages bruts par gène, en reads bruts et en UMI

Pour chaque comparaison, en cas d'analyse différentielle :

- fichiers Excel contenant les gènes présentant une expression différentielle

- comptages bruts et comptages normalisés
- images des MA-plots et des nuages de points entre paires d'échantillons
- images des volcano-plots des gènes différentiellement exprimés entre conditions

Afin de faciliter le croisement des résultats de diverses comparaisons, un diagramme de Venn dynamique est produit grâce à l'outil Vennt ; celui-ci est consultable avec un navigateur web.

Pour chaque comparaison, en cas d'analyse Gene Ontology :

- fichiers html contenant les termes significativement enrichis
- fichiers csv contenant les termes significativement enrichis
- images des sous-graph GO des termes significativement enrichis

Pour chaque comparaison, en cas d'analyse GSEA, un dossier contenant l'ensemble des résultats qui pourront être parcourus grâce à un fichier html interactif :

- liste des signatures enrichies dans une des conditions biologiques
- images des graphiques d'enrichissement pour chacune des signatures
- heatmap de l'analyse '*leading gene*' dans le cas ou elle a été réalisée

Pour chaque comparaison, en cas d'analyse de recherche de différences dans l'usage des exons :

- fichiers Excel contenant les exons présentant des différences d'usage significatives
- dossier contenant l'ensemble des résultats et des figures qui pourront être parcourus grâce à un fichier html interactif

L'ensemble de ces fichiers est rendu disponible sur le serveur SFTP du plateau, à partir de la mise en ligne du rapport sur le gestionnaire de projet et pour une durée de 2 semaines. Les fichiers peu volumineux (< 100 Mo) sont mis à disposition sur le gestionnaire de projets. Le serveur SFTP 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 14 jours à compter de l'édition du rapport de résultat.

Nous ne nous engageons pas à conserver les données de façon pérenne. Il appartient à l'utilisateur de vérifier l'intégrité des données téléchargées depuis le serveur SFTP (à l'aide de clés MD5SUM fournies par le plateau technique) et de veiller à leur sauvegarde, sur un disque dur externe ou sur un espace serveur dédié.