


E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 1/7

La technique d'analyse RNA-seq permet de faire une étude qualitative et quantitative des différents transcrits d'un échantillon.

Prestation proposée

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

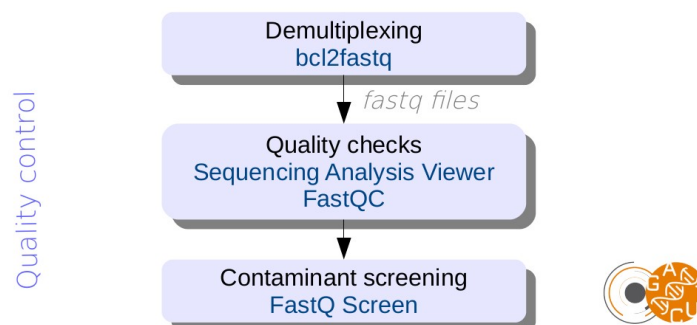
Production des fichiers fastq et démultiplexage des données

Le démultiplexage et la production des fichiers fastq est réalisée grâce au logiciel Illumina bcl2fastq.

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,
- recherche de contaminants.



Cette étape est réalisée systématiquement (bioinfo niveau 1). Les étapes suivantes sont réalisées sur demande (bioinfo niveau 2).

Trimming des reads


Le trimming des adaptateurs est réalisé avec le logiciel Cutadapt (Martin, 2011).

Alignement des reads

+ Annotation du génome disponible :

Si l'annotation du génome est disponible au format gtf (ou gff), les séquences obtenues sont alignées sur les jonctions d'épissage et le génome de référence (fichier au format fasta, la version du génome doit être indiquée par le client). L'alignement est alors réalisé avec le logiciel HISAT2 (Kim *et al.*, 2019) qui fonctionne en différentes étapes :

- Indexage du génome par la méthode *Hierarchical Graph FM index* (HGFM, basée sur les méthodes GCSA (extension de BWT adaptée sous forme de graphe) et FM index). Deux types d'index sont créés : un index global avec la totalité du génome et plusieurs milliers d'index locaux (qui couvrent mutuellement tous le génome). Chaque index local représente une région génomique d'environ 57 kb et chevauche ses voisins sur une fenêtre de 1 kb.
- Alignement des reads en utilisant les index créés à l'étape précédente.

E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 2/7

HISAT2 est un aligneur *splice-aware*, c'est à dire qu'il prend en considération le fait que lorsqu'un read est aligné, cet alignement peut être divisés sur plusieurs exons avec des gaps introniques plus ou moins grands selon les régions.

+ Annotation du génome non disponible :

Si le génome et son fichier d'annotation ne sont pas disponibles, le client peut fournir la liste des transcrits (au format fasta); les fichiers rendus sont alors des comptages bruts par transcrit.

Nous utilisons le logiciel d'alignement BWA (Li & Durbin, 2009), qui se présente sous deux formes algorithmiques différentes. Le choix de l'algorithme "backtrack" ou "MEM" se fera en fonction de la longueur de lecture des séquences, de la manière suivante :

- BWA-backtrack (`bwa aln`) si les *reads* ont une longueur inférieure à 70 pb.

BWA-backtrack autorise un maximum de 2 mismatches lors de l'alignement sur les 32 premières bases puis réalise l'alignement du *read* dans son intégralité.

- BWA-MEM si les *reads* ont une longueur supérieure ou égale à 70 pb.

BWA-MEM travaille en créant des *seeds* à partir des alignements ayant le Maximum de Matches Exactes (MEM). Une *seed* est une correspondance (ici exacte) d'une partie du *read* avec la référence. Cela signifie que, pour chaque position du *read*, BWA recherche le plus long alignement sans mismatches couvrant cette position. Ensuite, l'algorithme réalise une extension de la *seed* qui consiste à aligner le *read* dans son intégralité.


Cette méthode produit des alignements locaux (c'est-à-dire différents alignements à partir de différentes courtes séquences d'un même *read*) qui sont particulièrement adaptés aux longs *reads*.

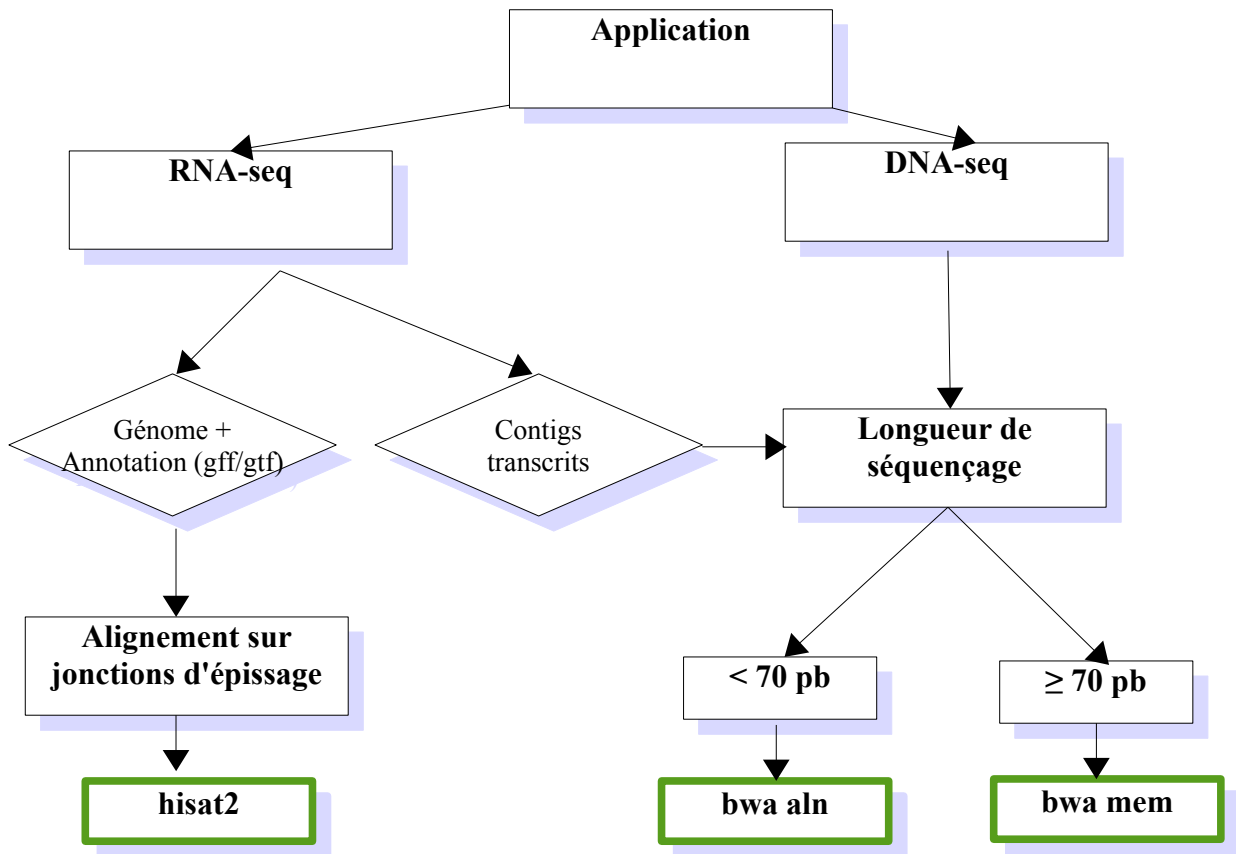
BWA reporte un score de qualité de l'alignement dans la colonne 5 (MAPQ) du fichier SAM/BAM. Le MAPQ dépend :

- de la différence entre le meilleur alignement et le second meilleur
- du nombre de seconds meilleurs alignements.

Le score MAPQ donne la probabilité que l'alignement reporté soit correct et par extrapolation, ce score reflète également l'unicité de l'alignement. Plus le MAPQ est grand, plus le *read* a de chances d'être aligné de manière unique. Il n'existe pas de consensus au niveau du score MAPQ, le choix du seuil est arbitraire. Nous avons choisi de filtrer les *reads* ayant un MAPQ < 20 afin d'éliminer ceux qui ont trop de chances d'être alignés à de multiples positions.

Les fichiers rendus sont des fichiers d'alignement de type "bam" triés par position. Le format bam permet de visualiser l'alignement RNA-Seq avec un logiciel dédié comme IGV.


E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 3/7



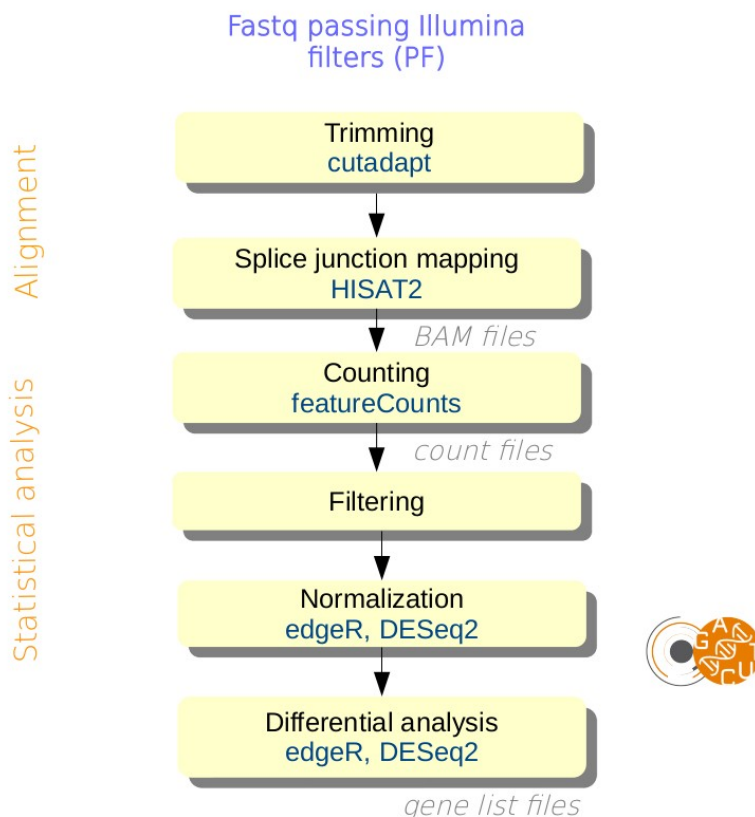
Comptage du nombre de *reads* par gène

Le logiciel featureCounts (Liao *et al.*, 2013) compte les *reads* alignés sur les gènes, un gène étant considéré comme l'union de ses exons. Ce logiciel permet de tenir compte de l'orientation des *reads* dans le cas de RNA-Seq orienté.

Les fichiers rendus sont des comptages bruts par gène (ou par exon).

E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 4/7


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.

E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 5/7

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 ou une liste de transcrits. 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.

E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 6/7

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

Pour 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

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 où 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 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.

E-EXP-11	Fiche prestation	Date : 23/03/2023 Version 10
	Analyse bioinformatique RNA-seq	Page 7/7

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.

Le dépôt des données brutes dans une base de données peut être demandé avant publication ; nous ne nous engageons pas à conserver ces données.