


E-EXP-12	Fiche prestation	Date : 10/10/2022 Version 8
	<b>Analyse bioinformatique des variants génomiques</b>	Page 1/5

A partir de données brutes de séquençage, le plateau technique propose une analyse bioinformatique des variants génomiques (indels et SNP) de types germinaux et somatiques. Pour une analyse germinale, la couverture moyenne requise est au minimum de 20X, une couverture de 30X est recommandée.

Dans le cadre d'une analyse somatique appariée, la couverture moyenne requise est au minimum de 60X pour les données tumorales (si la pureté est <80%) et une couverture de 40X environ pour les données normales. Dans le cadre d'une analyse à un seul échantillon, la couverture recommandée est de 80X minimum.

Une pureté plus faible ou si l'on s'attend à une fréquence d'apparition de variants très faible nécessitent une augmentation de la couverture de séquençage.

## Prestation proposée

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

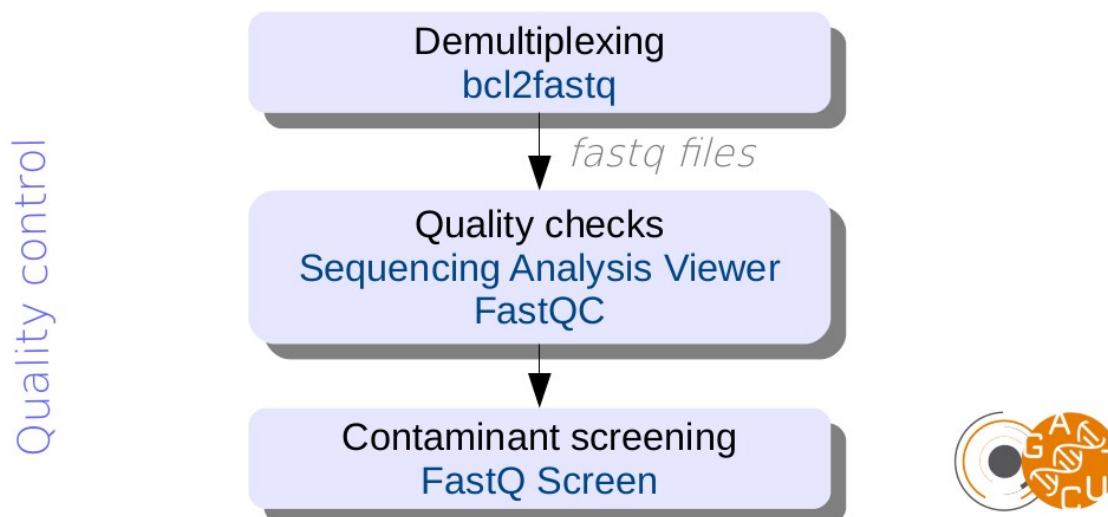
### Production des fichiers fastq et démultiplexage des données

Le démultiplexage et la production des fichiers fastq est réalisée grâce au logiciel Illumina bcl2fastq. Lorsque le kit utilisé pour la construction de la banque est connu, les adaptateurs présents dans les séquences sont recherchés et supprimés (*trimming*) pendant le démultiplexage avec le logiciel 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 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-12	Fiche prestation	Date : 10/10/2022 Version 8
	Analyse bioinformatique des variants génomiques	Page 2/5

### 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.

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 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.

### Analyse des variants


Le logiciel Octopus est utilisé pour la détection des SNP et des indels.

Une seule étape préliminaire est requise avant le processus de détection, l'ajout de groupes de *reads* avec Picard `AddOrReplaceReadGroup`. Aucune autre étape préliminaire n'est requise (le marquage de duplicats et le réalignement est réalisé en interne par l'outil de détection).

L'utilisateur peut préciser le type de variants à détecter : variants germinaux ou variants somatiques de faible fréquence.

Les variants somatiques sont détectés par Octopus selon deux modes d'analyse possibles :

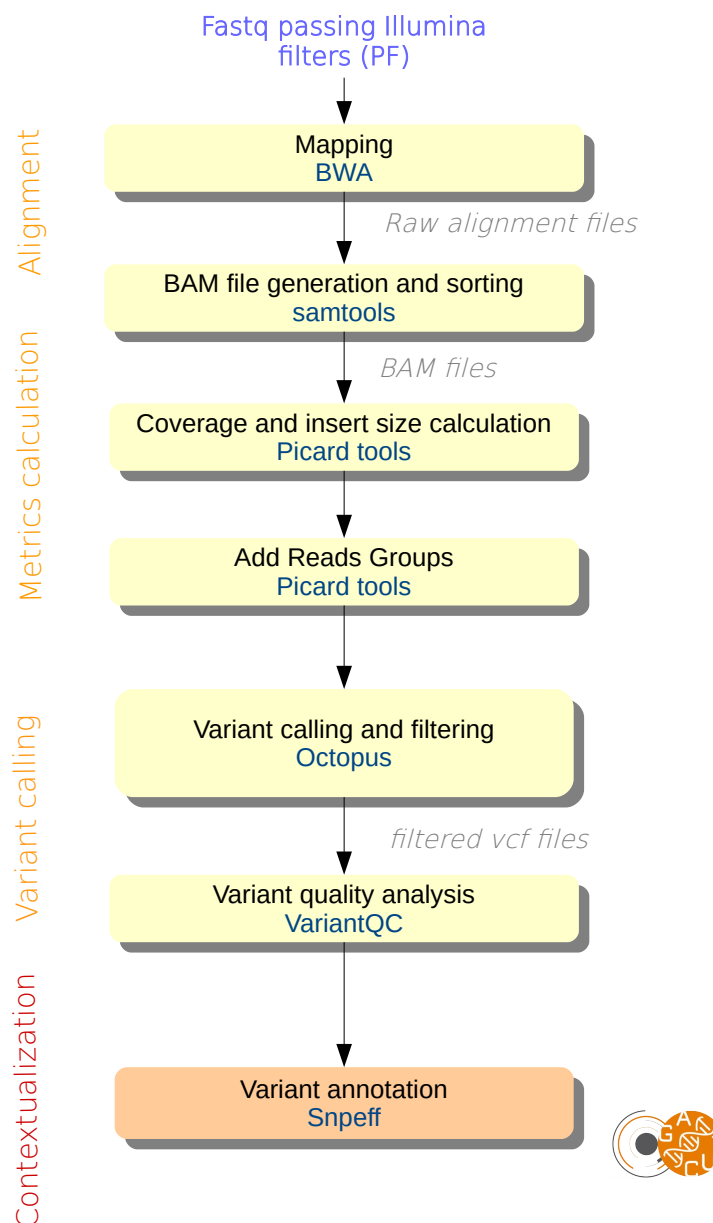
- Mode simple : analyse d'un seul échantillon (e.g. tumeur)
- Mode apparié : analyse d'un échantillon (e.g. tumeur) et des données appariées (e.g. échantillon sain)


E-EXP-12	Fiche prestation	Date : 10/10/2022 Version 8
	<h2>Analyse bioinformatique des variants génomiques</h2>	Page 3/5

Le mode individuel d'Octopus ne peut détecter que les variants germinaux (analyse « classique »).

Un filtrage de base est réalisé par Octopus (*hard filtering*), les différentes valeurs seuil utilisées pour le filtrage sont modifiables par l'utilisateur. Il est également possible de filtrer à l'aide d'un algorithme de type *random forest* pré-entraîné par les développeurs du logiciel, ou bien d'entraîner son propre modèle *random forest*.

Une analyse de la qualité des variants est également réalisée avec *VariantQC*.



E-EXP-12	Fiche prestation	Date : 10/10/2022 Version 8
	Analyse bioinformatique des variants génomiques	Page 4/5

## Prestations complémentaires

Des analyses complémentaires peuvent être effectuées selon la demande du client.

### Annotation des SNPs et indels

Les SNPs et indels sont annotés en fonction de leur position en terme de type de région génomique (*intergenic*, *exonic*, ...) et de gènes. L'annotation est réalisée grâce au logiciel SnpEff. Cela implique qu'une annotation des gènes soit disponible dans les différentes bases de données du logiciel.

Une information sur l'effet par type et par région ainsi qu'une évaluation de l'impact du variant est donnée par SnpEff.

### 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  $\geq 20$ .
- \*.fastq : fichier texte contenant les séquences nucléotidiques ainsi que les scores de qualité correspondants
- \*.vcf : fichier texte au format VCF. Ce format est le standard pour lister des variants. Chaque ligne correspond à un SNP ou un indel et contient notamment sa position, la base/séquence dans le génome de référence, la base/séquence alternative ainsi que la qualité du variant (format Phred).
- \*.stats : fichier texte contenant diverses statistiques sur les variants détectés
- \*\_VariantQC.html : fichier au format HTML contenant une série de tableaux ou de graphiques résumant différents aspects des variants détectés.

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

- \*\_snpEff\_summary.html : résumé au format HTML contenant différentes statistiques sur l'annotation ("Effects by type" vs "Effects by region", couverture...)

E-EXP-12	Fiche prestation	Date : 10/10/2022 Version 8
	<b>Analyse bioinformatique des variants génomiques</b>	Page 5/5

- \*\_genes.txt : SnpEff génère un fichier texte (séparé par des tabulations) contenant le nombre de variants affectant chaque transcrit et chaque gène
- \*.ann.vcf : fichier VCF annoté par SnpEff.

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.