TD1 : Traitement d'un Fichier Brut de Séquences Transcrites.

Dans le cadre de cette formation, nous allons utiliser des données Illumina (100-bases 'pair-ends') issues du transcriptome de plusieurs individus de l'espèce de riz Africain cultivé *Oryza glaberrima*

1. Prise en main de *Galaxy* :

Nous utiliserons ici l'outil *Galaxy* , qui permet de faciliter l'utilisation de plusieurs programmes couramment utilisés en bioinformatique, et ainsi de les rendre utilisable par un plus grand nombre de personne.

- Connectez-vous sur le serveur *Galaxy* du CIRAD, accessible à l'URL suivante : http://galaxy.southgreen.fr/galaxy/
- Utilisez les codes fournis pour accéder à votre compte.

Différentes possibilités s'offrent à vous pour rendre votre jeu de données accessible dans le serveur *Galaxy*. Ici, nous allons récupérer les données partagées (*Data libraries/Galaxy trainings 2015/NGS*).

Vous utiliserez dans un premier temps les données correspondant à l'individu RC1.

Le fichier noté *RC1_raw_1.fastq* correspond à la séquence *forward* de la polonie *Illumina* séquencé, celui noté *RC1_raw_2.fastq* à la séquence *reverse.* Ces fichiers sont dit pairés, c'est à dire que la première séquence du fichier *forward* correspond à la première séquence du fichier *reverse.*

2. Vérification de la qualité des séquences :

Les fichiers que vous avez récupéré sont des séquences issues d'un séquençage par la méthode Illumina. Avant de commencer à les utiliser, nous devons en contrôler la qualité.

Pour cela nous utiliserons le logiciel *FASTQC* (disponible gratuitement, http://www.bioinformatics.bbsrc.ac.uk/projects/download.html#fastqc, et implémenté sur *Galaxy*). Ce logiciel tourne sous toutes les plate-formes et permet de générer un rapport de la qualité moyenne sous forme *HTML.*

- Contrôlez la qualité des séquences à l'aide de *FASTQC*, disponible dans *NGS : Quality Control.*
- N'ajoutez pas de liste des contaminants, le logiciel prendra par défaut la liste classique.

- Quelles sont les critères importants à observer ?
- Quelle(s) analyse(s) majeure(s) pouvez-vous déduire de cette sortie par rapport à vos données ?

3. Trimming des extrémités 5' et suppression des adaptateurs/primers

Pour générer nos séquences, nous sommes passés par une étape d'amplification utilisant des hexaprimers aléatoires. Cependant, cette technique entraine un fort taux d'erreur sur les 7 premières bases des reads. Afin d'éviter la détection de faux SNPs du à ces erreurs, nous allons donc les supprimer

Dans la même étape, nous allons supprimer les adaptateurs/primers. Les adaptateurs/primers sont utilisés pour former la banque, créer les polonies et séquencer ces polonies. En fonction de la taille des séquences initiales (avant formation des polonies) et de la qualité de fabrication de la banque initiale, les adaptateurs peuvent être présents en plus ou moins grande quantité (dans la majorité des cas, si vous passez par un service extérieur pour effectuer votre séquençage, la compagnie peut vous proposer ces traitements).

Pour ces traitements, nous utiliserons le logiciel *CutAdapt*.

- Nettoyez les données avec la brique *CutAdapt*, en spécifiant le contaminant spécifique de votre fichier (Index 1 pour RC1, Index 2 pour RC2, ...) dans add new 5' or 3' adapters, une couverture commune de 7 bases, une qualité et une taille minimales de 20.
- Supprimer les 7 premières bases de l'extrémité 5'

Ces critères permettent d'éliminer les contaminants résiduels provenant du multiplexage (séquençage de plusieurs individus en même temps) sans risquer d'éliminer trop de vraies séquences. Le seuil de qualité de 20 permet aussi d'éliminer les bases de mauvaise qualités situées sur la fin des séquences, ce qui risquerait de générer des faux SNPs.

- Quels sont les risques à cette étape sur l'intégrité des données ?
- Pourquoi ne pas chercher toutes les séquences connues ?

4. Filtres des séquences sur leur qualité moyenne :

Afin de conserver uniquement des séquences de bonnes qualités, nous allons les filtrer sur leurs qualités moyenne.

- Utilisez la brique *NGS:Quality Control/Filter FastQ* pour retirer les séquences avec une qualité moyenne inférieure à 30 et une taille inférieure à 35.
- En quoi est-il gênant de conserver des séquences de mauvaise qualité ?
- Quelles ont été les modifications qui ont eu lieu dans la structure et la qualité des données après toutes ces étapes ?

5. Validation des Paires :

L'élimination des séquences de mauvaises qualité, ou de petites tailles après élimination des adaptateurs, a probablement invalidé la structure pairée des deux fichiers d'entrée. Or, cette structure en paire est essentielle pour la continuité des opérations, particulièrement dans le cas du Mapping.

- Utilisez l'outil *Synchronized paired FastQ*
- Quelle hypothèse fait l'outil ? Celà vous semble t'il correct ?

Pour la suite nous n'utiliserons que les reads encore pairés après cette étape

6. Mapping des données sur une référence de type CDS:

Ici, nous n'avons pas besoin de créer une référence, nous allons utiliser la séquence du riz Asiatique *Oryza sativa*, dont le génome entier est disponible et bien annoté. Nous allons utiliser la référence *reference.fasta* disponible dans *Data Libraries/Galaxy trainings 2015/NGS*

Pour le mapping nous allons utilisé l'outil BWA. Cet outil allie rapidité et précision, mais est très sensible aux différences entre les reads et la référence. Comme nous utilisons des reads courts (100 bases) et nettoyés de façons stringentes, cet outil est adapté à nos besoin. Pour des séquences plus longues, type 454, BWA a mis en place un outil (BWA MEM) utilisant un algorithme différent. Cet outil est plus lent, mais il est moins sensible aux erreurs de séquençage et aux polymorphismes.

- Dans la partie NGS, sélectionnez l'outil BWA – map short reads
- Choisissez d'utiliser une référence issue de votre historique, et sélectionnez *reference.fasta*
- Sélectionnez un mapping de séquences en pair
- Choisissez comme fichiers d'entrée les séquences *forward* et *reverse* nettoyées
- Lancer le mapping avec les autres conditions par défaut
- Quelles conditions de mapping aurions nous pu modifier, sachant que nous n'utilisons pas la référence exacte associée à nos échantillons ?

7. Tri des fichiers SAM

Les fichiers *SAM* issus du mapping sont triés dans l'ordre d'entrée des séquences, mais pas dans l'ordre de la référence (de la première base du premier gène à la dernière base du dernier gène). Il faut les trier par coordonnées (de la première base du premier gène à la dernière base du dernier gène). Les fichiers de sorties de mapping étant généralement très gros, cette étape est indispensable pour tous les post-traitements que vous souhaiterez faire. En effet, avec un fichier ordonné, les logiciels peuvent accéder très rapidement à l'information qui les intéressent.

Dans le même temps, nous allons compresser notre fichier SAM en le transformant en fichier BAM. Ce fichier n'est plus lisible à l'oeil, mais il est compressé et indexé.

- Dans la partie *Picard tools*, sélectionnez l'outil *SortSam*
- Comme méthode de tri, choisissez *COORDINATES*
- Quelle est l'intérêt de travailler sur des fichiers compressés ? Indexés ?

8. Élimination des duplicats techniques:

Une fois le fichier SAM unique créé, il faut éliminer les duplicats techniques, qui risqueraient de fausser la détection de SNPs en accroissant artificiellement la profondeur à leur localisation. Cette manipulation s'effectue sur un fichier *BAM*, version binaire des *SAM*. Ces fichiers sont compressés, donc moins volumineux, et indexés ce qui permet une accession rapide aux information qu'ils contiennent.

- Dans la partie *Picard tools*, sélectionnez l'outil *MarkDuplicates*

9. Ajout des ReadGroups

Les ReadsGroups vont permettre de relier les reads d'une fichier SAM/BAM à leur échantillon d'origine. Ceci est important afin de garder une traçabilité de l'origine de chaque reads dans un SAM/BAM multi-échantillon.

- Dans la partie *Picard tools*, sélectionnez l'outil *Add or Replace Groups*. Ne laisser aucun champs vide.

10. Réalignement local autour des Indels

L'algorithme utilisé pour le mapping est très performant. Malheureusement, il arrive qu'autour des indels il manque légérement de précision. Si ce n'est pas corrigé, ceci peut être à l'origine d'erreur lors de la détection de SNPs. Le réalignement local fait partie de la suite d'outil GATK. Dans un premier temps les zones à réalignés sont définies, puis le réalignement est effectué.

- Dans la partie *NGS:GATK2 Tools*, sélectionnez l'outil *Realigner Target Creator*
- Utiliser ensuite l'outil *Indel Realigner*

11. Création de Workflows:

Comme vous avez pu le constater, enchaîner toutes ces étapes est fastidieux. Hors toutes ces étapes doivent être lancé sur chaque échantillon de manière individuel. C'est pourquoi il est intéressant d'automatiser leur enchaînement. Pour cela nous allons créer un petit workflow.

- Dans les options de votre historique sélectionnez *Extract workflow*
- Précisez les différentes étapes que vous souhaitez y intégrer
- Renommer le et sauvegarder
- Allez le visualiser dans la partie workflow
- Modifiez certains paramètres pour qu'il soit spécifié lors du lancement du workflow
- Lancez votre workflow sur les données de l'individu RC2

12. Fusion de plusieurs BAM

Vous avez généré deux fichiers BAM, prêt pour la détections de SNPs. Nous allons maintenant les fusionner.

- Dans la partie *Picard tools*, sélectionnez l'outil *Merge SAM Files*
- Cliquez sur *yes* dans *Merge sequence dictionnaries*
- Fusionnez les BAM que vous avez généré pour RC1 et pour RC2
- Récupérez le fichier RC3-10.bam dans un historique publié et fusionnez le avec le BAM contenant les individus RC1 et RC2

13. Visualisation sous Tablet

- Récupérez le BAM contenant les 10 individus et ouvrez le avec le logiciel Tablet
- Chercher des SNPs
- Trouver un individu hétérozygote pour une position polymorphe
- Intuitivement, comment estimeriez vous la fiabilité d'un SNP? Comment feriez-vous la différence entre une position hétérozygote et une erreur de séquençage?

14. Détection des SNPs

Nous allons maintenant lancé la détection de SNPs à l'aide du module Unified Genotyper de GATK

- Dans la partie *NGS:GATK2 Tools*, sélectionnez l'outil *UnifiedGenotyper*
- Retrouvez vous les SNPs "détectés" avec tablet ?
- Récupérez les *BAM* avant le passage l'outil *IndelRealigner*
- Fusionnez les et lancer l'outil *UnifiedGenotyper* dessus
- Que remarquez vous ?

15. Filtrage des variants obtenus

Le module GATK est connu pour son taux relativement élevé de faux positifs. Pour cette raison, il est indispensable de filtrer les variants obtenus. Dans l'idéal, on peut utiliser la partie *Variant Recalibration* de GATK. Malheureusement, celle-ci nécessite de disposer de polymorphisme déjà connu afin d'entrainer son modèle. Dans notre cas, nous devrons nous contenter de filtres brutaux.

- Dans la partie *NGS:GATK2 Tools* sélectionnez l'outil *Variant Filtration*
- Ajouter les filtres sur la qualité de mapping, la qualité par profondeur et la profondeur global
- Ajouter un filtre sur les clusters de SNPs

16.