Rapport de Stage

CHARRIAT Florian

29 mars 2018

Matériels et Methodes

1.Matériels

- 201 génome (81 litterature et 120 projet)
- Définir les 32 RNAseq
- Définir protéine de référence de 70-15 (expliquer 70-15)

2.Methodes

2.1 Assemblage

Les données de séquençage étant des données short read, l'outils ABySS (Assembly By Short Sequences) [4] à été utilisé pour réaliser les assemblages. L'assemblage avec cet outil se réalise en deux étapes. Une première étape consiste à generer touts les k-mère de longeur donnée à partir des reads. Cette ensemble de k-mères est ensuite utilisée pour construire les contigs initiaux. Puis la deuxième étape utilise les informations de pair end pour étendre les contigs et résoudre les ambiguités de chevauchement des contigs.

Pour obtenir le meilleur assemblage possible, chaque souche est assemblée 8 fois avec une taille de kmère différents. Les tailles de kmère utilisées vont de 20 a 90 avec un pas de 10. Le meilleur assemblage de chaque souche est ensuite sélectionné après visualisation des données de qualité.

Une fois la selection réalisé, le script formatFastaName est utilsé. Pour chaque assemblage selectionné, le script élimine, dans une premier temps, les scaffolds de taille inférieur à 500 pb qui risque de poser problème pour l'annotation. Puis les scaffolds seront numerotés en fonction de leur longueur afin d'homogeneisé la nomenclature des scaffolds. Le plus grand scaffold sera alors nommé scaffold_1.

2.2 Eléments répétés

Une fois l'assemblage réalisé, il est necessaire de masquer les éléments répétés. En effet, les éléments répétés peuvent poser problèmes lors d'annotation automatique. Ils peuvent être confondus avec des gènes codant pour des protéines. Mais peuvent aussi perturber la structure des modèles de gène, en s'insérant dans les introns par exemple. L'outil repeatMasker est donc utilisé pour masquer les éléments répétés des assemblages selectionnés. Cet outil compare les sequences provenant d'une base de données d'éléments répétés avec ceux provenant de l'assemblage. Il va ensuite masquer ces éléments trouvés en les remplaçant par des N. L'outils repeatMasker permet aussi d'obtenir une annotation détaillée des répétitions présentes dans l'assemblage. La base de donnée utilisé pour repeatMasker est Repbase auquel à été ajouté les éléments répétées decouvert par l'equipe à l'aide de l'outil RepeatModeler.

Une fois les éléments masqués masqué, les scaffolds ne comportant que des éléments répétés sont élimner des assemblages. En effet, les outils utilisés pour l'annotation automatique n'arrive pas a annoter les assemblages comportant des scaffolds uniquement composé de N.

2.3 Annotation automatique

M.oyzae étant un eukaryote, l'outil Braker à été utilisée pour l'annotation automatique. Cet outil nescessite un fichier d'assemblage au format fasta et un fichier d'alignement de donnée RNA-seq sur l'assemblage au format bam. Braker utilise un pipeline d'annotation qui utilise les outils GeneMark-ET et AUGUSTUS.

Tout d'abord, GeneMark-ET génère des structures géniques. Pour cela l'outil utilse un ensemble de paramètre du hidden semi-Markov model (HSMM) définit initiallement pour prédire des régions codante dans l'assemblage. Une fois ces gènes prédites, un sous ensemble d'exon et intron est utilisé pour re-estimer les paramètres du HSMM. L'intégration des région codante dans le set d'entrainement nécessite que l'exon prédit possède au moins un site d'épissage. Les sites d'épissage sont ceux prédits indépendamment par les deux méthodes, l'ab initio et l'alignement des données RNA-Seq. Les étapes de prédiction des régions condantes et de la ré-estimation des paramètres du HSMM sont réalisées de manière itérative tant que les regions condantes prédites varies.

Deuxièmement, AUGUSTUS utilise les gènes prédits par GeneMarker-ET pour l'entraînement, puis intègre les informations de mapping des données RNA-seq dans les prédictions de gènes finaux.

Pour enrichir les données utilisées par Braker le fichier d'alignement bam est remplacé par un fichier hints intronique. Ce fichier correspond au position de tous les introns récupérés à partir de fichiers d'alignements. Le fichier hints utilisés est issue de l'alignement des 32 RNA-seq et du fichier de protéine de références issue de l'annotation de 70-15.

Pour l'annotation automatique des assemblages, le pipeline BRAKER_pipeline à été réalisé en snakemake. La figure X présente un exemple du pipeline d'annotation pour la souche XXX. Les alignements des données RNA-seq et l'alignement des protéines de référence vont être traité en parrallèle par le pipeline. Pour l'alignement des 32 données de RNA-seq Braker_pipeline va être utilisé l'outil topHat2. Puis les fichiers bam obtenue seront mergé à l'aide de l'outils samtools merge. En suite le fichier contenant tous les alignements va être trié à l'aide de l'outils picard SortSam. Enfin l'alignement au format bam va être converti en fichier hints à l'aide de l'outil bam2hints fournis avec l'outil Braker. Pour l'alignement des protéines de référence, l'outil exonerate va être utilisé. Une fois l'alignement réalisé l'outil exonerate2hints est utilisé pour convertir le fichier d'alignements en fichier hints.