université de BORDEAUX

ANNEE UNIVERSITAIRE 2020 / 2021
SESSION 1

PARCOURS / ETAPE : Master Bioinformatique/Agrosciences
Code UE : 4TBI804U NGS
Epreuve : Cours
Date : 08/04/20 **Heure : 14h30-16h30** **Durée : 2h**
Documents : non autorisés
Epreuve de M/Mme : P. Sirand-Pugnet et P. Thébault

Collège Sciences et technologies

Exercice 1 :

Une banque plasmidique comportant une séquence dégénérée sur 7 positions a été construite et amplifiée dans *Escherichia coli* (Figure 1). Les bases A, C, G et T doivent donc être théoriquement présentes à la même fréquence sur ces 7 positions. Afin de vérifier la représentativité de la banque, un amplifiat de 50 pb incluant les positions dégénérées a été généré par PCR à partir de l'ADN extrait de la banque plasmidique. Cet amplifiat a ensuite été séquencé sur un séquenceur MiSeq en paired-end.

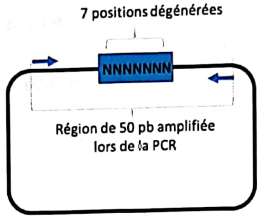


Figure 1: schéma du plasmide utilisé pour générer la banque.

1. **Expliquez, sous forme d'un schéma annoté, le principe du sequençage paired ends.** ***(2 points)***

Les fichiers fastq obtenus sont analysés avec le logiciel FASTQC. Certains graphes issus de cette analyse et correspondant au Read 1 sont présentés dans la Figure 2A. Des figures similaires sont obtenues après analyse du Read 2.

2. **Quel est le but d'une analyse avec le logiciel FASTQC ?** ***(1 point)***

Le format des fichiers fastq des Read 1 et Read 2 est vérifié avec le logiciel Groomer, puis les deux fichiers fastq sont traités avec le logiciel Trimmomatic. Ce logiciel permet de traiter en parallèle les deux fichiers et produit 4 fichiers fastq :

- Trimmed Read 1 (paired)
- Trimmed Read 2 (paired)
- Trimmed Read 1 (unpaired)
- Trimmed Read 2 (unpaired)

3. **A quoi correspondent ces 4 fichiers?** ***(2 points)***

Une nouvelle analyse FASTQC est alors réalisée sur ces 4 fichiers. Les graphes présentés en Figure 2B correspondent à l'analyse du fichier Trimmed Read 1 (paired).

4. **Comparer les graphes avant et après traitement par l'outil Trimmomatic. D'après ces graphes, déduisez les différents traitements qui ont été effectués par l'outil.** ***(2 points)***

5. **La banque de plasmides vous semble-t-elle représentative de la dégénérescence souhaitée sur 7 positions ? Justifiez votre réponse.** ***(3 points)***

Exercice 2:

L'analyse bioinformatique suivante (Popin et al., DOI: 10.1101/201178) propose de confronter les prédictions de variants obtenus avec plusieurs logiciels bioinformatiques (GATK HC, GATK UG, Platypus, SAMtools). Dans le cadre de cette analyse, deux types de données ont été utilisées, chacune provenant d'un séquençage paired-end avec des lectures de taille différentes, respectivement 250 pb pour WGS et 76 pb pour WES. La figure suivante représente les résultats de comparaisons, en analysant les faux négatifs (FN, nombre de variants réels qui n'ont pas été prédits), faux positifs (FD, nombre de faux positifs qui ont été prédits) et variants réels (GT concordance).

1. **Définissez les termes SNP et INDELS (2 points),**

En analysant les résultats pour les SNPS, vous répondrez aux questions suivantes en vous servant pour chaque question de la figure correspondante (par exemple, d'après SNP/FNR ou SNP/GT concordance ...)

2. **Proposez un schéma résumant les déifférentes étapes nécessaires à la prédiction de SNPS à partir de données FASTQ. (2 points)**
3. **Quels problèmes identifiez-vous avec le choix de SAMTools et le choix de Platypus et pour quel type de données? (2 points)**
4. **Mêmes questions pour la prédiction des indels. (2 points)**
5. **Que pensez vous de ces résultats et quel outils choisiriez-vous ? Pour répondre vous devez vous appuyer sur la sensibilité et spécificté des outils de prédiction. (2 points)**

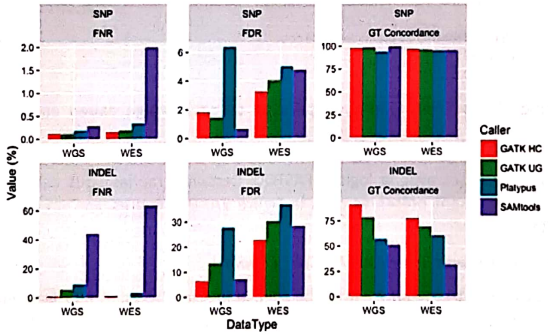
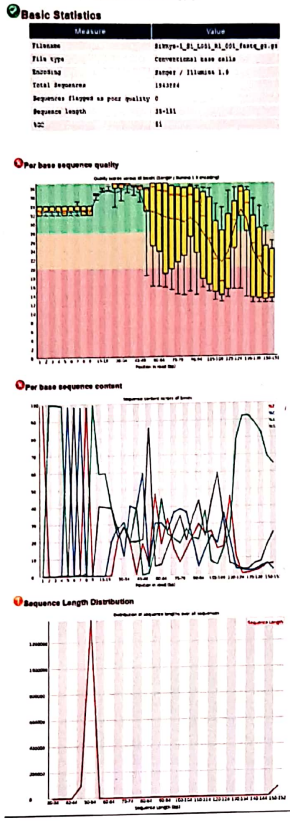


Figure 3: Variant caller comparison. Accuracy comparisons of both SNP and indel variant calls over all three samples in the CEU trio by the GATK HaplotypeCaller, GATK and Unified Genotyper, Platypus and SAMtools.
WGS data was 250bp paired-end reads sequenced by an Illumina HiSeq.
WES data was 76bp paired-end reads sequenced by an Illumina HiSeq.
Sensitivity is plotted as false negative rate (FNR), for which lower values gives to better sensitivity. Specificity is plotted as false discovery rate FDR, for which lower values are also better. For genotype concordance higher values indicate better genotype call accuracy (these results take into account the FNR and FDR scores.

Figure 2: Analyse FASTQC de fichier fastq

A - Avant traitement par Trimmomatic



B - Après traitement par Trimmomatic