La ChIP-seq (*Chromatin Immuno Precipitation – Sequencing*) est une technique qui permet d'analyser les interactions ADN / protéine. A partir de données brutes de séquençage, le plateau technique propose plusieurs traitements bioinformatiques pour vous aider dans l'interprétation des données.

Prestation proposée

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

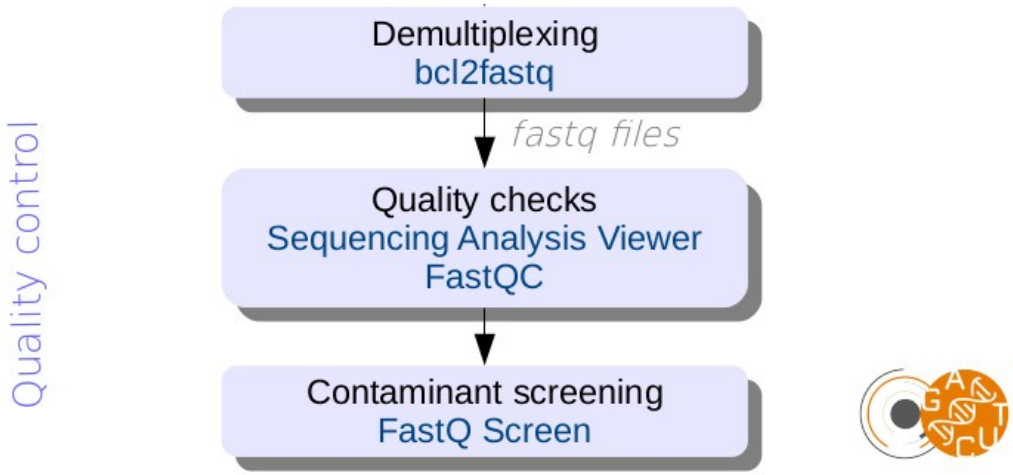
Production des fichiers fastq et démultiplexage des données

Le démultiplexage et la production des fichiers fastq 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 adaptateurs

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

Alignement des *reads* sur le génome de référence

Les séquences obtenues sont alignées sur le génome de référence (la version du génome doit être indiquée par le client) ou sur un ensemble de séquences fourni au format fasta.

Nous utilisons le logiciel d'alignement BWA, 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 a un algorithme équivalent à Eland (CASAVA) d'Illumina ; un maximum de 2 mismatches est autorisé lors de l'alignement sur les 32 premières bases suivi de 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 Exacts (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.

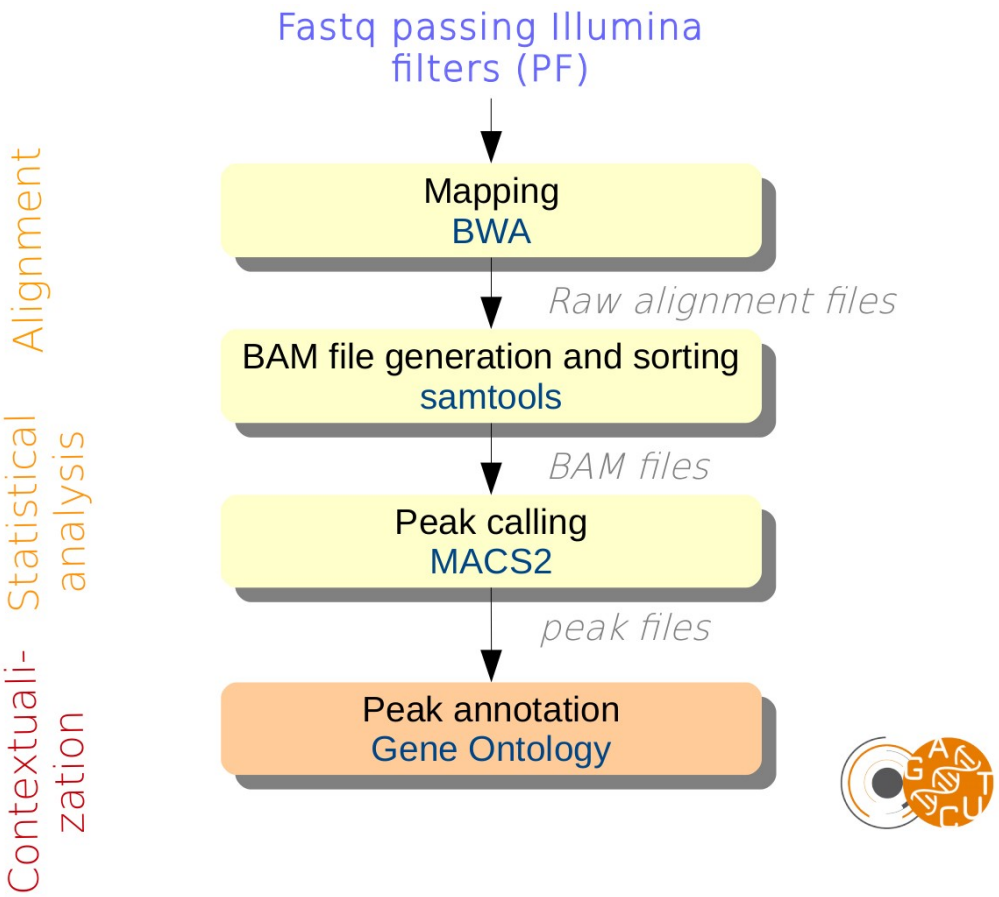
Le score MAPQ donne la probabilité que l'alignement reporté soit correct et par extrapolation, ce score reflète également l'unicité de l'alignement. Plus le MAPQ est grand, plus le *read* a de chances d'être aligné de manière unique.

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.

Détection de pics d'enrichissement dans les données de ChIP-seq

Une analyse statistique peut être réalisée pour rechercher les sites d'enrichissement. Le logiciel MACS (Model-based Analysis of ChIP-Seq) utilisé pour l'analyse permet d'identifier des régions enrichies dans les IP par rapport aux Inputs. Les résultats se présentent sous la forme de fichiers Excel pour les pics positifs et négatifs, et également sous format bed pour les pics positifs ainsi que pour les sommets des pics.

La figure ci-dessous illustre les étapes d'analyse réalisées sur demande (bioinfo niveau 2, cases en jaune) décrites ci-dessus, ainsi que les prestations complémentaires (bioinfo niveau 3, case en orange) détaillées dans le paragraphe suivant.



Prestations complémentaires

Des analyses complémentaires peuvent être effectuées sur demande :

Générer une liste de séquences à partir des pics détectés

Inférer les motifs de fixation ("binding motifs")

Des programmes de recherche de motifs (cisgenome 1.2 et MEME-Chip) pour générer une liste de motifs de fixation putatifs à partir de vos séquences peuvent être utilisés.

Annotation des pics

Une région génomique (correspondant à un pic d'enrichissement) est annotée en utilisant le gène le plus proche. La distance est calculée en amont du gène à partir du site "start" de transcription et en aval du gène à partir du site "end" de transcription.

Analyse Gene Ontology

A partir de l'annotation des pics, une annotation fonctionnelle peut être réalisée grâce à une analyse statistique sur les termes Gene Ontology.

Matériel initial

Les données nécessaires à l'analyse bioinformatique sont directement issues d'un des séquenceurs 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 (*) :

- *.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. Nous vous fournissons 2 fichiers bam par échantillon, un avec la totalité des *reads* (sortie de BWA) et un autre avec le filtrage MAPQ >= 20.

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

Des fichiers complémentaires peuvent être générés sur demande :

- *.bed : fichier contenant tous les tags alignés de manière unique sur le génome, au format bed, utile par exemple si vous voulez visualiser les données dans UCSC.

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 pour une durée limitée.

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.