


| | | |
|---|--|-----------------------------------|
| E-EXP-23 | Fiche prestation | Date : 24/02/2020 Version 1 |
|  | <h2 style="text-align: center;">Analyse bioinformatique Single Cell RNA-seq</h2> | Page 1/3 |

La technique d'analyse Single Cell RNA-seq permet de faire une étude qualitative et quantitative des différents transcrits d'un échantillon, à l'échelle de la cellule. La technologie de 10X Genomics étant utilisée ici, seuls les transcrits les plus abondants pourront être quantifiés au sein de chaque cellule.

Prestation proposée

A partir des données de séquençage des échantillons, le plateau technique réalise les étapes de :

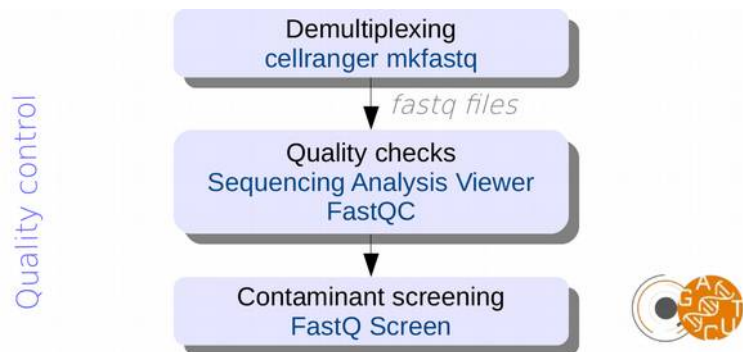
Production des fichiers fastq et démultiplexage des données

Cellranger mkfastq (10X Genomics) appelle bcl2fastq d'Illumina pour démultiplexer correctement les échantillons préparés avec le Chromium et convertir les *reads*, barcodes cellulaires et UMI (*Unique Molecular Identifier*) en fichiers fastq.


Contrôle qualité des données

Le contrôle qualité s'appuie sur plusieurs critères :

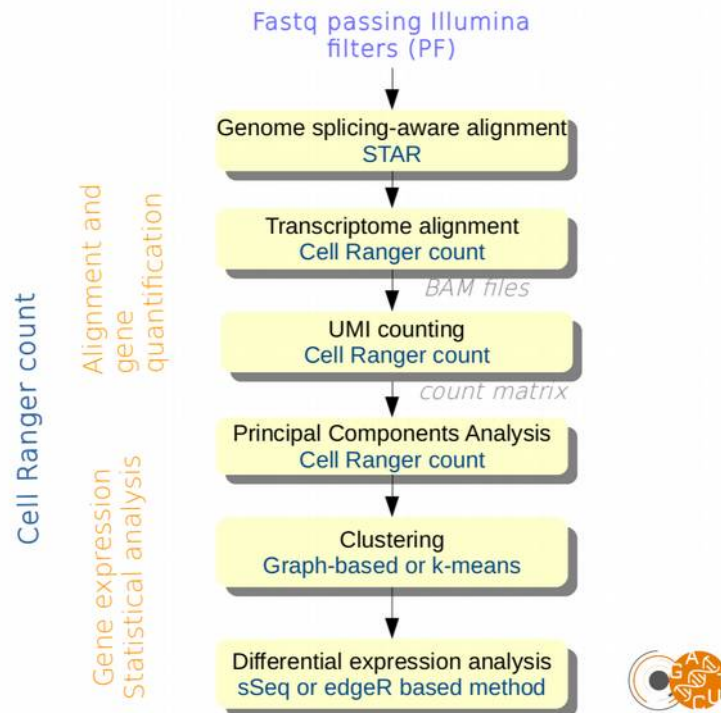
- validation du run en utilisant une série de critères associés
- distribution des scores de qualité à chaque cycle
- distribution des scores moyens de qualité par séquence
- pourcentage de bases "N" par cycles
- recherche de contaminants / adaptateurs



Ces étapes sont réalisées systématiquement (bioinfo niveau 1). Les étapes suivantes sont réalisées sur demande (bioinfo niveau 2).

| | | |
|---|--|-----------------------------------|
| E-EXP-23 | Fiche prestation | Date : 24/02/2020 Version 1 |
|  | <h1 style="text-align: center;">Analyse bioinformatique Single Cell RNA-seq</h1> | Page 2/3 |

Alignement des *reads* et comptages du nombre de *reads* par gène




Cellranger count est utilisé pour l'analyse échantillon par échantillon aboutissant à la production d'une matrice contenant pour chaque échantillon le comptage du nombre de *reads* par gène et par cellule. Les étapes réalisées par Cellranger count sont les suivantes :

- Alignement des *reads* correspondant aux cDNA sur le génome et le transcriptome en utilisant STAR, et production des fichiers BAM
- Génération des matrices de comptage pour chaque cellule en comptant le nombre de molécules uniques par gène
- Suppression des barcodes cellulaires provenant de gouttelettes (GEM) ne contenant pas de cellule (donc pas de cDNA)
- Analyse en composantes principales (PCA) pour réduire l'expression des gènes à ses composantes les plus variables
- Application de la méthode t-Stochastic Neighbor Embedding (t-SNE projection) pour visualiser les résultats de la PCA en deux dimensions
- Application de la méthode Uniform Manifold Approximation and Projection (UMAP) pour visualiser les résultats de la PCA en deux dimensions
- Clustering des cellules (pour classer les cellules en sous-populations) sur la base des résultats de la PCA, en utilisant la méthode k-means ou l'algorithme *graph-based*
- Analyse statistique d'expression différentielle : identification des gènes les plus différentiellement exprimés entre chaque cluster par rapport au reste de la population

Agrégation des résultats

Cellranger aggr est utilisé pour combiner différents échantillons dans une même analyse. Cette étape réalise une normalisation de façon à ce que tous les échantillons présentent la même profondeur.

| | | |
|---|--|-----------------------------------|
| E-EXP-23 | Fiche prestation | Date : 24/02/2020 Version 1 |
|  | Analyse bioinformatique Single Cell RNA-seq | Page 3/3 |

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.

Restitution des résultats

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

- Un rapport d'analyse au format PDF disponible depuis le logiciel de gestion de projet redmine

Pour chaque échantillon (*) :

- *_R1_*.fastq : fichier texte contenant les séquences nucléotidiques des barcodes cellulaires et des UMI, ainsi que les scores de qualité correspondants
- *_R2_*.fastq : fichier texte contenant les séquences nucléotidiques de l'ARN, ainsi que les scores de qualité correspondants
- *_I1_*.fastq : fichier texte contenant les séquences nucléotidiques de l'index, ainsi que les scores de qualité correspondants

Pour chaque échantillon (*), en cas d'analyse :

- *.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 un logiciel comme 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
- *_feature_bc_matrix : dossier contenant les comptages par gènes et barcodes cellulaires
- web_summary.html : fichier html contenant un résumé des métriques de l'analyse
- cloupe.cloupe : fichier d'entrée pour le logiciel Loupe Cell Browser (10X Genomics) permettant de visualiser et d'analyser les résultats obtenus de façon interactive
- analysis/pca : dossier contenant les résultats de l'ACP
- analysis/tsne : dossier contenant les résultats de la méthode tSNE
- analysis/umap : dossier contenant les résultats de la méthode UMAP
- analysis/clustering : dossier contenant les résultats des *clustering*
- analysis/diffexp : dossier contenant les résultats des analyses statistiques d'expression différentielle

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.

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