**Analyse de données de séquençage haut-débit obtenues par la technologie Illumina HiSeq2000**

Echantillons : Ech1 et Ech2

A l’attention de : Machin BIDULE

**Résumé**

Les échantillons Ech1 et Ech2, séquencés par la société BiduleTruc ont été analysés à l’aide du pipeline développé par l’équipe bioinformatique de l’IJPB.

Table des matières

Mutant étudié 3

Génome de référence 3

Contrôles qualité et nettoyage 3

Estimation de la qualité des lectures brutes – Outil : FastQC version x.x 3

Nettoyage des lectures – Outil : Trimmomatic Version x.x 4

Estimation de la qualité des séquences nettoyées – Outil :FastQC 4

Etape d’alignement 4

Alignement sur la séquence de référence – Outil : BWA, version x.x 5

Filtrage de l’alignement – Outil : Script maison 5

Contrôle de la qualité  de l’alignement – Outils : samtools flagstat / Samstat 5

Calcul de la couverture – Outil : samtools mpileup 6

Les variants 6

Détection des variants – Outil : bcftools view 6

Filtrage des variants – Outil : Script R 6

Effets des mutations – Outil : SnpEff 6

Annexes 7

Annexe 1 : Schéma du pipeline Version x.x 7

Annexe 2 : Critères associes aux « failure » et « warning » de FastQC 7

Annexe 3 : Rapport HTML FastQC sur données brutes 7

Annexe 4 : Fichier log du traitement Trimmomatic 7

Annexe 5 : Rapport HTML FastQC sur données nettoyées 7

Annexe 6 : Log de l’alignement 7

Annexe 7 : Log filtrage de l’alignement 7

Annexe 8 : Log contrôle qualité de l’alignement 7

Annexe 9 : Log couverture 7

Annexe 10 : Log détection des variants 7

Annexe 11 : Log filtration des variants 7

Annexe 12 : Summary HTML de SnpEff 7

Annexe 13 : Fichier tabulé de SnpEFF 7

Analyse des données de séquençage des échantillons Ech1 et Ech2

Les échantillons ont été séquencés par la société XXX selon la technologie HiSeq 2000 d’Illumina (score Phred 22/64) en paired-end 2 x 100 pb.

La couverture attendue est de X.

Les séquences fournies sont nettoyées des adaptateurs et répondent au critère : 75 % des séquences ont un score qualité > Q30 (le score Q30 est associé à un taux d’erreur de 1/1000).

Les analyses sont réalisées à l’aide du pipeline V°x.x développé par l’équipe bioinformatique. Le schéma du pipeline est détaillé en annexe 1.

# Mutant étudié

Le séquençage en paired-end génère 2 fichiers au format FasqtQ, un pour chaque sens de lecture.

L’intégrité des données récupérées est vérifiée : checksum

Annexe 1 : Log du checksum

Les résultats obtenus pour le séquençage sont présentés dans le tableau suivant :

|  |  |  |  |
| --- | --- | --- | --- |
| Echantillons | Taille du fichier (Go) | Nombre de lectures | Nb de bases séquencées |
| Ech1.fastq |  |  |  |
| Ech2.fastq |  |  |  |

# Génome de référence

* Nom du génome de référence : *Arabidopsis thaliana*, écotype Columbia.
* Taille du génome : 119,67 Mbases
* Version du génome utilisée : TAIR 10
* Version/date de l’index :

# Contrôles qualité et nettoyage

Un premier contrôle qualité est effectué sur les lectures brutes, un deuxième sur les lectures après nettoyage.

## Estimation de la qualité des lectures brutes – Outil : FastQC version x.x

L’outil FastQC permet de passer en revue la qualité des données brutes selon différents critères et génère un rapport HTML des résultats. L’annexe 2 décrit les critères associés aux « failure » et « warning » des sorties de FastQC.

Les critères retenus pour le pipeline sont :

- Per base quality score : doit être supérieur à 20

Un score de qualité (phred quality score) est associé à chaque nucléotide des lectures fournies, qui reflète la probabilité que ce nucléotide soit une erreur de séquençage. Plus le score est grand, meilleure est la qualité de la base.

Le score de 20, correspondant à une chance sur 100 de commettre une erreur, est communément retenu.

- Sequence duplication levels : doit être inférieur à 50 %

Un trop grand nombre de duplicats nuit à la qualité de détection des SNPs en ccréant des couvertures artificielles, localement anormalement élevées.

Si le seuil dépasse 50 %, arrêt du pipeline.

Les duplicats seront éliminés après alignement si seuil entre 20 et 50 %.

Annexe 3 : rapport HTML sur données brutes

## Nettoyage des lectures – Outil : Trimmomatic Version x.x / fastx ???

Cette étape permet l’élimination des lectures de mauvaises qualités, qui peuvent altérer la qualité de l’alignement, des lectures trop courtes qui sont sources d’erreur, de même qu’une trop grande abondance de N. On grignote en 5’ et en 3’ les bases dont la qualité est inférieure à 20.

Critères retenus :

LEADING :20 élimine les bases en 5’ dont la qualité est inférieure à 20

TRAILING :20 élimine les bases en 3’ dont la qualité est inférieure à 20

SLIDING WINDOW :4 :20 une qualité moyenne de 15 est requise dans une fenêtre de 4 bases

MINLEN :50 élimine les lectures dont la taille est inférieure à 50 paires de bases après le nettoyage.

L’outil génère 5 fichiers :

* 2 pour les sorties pairées où les 2 lectures ont passé le nettoyage
* 2 pour les sorties non pairées
* 1 fichier log du traitement des séquences brutes (Annexe 4)

|  |  |  |
| --- | --- | --- |
|  | Ech1 | Ech2 |
| Nombre de lectures pairées |  |  |
| Nombre de non pairées |  |  |

## Estimation de la qualité des séquences nettoyées – Outil :FastQC

La qualité des séquences nettoyées est revérifiée.

Annexe 5 : Rapport HTML FastQC sur données nettoyées

# Etape d’alignement

A partir du jeu de données nettoyé, on procède à l’alignement de chaque lecture sur le génome de référence. En paired-end, les séquences doivent être alignées au même endroit pour être validées.

## Alignement sur la séquence de référence – Outil : BWA, version x.x

Le logiciel BWA autorise la présence de gaps. Il aligne d’abord séparément les paired-ends (aln) puis les combine ensemble en un alignement SAM (sampe).

On attend que ~1000 mutations EMS max par génome

Options choisies :

aln -n 2 -R 30 -o 1 (prioritaire sur n)

* -n : distance maximale d’édition = 2 => on n’accepte pas plus de 2 mismatchs dans une lecture
* -R : si moins de 30 best hits récupérer les alignements subotpimaux (ex si 3, va chercher 1 suboptimal …
* -o 1 On tolère jsq 2 évènements indépendants, sachant qu’on peut en avoir plus mais on filtre après

Niveau paire, une des 2 respectent NM de 2 max,dc un indel max de lgr 2 max car –o = 1

Apres sa copine, c’est n’importe quoi

Donc on accepte sur autre 2 ev indpdts de taille < 5 NM max 10

Paire = best hit ar filtre des subopt

Qualité = 0 ou sup 20

R>=Nb repetitions

si R = 30, on aura un defaut de couverture si la mutation est ds une region repete plus e 30 fois (donc potentiellement 30 paires correctes)

~~Prerun sampe par défaut sur 100 000 lectures pour trouver le « inferred maximum intersize » dans le log et le réinjecter en paramètre dans le vrai sampe …~~

Sampe –n 0–N 0

-n : nombre maximal d’alignements à reporter dans le tag XA pour les lectures pairées correctement . On ne veut qu’un best hit donc aucun reporter ds le XA

-N : aucun discordant read

pas d’option –a car il reporte tout ! Estime si suff de bons alignements .

samse –n 4

Annexe 6 : Log de l’alignement

## Filtrage de l’alignement – Outil : Script maison

1)

MAPQ = 0 ou est >= 20 ( mapping quality : proba erreur =1 %). On garde les lectures avec plusieurs occurrences (MAPQ = 0) et celles dont la mapping quality est supérieure à 20.

2)

XO + XM <= nb évènements indépendants (2 par défaut pour nous )

3)

CIGAR supprimer read où I ou D > taille microindel passé en paramètre

Filtrage sur les balises :

XM <=2 (nombre maximum de mismatchs autorisés)

X0 4 (nombre maximum de bests hits)

X1 0 (on n’accepte aucun hit suboptimal)

Annexe 7 : Log filtrage de l’alignement

## Contrôle de la qualité  de l’alignement – Outils : samtools flagstat / Samstat

|  |  |  |
| --- | --- | --- |
| Echantillon | Ech1 | Ech2 |
| Nombre de lectures |  |  |
| Nb de lectures mappées (%) |  |  |
| Nb de lectures pairées (%) |  |  |
| Nb de singletons (%) |  |  |
| Nb de lectures mappées sur un chromosome différent |  |  |

Annexe 8 : Log contrôle qualité de l’alignement

* Camembert samstat : % de lectures et qualité
* Nb lectures / nb MM
* Nb lectures/ nb de hits

## Calcul de la couverture – Outil : samtools mpileup

Options :

-B : disable proba realignemnt

-Q : base quality thresold 20

-u : for bcf uncompressed

-f ref indexed in fasta

[-6] si illumina 1.3 ( a determiner ds fastqc)

Annexe 9 : Log couverture

# Les variants

## Détection des variants – Outil : bcftools view

Options : -v (variant sites only)

-c :

-g

Annexe 10 : Log détection des variants

## Filtrage des variants – Outil : Script R

Filtrer sur le DP4, indicateur de couverture

DP4 (MF, MR, MMF, MMR)

MF nb de match Forward

MR nb de match Reverse

MMF nb de mismatch Forward

MMR nb de mismatch Reverse)

MF et MR doivent être < 2 (risque de rater snp causal)

Si MMF+ MMR > 10, vérifier le PV4 strand biais

|  |  |  |
| --- | --- | --- |
|  | Ech1 | Ech2 |
| Nb de variants avant filtrage |  |  |
| Nb de variants après filtrage |  |  |

Sortie : positionner les mutations / chromosome fait par snpeff

Annexe 11 : Log filtration des variants

## Effets des mutations – Outil : SnpEff

Annexe 12 : Summary HTML de SnpEff

Annexe 13 : Fichier tabulé de SnpEFF

# Annexes

## Annexe 1 : Log du checksum

## Annexe 2 : Schéma du pipeline Version x.x

## Annexe 2 : Critères associes aux « failure » et « warning » de FastQC

## Annexe 3 : Rapport HTML FastQC sur données brutes

## Annexe 4 : Fichier log du traitement Trimmomatic

## Annexe 5 : Rapport HTML FastQC sur données nettoyées

## Annexe 6 : Log de l’alignement

## Annexe 7 : Log filtrage de l’alignement

## Annexe 8 : Log contrôle qualité de l’alignement

## Annexe 9 : Log couverture

## Annexe 10 : Log détection des variants

## Annexe 11 : Log filtration des variants

## Annexe 12 : Summary HTML de SnpEff

## Annexe 13 : Fichier tabulé de SnpEFF