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

Dans le cadre de cette formation, nous allons utiliser des données Illumina (75-bases 'pair-ends') issues du transcriptome de plusieurs individus de l'espèce de riz Africain cultivé *Oryza glaberrima* et de son ancêtre sauvage *O. barthii*.

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://gohelle.cirad.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/Formation/Preprocessing and Mapping 2012*).

Chaque groupe va se voir affecter un couple de fichier d'entrée différent, représentant chacun un individu (1 à 10), soit sauvage (RS) soit cultivé (RC).

Le fichier noté *_1.fastq* correspond à la séquence *forward* de la polonie *Illumina* séquencé, celui noté *_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. Suppression des 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 présent 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 supprimer ces séquences nous utiliserons le logiciel *CutAdapt,* capable de couper les séquences adaptatrices (entières ou non, avec indel/mismatches) que l'utilisateur lui donne en entrée.

- Nettoyez les données avec la brique *CutAdapt*, en spécifiant le contaminant spécifique de votre fichier (affiché au tableau) dans add new 5' or 3' adapters, une couverture commune de 7 bases, une qualité et une taille minimales de 20.

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 Filter FastQ (ARCAD) (Tools, NGS:Quality Control) 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 Separate Fastq pair and single

6. Assemblage *de novo* avec Mira:

Maintenant que nos séquences ont été nettoyées, nous pouvons sans risque commencer à les utiliser. En général, si l'on ne dispose pas de séquence de référence, il convient de créer cette dernière *via* un assemblage *de novo*.

Pour créer cette séquence, nous utiliserons le logiciel *Mira. Mira* permet de faire des assemblages quelque soit la technologie utilisé pour générer les séquences. Il permet également de faire des assemblages mixtes entre séquences issues de différentes technologies.

Mira prend un fichier unique en entrée pour chaque technologie. Or, nous avons trois fichiers de séquences. Nous allons donc concaténer les fichiers d'entrée.

- Utilisez *Concatenate datasets* sur les fichiers nettoyés *forward et reverse.*
- Lancer *Untested Tools/NGS/Assembly/Assemble with Mira* avec les paramètres appropriés, *i.e. EST assembly* et *Solexa/Illumina reads = YES*.

- Combien de contigs obtenez-vous?
- Qu'en est-il des assemblages des groupes autour de vous ?
- Connaissez-vous la raison de ces différences ?

7. Validation des contigs via BLAST sur une base de référence:

Une fois vos contigs créés, il serait bon d'essayer d'identifier à quels gènes ils appartiennent. Attention, un même gène peut contenir plusieurs contigs.

- Utilisez l'outil *BLAST+ blastn (MC)* de la partie *Sequence Comparisons*.
- Effectuez une analyse de type *MegaBlast* sur la base *nt,* avec les conditions par défaut.

- Un même contig peut-il appartenir à deux gènes ?
- Si oui sous quelles conditions ?
- Si non dans quelles conditions ce type de résultat peut-il apparaître ?

8. 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/Formation-NGS/References*

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 for long reads) 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 for Illumina
- 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 ?
-

Il faut maintenant mapper aussi les séquences *single*.

- De la même manière qu'auparavant, utiliser l'outil *BWA*
- Choisissez d'utiliser une référence issue de votre historique, et sélectionnez *reference.fasta*
- Sélectionnez un mapping de séquences en *single*
- Choisissez comme fichiers d'entrée les séquences *single* nettoyées
- Lancer le mapping avec les autres conditions par défaut

9. 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) pour pouvoir regrouper les fichiers *SAM pair* et *SAM single* en un seul fichier représentant l'individu. 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 la partie *NGS : SAM/BAM manipulation*, sélectionnez l'outil *SortSam* de la suite *PicardTools*
- Comme input format, choisissez *SAM*
- Comme méthode de tri, choisissez *COORDINATES*

10. 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 *NGS : SAM/BAM manipulation*, sélectionnez l'outil *SAM-to-BAM* de la suite *SamTools*
- Choisissez un fichier *from history*, et indiquez le fichier *SAM pair* trié la bonne référence *reference.fasta.*
- Une fois la conversion effectuée, dans la partie *NGS : SAM/BAM manipulation*, sélectionnez l'outil *RemoveDuplicates* de la suite*PicardTools*
- Réiterez l'opération avec le fichier *SAM single*

A notez qu'à l'étape précédente (Tri des fichiers SAM), nous aurions pu en même temps faire la transformation au format BAM

11. Assemblage des données *single* et *pair:*

Une fois les deux fichiers triés dans le même ordre et les duplicats techniques éliminés, il faut les associer pour créer un fichier unique pour l'individu.

- Dans la partie *NGS : SAM/BAM manipulation*, sélectionnez l'outil *MergeSam* de la suite *PicardTools*

- Sélectionnez *BAM* comme format d'entrée et de sortie
- Indiquez le premier *BAM* d'entrée *(BAM pair)*
- Ajoutez un *dataset*
- Indiquez le deuxième *BAM (BAM single)*

12. Assemblage des sorties des différents individus

Avant de pouvoir analyser la variabilité au sein de nos échantillons, nous devons les regrouper au sein d'un fichier de type SAM/BAM unique. Pour pouvoir reconnaître l'origine des variations (*i.e.* quel individu est affecté par tel ou tel SNP), il nous faut adresser une information supplémentaire à chaque séquence, le *readgroup*.

- Il faut convertir le fichier *BAM* en *SAM,* via l'outil *SAM-to-BAM* de la suite *SamTools*
- Ensuite, dans la partie *NGS :* SAM/BAM manipulations sélectionnez l'outil *Add or Repalce Groups* de la suite *Picard Tools*
- Indiquez votre fichier *SAM* et renseignez votre R*ead group ID, Read group sample name* et *Read groups library* avec la même information (ex RC1)
- Renseigner les champs *Read group platform* et *Read group platform unit* avec le texte illumina

Une fois cette information ajoutée au fichier *SAM,* partagez le avec les autres membres de la formation.

- Dans les options de l'historique, choisissez *Share or Publish*
- Choisissez ensuite *Make History Accessible and Publish*

Quand chaque groupe aura partagé son historique, récupérez les données des fichiers *SAM* de tous les individus.

- Accédez ensuite aux historiques des autres groupes via *Shared Data/Published Histories*
- Sélectionnez un historique de la formation
- Téléchargez en local le fichier B*AM* correspondant
- Insérez le dans votre historique *via Get Data/ Upload Files from your computer*
- Recommencez l'opération pour récupérer les 18 fichiers B*AM* supplémentaires
- Joignez les 20 fichiers en un seul en utilisant l'outil *MergeSam* comme précédemment

13. Création de Workflows:

Comme vous avez pu le constater, enchaîner toutes ces étapes est fastidieux. C'est pourquoi il est intéressant d'automatiser leur enchaînement. Pour cela nous allons créer un petit workflow.

- Aller dans l'onglet *Workflow* et créer un nouveau workflow
-
- Ajouter le programme *CutAdapt* deux fois et modifier les options
-
- Ajouter *Filter Fastq (ARCAD)* deux fois
- Lier les sorties des *CutAdapt* aux *FilterFastq (ARCAD)*
-
- Ajouter le programme *Separate Fastq pair and single*
- Lier les sorties de *Filter Fastq (ARCAD)* à *Separate Fastq pair and single*

Votre workflow de nettoyage est maintenant fonctionnel.

- Lancer votre workflow avec en entrée les fichiers d'origine.
- Récupérer le résultat du workflow et lancer *FASTQC* dessus.

- Quel est l'avantage d'enregistrer/créer un tel workflow ?

- De la même manière, créez un workflow de mapping et un workflow d'élimination des duplicats techniques.

- Enfin, créez un workflow complet partant des données brutes à un fichier *SAM* avec *readgroups,* trié et nettoyé.