Tutoriel assemblage sur Galaxy

1°) uploader les fichiers fastq au format fastqsanger (attention il existe un fastqcsanger)

2°) utiliser « create assembly with Unicycler » en laissant les paramètres par défaut

Si besoin pour combiner plusieurs fichiers fastQ en un seul : concatenate datasets en faisant attention de concatener les bonnes pistes. On utilise cette commande quand le prestataire a fourni les données de deux pistes pour un même ADN.

Tutoriel détection de mutations/polymorphismes sur Galaxy

1°) uploader le génome de référence au format fasta

2°) uploader les fichiers fastq au format fastqsanger

3°) « Mapper » les fichiers fastq sur le génome de référence en utilisant Bowtie2 en laissant les paramètres par défaut

Note : il est possible de demander de sauvegarder les reads non mappés pour refaire un assemblage et avoir ainsi les séquences supplémentaires absentes de la souche de référence

4°) télécharger les fichiers BAM de mapping et les charger avec Artemis en utilisant le génome de référence comme support. Attention le fichier BAM est accompagné d'un fichier d'index. Pour être chargé correctement par Artemis, il faut que les deux portent le même nom soit respectivement nom.bam et nom.bam.bai

5°) regarder si le mapping s'est bien passé et déterminer la couverture approximative (e.g., 10X, 20X, 50X...). La couverture sera importante pour paramétrer le programme VARScan. Il est possible de la calculer en utilisant la formule suivante : nombre de reads mappés x taille des reads/taille du génome. Pour obtenir le nombre de reads mappés, il faut avoir cliqué yes en dessous de « Save the bowtie2 mapping statistics to the history »

6°) utiliser Samtools Mpileup en utilisant le fichier BAM et le génome de référence ~~avec l'option « Do not perform genotype likelihood computation (output pileup)»~~

7°) utiliser « VarScan mpileup » version 2.4.3.1avec le pileup obtenu.

* Choisir le type d'analyse. Il est bon de faire l'analyse de « single nucleotide variation » et des « indels »...

* le « Minimum coverage » et le « Minimum supporting reads » dépendent de la couverture. Mettre respectivement des valeurs légèrement inférieures à la moitié et au quart de la couverture. Pour un 40X, mettre donc Minimum read depth = 20 et Minimum supporting reads = 10.

* le « Minimum variant allele frequency» depend de la qualité. Pour des séquences typiques, mettre 90%. Mettre moins si elles sont de mauvaise qualité et plus pour

des séquences de qualité exceptionnelle. Cela se voit avec Artemis en utilisant l'option « Show SNP marks ».

Nouveau parameter: Pour le « Default p-value threshold for calling variants" mettre 95%

8°) ouvrir le fichier obtenu avec Excel ou autre lecteur de tables. Le plus simple est ensuite d'aller inspecter directement la séquence avec Artemis en n'oubliant pas de mettre l'option « Show SNP marks ».

Tutoriel d'analyse RNAseq

1°) s'assurer que les fichiers des séquences des ARNs sont sous format fastqsanger (éventuellement sous forme de fichier gzip : fastq.gz). S'ils sont sous format sra car provenant de GenBank, les convertir avec fastq-dump sous windows, linus ou mac en utilisant l'option gzip (la commande est alors fastq-dump.exe --gzip nom.sra ; attention l'exécution du programme est plus lent que sans l'option --gzip). Pour ceci, il faut télécharger le sra-toolkit pour votre OS (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/)

2°) uploader le génome de référence et les fichiers fastq dans galaxy

3°) mapper les fichiers fastq sur le génome de référence en utilisant TopHat en utilisant le full parameter list et changer dans cette liste le « minimum intron length » et mettre 45 au lieu de 70, le « maximum intron length » et mettre 1000 au lieu de 500 000 car chez les champignons les introns sont petits (chez *P. anserina* le plus petit fait 46 pb et seuls quelques introns font plus de 1 000 pb). Changer le « Minimum length of read segments » et mettre 15 au lieu de 25. Cela permet de mieux mapper les reads qui recouvrent les jonctions intron/exon. Ils définissent alors correctement les jonctions dans Artemis. Pour les visualiser ne pas oublier la commande « colour by RNAstrand specific tag »

4°) calculer la couverture de chaque exon en utilisant bedtools MultiCov Bed avec les paramètres par défaut. Pour ceci, il faut fournir les fichiers BAM de sortie de TopHat et un fichier GFF des « features ». Celui-ci peut être obtenu avec Artemis à partir de fichiers au format Artemis, GenBank, EMBL... Il faut sauver au format GFF, puis éliminer avec un éditeur de texte tout ce qui se trouve en dessous de ##FASTA (inclus).

5°) le fichier obtenu peut être analysé en utilisant Excel (cas de deux échantillons) ou DESeq2 dans Galaxy. Dans ce dernier cas, il faut fournir les fichiers de comptage des différentes conditions, incluant les réplicats, et le fichier GTF d'annotation. Il sera nécessaire de définir le dessin expérimental en indiquant les facteurs de l'expérience (« Specify a factor name »), a priori condition et réplicat, et les niveaux de chacun de deux facteurs (e.g., 1, 2 et 3 pour des triplicats ; control versus treated pour la condition). Conserver les autres paramètres à leurs valeurs par défaut et s'assurer que les sorties incluent bien la visualisation, les données normalisées et log-transformées (rLog normalized table). Pour avoir une comparaison de l'ensemble des conditions l'une contre l'autre, il faut cocher la case « Output all levels vs all levels of primary factor (use when you have >2

levels for primary factor) ». Le résultat de DESeq2 consiste en un fichier PDF avec des graphiques diagnostiques et un tableau avec les gènes et leur p-value corrigée pour le taux de fausses découvertes. Les gènes avec un faible nombre de reads sont automatiquement filtrés.