Conception d'un projet de recherche et de développement

Cahier des charges du projet : Analyse de données RNA-seq (champignons *Fusarium*)

Clients :
Nadia Ponts
Fabien Dumetz

Laboratoire :
Laboratoire Bordelais de Recherche en Informatique

Auteurs :
Djemilatou Ouandaogo
Linda Khodja
Lucien Piat
Maroa Alani

Superviseur :
Marie Beurton-Aimar

Table des matières

Chapitre 1

Introduction

L'INRAE, ou Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, est un organisme public de recherche français. En son sein, l'unité de recherche Mycologie et Sécurité des Aliments cherche à comprendre les mécanismes de contamination des aliments par les mycotoxines. [1]

Les champignons du genre Fusarium infectent le blé et produisent des toxines sur les épis destinés à la consommation. Ces derniers élevés en coculture forment un Meta organisme, le Meta-Fusarium. Ses composantes peuvent communiquer grâce à des petits ARN appelés smRNA.

La problématique du projet est d'analyser des données de smRNA-seq en sortie de séquenceurs short read Illumina, les aligner sur les génomes de référence, identifier les petits ARN produits dans chaque scénario de coculture et les quantifier.

Un pipeline de traitement sera produit par les étudiants du Master de Bio-informatique de Bordeaux pour répondre à la problématique et un rapport lui sera associé.

Mots-clés : Fusarium, mycotoxines, coculture, Meta-Fusarium, communication, smRNA, RNAseq

Chapitre 2

Analyse

1 Contexte

1.1 Contex biologique du genre *Fusarium*

Depuis un certain temps, les biologistes portent un vif intérêt aux champignons filamenteux du genre *Fusarium*. En effet, les *Fusarium* sont des phytopathogènes qui contaminent, entre autres, les céréales que consomme l'homme comme le blé. Chez ce dernier, ils entraînent la fusariose de l'épi qui détruit les cultures et entraîne des pertes économiques conséquentes. Ces mycètes, sont aussi à l'origine de la contamination des grains par des mycotoxines constituant un problème majeur de sécurité alimentaire. Ces toxines comme les B-trichothécènes sont très stables et se retrouvent dans les grains qui finiront dans l'alimentation.

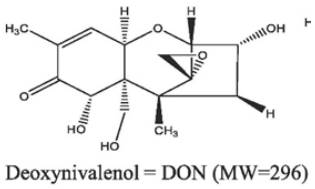


Figure 2.1 – Formule semi développée du Deoxynivalenol, molécule de la famille des B-trichothécènes produite par les *Fusarium*[2]

Jusqu'à lors, le processus de production des mycotoxines a été étudié en ne considérant "qu'un pathogène - une maladie". Cependant, des preuves irréfutables des interactions entre les espèces de *Fusarium* responsables de la fusariose, laisse suggérer que la communication entre ces champignons puisse moduler la régulation de production des toxines.

Afin de mettre en exergue les mécanismes de production de ces molécules inter-individus, il est nécessaire changer d'échelle d'analyse d'observer plus globalement le " Meta-Fusarium sp."qui comprend les principales espèces impliquées dans l'infection.[3, 4]

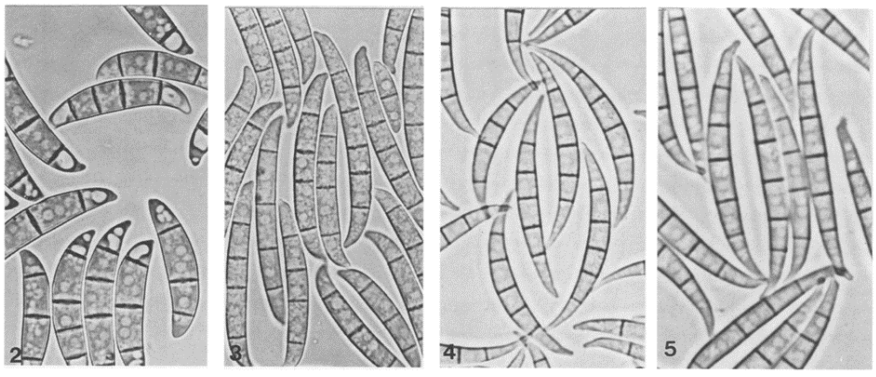


Figure 2.2 – Observation des conidies de 2 : F.culmorum ; 3 : F.solani ; 4 : F.equiseti ; 5 : F.graminearum qui sont des champignons qui font partie du Meta-Fusarium sp. (950X) [5]

1.2 La Communication au sein du Meta-Fusarium sp.

Au sein du Meta-Fusarium sp. la communication s'effectue grâce à des petits acides ribonucléiques comme les small ARN (smRNA) ou les micro ARN (miRNA). Ces derniers sont des courtes suites de bases qui peuvent être séquencées.

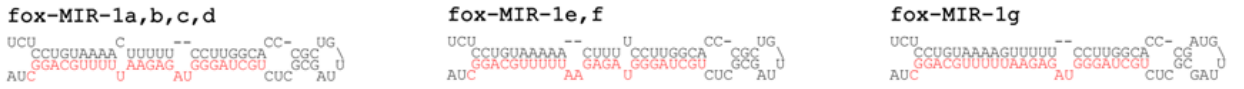


FIGURE 2.3 – Structure secondaire de quelques précurseurs miARN déjà mis en évidence chez les champignons du genre *Fusarium*.[6]

Ces petites molécules d'ARN non codantes d'environ 22 nucléotides, sont impliqués dans la régulation post-transcriptionnelle de l'expression génique. En identifiant les miARN spécifiques produits dans chaque scénario de co-culture et en comprenant leurs origines. Nous pourons mieux comprendre les mécanismes sous-jacents aux interactions hôte-pathogène et de développer des stratégies de prévention contre la contamination des cultures par les champignons *Fusarium*. [3, 4]

L'objectif du projet est d'analyser les données séquencées et d'identifier les microARN produits dans chaque scénario de culture, d'en repérer les spécificités et de les quantifier.

2 Etat de l'art/Existant

2.1 Contrôle qualité des reads

Contrôle qualité des fichiers fastQ pour supprimer ou modifier les artéfacts et les reads de mauvaise qualité.Utilisation de logiciels tels que FastQC. FastQC est un outil utilisé pour évaluer rapidement la qualité des données de séquençage. Il fournit des statistiques détaillées et des graphiques pour identifier les problèmes potentiels tels que la qualité des bases, la distribution des tailles d'insertion, la présence d'adaptateurs, etc. FastQC est souvent utilisé comme première étape dans l'analyse des données de séquençage pour évaluer rapidement la qualité globale des échantillons. Il est gratuit est disponible sur Github.[7] Il existe un autre outil utilisé MultiQC [8] qui est un outil qui permet d'agréger les résultats de plusieurs analyses FastQC (ou d'autres outils similaires) en un seul rapport.MultiQC est particulièrement pratique lorsque vous effectuez une analyse à grande échelle impliquant de nombreux échantillons, car il permet de consolider les résultats de manière efficace. Le choix entre FastQC et MultiQC dépend de notre besoin d'une évaluation détaillée de la qualité des données pour chaque échantillon individuel, ce qui signifie que FastQC est souvent le meilleur choix.

2.2 Nettoyage des reads

Trimmomatic [9] est un outil largement utilisé pour nettoyer les reads en éliminant les bases de faible qualité, en coupant les adaptateurs et en éliminant les séquences de mauvaise qualité à partir des extrémités des reads. Trimmomatic produit des fichiers de sortie contenant les reads nettoyés, prêts à être utilisés pour l'alignement ou d'autres analyses. Il existe un autre outil utilisé Cutadapt [10] Il est également capable de détecter et de supprimer les régions de faible qualité, mais sa principale force réside dans la gestion précise des adaptateurs.Nous préférons Trimmomatic pour sa facilité d'utilisation et sa capacité à nettoyer rapidement les données de séquence.

2.3 Alignement des reads sur le genome

Bowtie2 [11] est un outil de cartographie de reads à haut débit qui aligne efficacement les reads sur un génome de référence en utilisant l'algorithme de l'alignement parfaite. Il génère des fichiers d'alignement indiquant où chaque read s'aligne sur le génome de référence.BWA [12] est un autre outil largement utilisé pour l'alignement de séquences d'ADN à courte lecture.Bowtie2 est généralement plus rapide que BWA, ce qui peut être bénéfique pour les analyses de séquençage d'ARN à grande échelle avec un grand nombre d'échantillons.

2.4 Identification des petits ARN produits

À partir des données de smRNA-seq, l'objectif est d'identifier les petits ARN (micro ARN - miARN) présents dans chaque échantillon issu de différentes situations de co-culture.Utilisation d'outils spécialisés comme miRDeep2 [13] ou miRBase [14] pour détecter et annoter les petits ARN dans les données alignées.MiRDeep2 est un outil spécifiquement conçu pour la prédiction de nouveaux microARN (miARN) à partir de données de séquençage.Il utilise à la fois les données de séquençage des petites ARN et les informations sur la structure secondaire des précurseurs de miARN pour prédire de nouveaux miARN et évaluer leur fiabilité.Mais MiRBase est une base de données de référence qui répertorie les séquences de miARN connues.MiRDeep2 est souvent utilisé lorsque l'on souhaite identifier de nouveaux miARN dans des données de séquençage, nous allons donc l'utiliser.

2.5 Quantification des petits ARN

Estimation des niveaux d'expression des petits ARN détectés à l'aide d'outils comme HTSeq[15] ou featureCounts[16].HTSeq est un package Python conçu pour analyser les données de séquençage à haut débit, y compris les petites données de séquençage d'ARN.featureCounts fait partie du package Subread, qui fournit des outils pour cartographier et quantifier les lectures de séquences à partir d'expériences de séquençage à haut débit.Semblable à HTSeq, featureCounts prend en entrée des lectures de séquence alignées (au format BAM) et un fichier d'annotation de référence. Il attribue chaque lecture alignée à des caractéristiques génomiques spécifiées dans le fichier d'annotation, telles que des gènes ou de petits loci d'ARN.La sortie de featureCounts est également une matrice de comptage, où chaque ligne représente une caractéristique génomique et chaque colonne représente un échantillon, avec des valeurs indiquant le nombre de lectures mappées à chaque caractéristique dans chaque échantillon.Et featureCounts offre des options de comptage de lectures spécifiques à un brin, ce qui peut être important pour quantifier avec précision l'expression des petits ARN, en particulier en présence d'une transcription antisens.C'est donc l'option la plus appropriée.

2.6 Analyse différentielle

Identification des petits ARN différentiellement exprimés entre les différentes conditions de co-culture à l'aide d'outils comme DESeq2 [17] ou edgeR[18].DESeq2 et edgeR sont tous deux des outils bioinformatiques largement utilisés pour effectuer une analyse d'expression différentielle des données de séquençage d'ARN.Ils sont particulièrement utiles pour identifier des gènes ou d'autres caractéristiques génomiques qui présentent des différences d'expression statistiquement significatives entre différentes conditions expérimentales.

2.7 Visualition des resultats

Matplotlib est une bibliothèque de visualisation de données pour Python. Elle offre une grande variété de graphiques et de styles pour représenter les données, ce qui en fait un choix populaire pour la visualisation de données génomiques. Matplotlib permet de créer une grande variété de graphiques, y compris des histogrammes, des graphiques linéaires, des diagrammes en boîte, etc., pour représenter vos données de manière visuelle. Les visualisations produites avec Matplotlib peuvent être utilisées pour comprendre et présenter les résultats de l'analyse de séquençage, faciliter la communication des résultats ou guider les analyses ultérieures.

Tous les outils mentionnés peuvent être utilisés depuis Python via des interfaces ou des bibliothèques Python. Pour les outils FastQC, Trimmomatic et Bowtie2, l'importation de la bibliothèque subprocess permet leur exécution. Matplotlib est une bibliothèque Python directe. Il peut être installé et importé dans les scripts Python pour créer des visualisations de données génomiques ou de séquençage.

3 Analyse des besoins

3.1 Besoins fonctionnels

Pour le prétraitement des données, les besoins comprennent :

— Acquisition des données issues du séquençage Illumina (Short Read) générées à partir de différentes conditions de culture de Fusarium au format FASTQ.
— Prétraitement et nettoyage des données brutes afin d'éliminer les séquences de faible qualité et les erreurs techniques en réalisant un contrôle qualité des lectures.
— Normalisation des données si nécessaire.

Pour l'analyse des données, les besoins incluent :

- Alignement des lectures prétraitées sur les génomes de référence des souches de Fusarium.
- Identifier les miARNs, déterminer leur origine ou leur localisation génomique, et évaluer leur quantité.
- Analyse comparative des profils d'expression des microARN entre les échantillons de cultures pures et ceux des confrontations entre souches.
- Visualisation des résultats à travers des tableaux et des graphiques codifiés présentant les profils d'expression des miARNs dans les différentes conditions de culture.

3.2 Besoins non fonctionnels

- Utilisation recommandée des langages de programmation adaptés à la bioinformatique, tels que Python, R et éventuellement Bash.
- Traitement efficace des données dans des délais raisonnables.
- Fiabilité et robustesse pour minimiser les risques de perte de données ou d'erreurs dans l'analyse.
- Intuitivité pour une utilisation aisée par les chercheurs non spécialisés en bioinformatique.
- Documentation claire et interfaces rationnelles pour faciliter la navigation et l'utilisation des fonctionnalités.

3.3 Contraintes

- Utiliser des outils open source afin d'assurer la transparence, la réutilisabilité du système.
- Effectuer nos tâches dans un cadre à accès pseudo-restreint.
- (a confirmer) Proposer des solutions de visualisation compatibles et integrables avec les bases de données fonctionnelles existantes pour les champignons filamenteux.

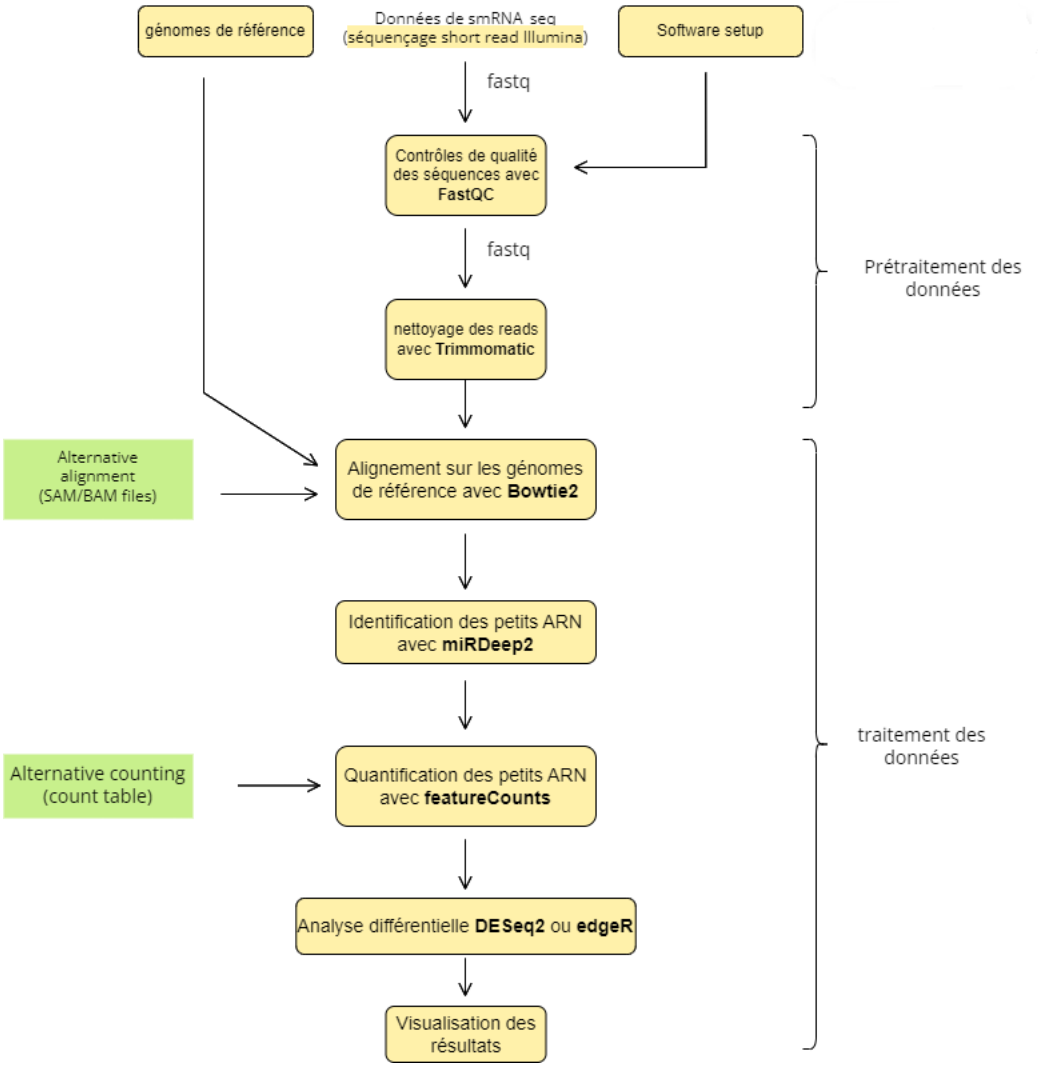
3.4 Ajouts optionnels

- Création d'une base de données pour stocker les résultats.
- Sauvegarde et reprise du processus d'analyse et reprise de l'analyse là où elle s'est arrêtée.
- Identification des cibles potentielles des miARNs dans les génomes des souches.

Chapitre 3

Flux opérationnel

Voici un diagramme reprenant les différents points à réaliser.



Chapitre 4

Organisation

1 Diagramme de Gantt

Bibliographie

[1] About us. 2020.

[2] Mohamed A Gab-Allah, Yeshitila Getachew Lijalem, Hyeonji Yu, Seulgi Lee, Se-Yeong Baek, Jaejoon Han, et al. Development of a certified reference material for the accurate determination of type b trichothecenes in corn. *Food Chemistry*, 404, 2023.

[3] Nadia Ponts, Leslie Couedelo, Laetitia Pinson-Gadais, Marie-Noelle Verdal-Bonnin, Christian Barreau, and Florence Richard-Forget. Fusarium response to oxidative stress by h2o2 is trichothecene chemotype-dependent. *FEMS Microbiology Letters*, 293 :255–262, 2009.

[4] Communication MycSA. Le projet anr 2022-2026 teamtox. 2023.

[5] Peter E Nelson, Maria C Dignani, and Elias J Anaissie. Taxonomy, biology, and clinical aspects of fusarium species. *Clinical microbiology reviews*, 1994.

[6] Rui Chen, Nanyu Jiang, Qing Jiang, Xia Sun, Yulin Wang, Hong Zhang, et al. Exploring microrna-like small rnas in the filamentous fungus fusarium oxysporum. *PLoS ONE*, 9, 2014.

[7] Simon Andrews. Fastqc : A quality control tool for high throughput sequence data. https ://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010.

[8] Ewels, Philip and Magnusson, Måns and Lundin, Sverker and Käller, Max. Multiqc : summarize analysis results for multiple tools and samples in a single report. https ://multiqc.info/, 2016.

[9] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic : a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15) :2114–2120, 2014.

[10] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1) :10–12, 2011.

[11] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4) :357–359, 2012.

[12] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14) :1754–1760, 2009.

[13] Martin R Friedlander, Sebastian D Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. Discovering micrornas from deep sequencing data using mirdeep. *Nature biotechnology*, 26(4) :407–415, 2012.

[14] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. mirbase : from microrna sequences to function. *Nucleic acids research*, 47(D1) :D155–D162, 2019.

[15] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2) :166–169, 2015.

[16] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts : an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7) :923–930, 2014.

[17] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12) :550, 2014.

[18] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1) :139–140, 2010.