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 :
L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

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 co-culture forment un Meta-organisme, le Meta-Fusarium. Ses composantes peuvent communiquer grâce à des petits ARN appelés miRNA.

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 co-culture et les quantifier.

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

Mots-clés : Fusarium, mycotoxines, co-culture, Meta-Fusarium, communication, smRNA, RNAseq

Chapitre 2

Analyse

1 Contexte

1.1 Contexte 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.

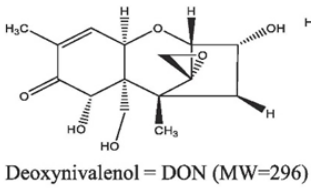


Figure 2.1 – Formule semi développée du Deoxynivalenol, molécule de la famille des B-trichothécènes produite par les *Fusarium*[2]

Jusqu'à présent, 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 afin d'observer plus globalement le "Meta-Fusarium sp." qui comprend les principales espèces impliquées dans l'infection.[3, 4]

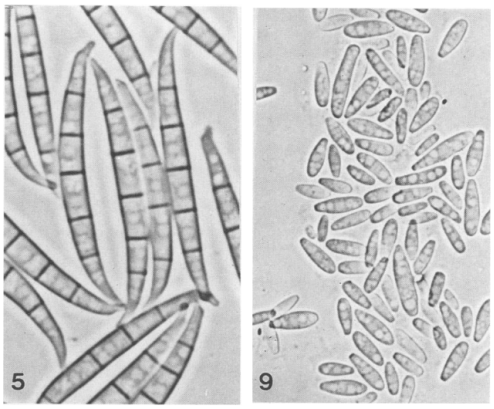


Figure 2.2 – Observation 5 : de macroconidies de *F.graminearum* (950X), 9 : de microcondies de *F.verticillioides* (avant *F.moniliforme* [5]) (1000X) qui sont des champignons qui font partie du Meta-*Fusarium sp.*. [6]

1.2 Communication au sein du Meta- *Fusarium sp.*

Au sein du Meta-*Fusarium sp.*. la communication s'effectue en partie par le biais de petits acides ribonucléiques comme les small ARN (smRNA) ou les micro ARN (miRNA). Ces derniers sont des courtes séquences de bases non codantes qui peuvent être séquencées.

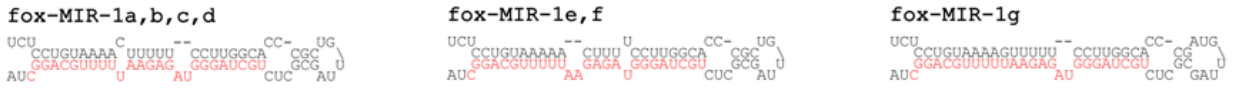


FIGURE 2.3 – Structure secondaire de quelques précurseurs miARN déjà mis en évidence chez les champignons du genre *Fusarium*.[7]

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 spécifiquement produits lors de la communication induite par la rencontre entre deux Fusaria de la même souche, nous pourrons mieux comprendre les mécanismes sous-jacents aux interactions entre les espèces de *Fusarium* toxinogènes. Et ainsi, 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 État de l'art

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.[8]

Il existe un autre outil utilisé MultiQC [9] 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.

Le QC doit être effectué à deux moments cruciaux : avant et après le nettoyage des données. L'utilisation de MultiQC peut illustrer les différences de qualité avant et après le nettoyage, permettant une évaluation précise de l'efficacité des procédures de nettoyage. Les métriques de qualité attendues en sortie de QC comprennent le nombre de séquences initiales, le nombre après nettoyage, le pourcentage d'ARN ribosomique (rRNA), le pourcentage d'ARN de petite taille (15-17 nt) qui peuvent indiquer la présence de produits de dégradation, le pourcentage de séquences de moins de 15 nt, et la distribution des tailles des miARN (par exemple, 21, 22, 23 nt). Ces métriques fournissent des indicateurs essentiels de la qualité des données et de l'efficacité du nettoyage, facilitant l'optimisation des étapes subséquentes du processus d'analyse.

Sorties attendues :

- Statistiques détaillées sur la qualité des données de séquençage
- Graphiques pour identifier les problèmes potentiels tels que la qualité des bases, la distribution des tailles d'insertion, la présence d'adaptateurs, etc.
- Format de sortie : Rapport HTML généré par FastQC.

2.2 Nettoyage des reads

Pour le nettoyage des lectures (reads) de séquençage, deux outils principaux se distinguent : Trimmomatic [10] et Cutadapt [11].

Tandis que Cutadapt est reconnu pour son efficacité à éliminer les adaptateurs et les bases de faible qualité, Trimmomatic offre une solution complète pour le nettoyage des lectures, incluant la suppression des adaptateurs, la coupe des bases de faible qualité, et l'élimination des lectures de mauvaise qualité à partir des extrémités.

Cutadapt est un outil performant destiné au nettoyage des séquences en éliminant les adaptateurs, les bases de faible qualité, ainsi que les séquences de mauvaise qualité des extrémités des lectures (reads). Il se distingue par sa spécialisation dans la détection et la suppression efficace des séquences d'adaptateurs. Cette spécificité fait de Cutadapt un choix privilégié pour les projets nécessitant une élimination précise et efficiente des adaptateurs.

Cutadapt génère des fichiers de séquences nettoyées, optimisant ainsi les données pour l'alignement ou d'autres types d'analyses subséquentes. Sa réputation s'appuie sur sa flexibilité et sa capacité à s'adapter à divers besoins spécifiques de nettoyage de séquences, rendant souvent Cutadapt plus simple à configurer et à utiliser par rapport à Trimmomatic pour certains projets.

Paramètres recommandés pour Cutadapt :

- Élimination des adaptateurs : Spécifiez les séquences d'adaptateurs à éliminer. Par exemple, -a ADAPTATEUR pour les adaptateurs en 3' ou -g ADAPTATEUR pour ceux en 5'.
- Qualité minimum : Utilisez -q QUALITÉ, pour couper les bases de faible qualité des extrémités des reads. QUALITÉ doit être remplacée par le score de qualité souhaité, comme -q 20 pour supprimer les bases ayant un score de qualité inférieur à 20.
- Longueur minimale : Définissez la longueur minimale d'un read pour être conservé avec -m LONGUEUR, où LONGUEUR est la taille minimale requise. Les reads plus courts que cette valeur seront éliminés.
- Recadrage : Si nécessaire, vous pouvez recadrer les reads à une longueur spécifique avec -l LONGUEUR, garantissant que tous les reads ont une taille uniforme après le traitement.

Sorties attendues :

- Fichiers de sortie contenant les reads nettoyés
- Format de sortie : Fichiers FASTQ prêts à être utilisés pour l'alignement ou d'autres analyses.

En conclusion, bien que Trimmomatic soit un outil puissant pour le nettoyage des lectures de séquençage, Cutadapt a été choisi pour sa spécialisation dans l'élimination des adaptateurs, sa facilité d'utilisation, et sa capacité à être finement ajusté pour répondre aux besoins spécifiques de notre projet.

2.3 Alignement des reads sur le génome

Pour ce projet qui se concentre sur l'analyse des petits ARNs (smRNA) non codants, environ 22 nucléotides de long, issus de co-cultures de céréales et de champignons du genre Fusarium, la sélection de l'outil d'alignement est cruciale pour identifier précisément les origines et les cibles potentielles de ces smRNAs dans les génomes des protagonistes. Les caractéristiques clés de ce projet—la courte longueur des séquences d'intérêt et la nécessité de distinguer finement les smRNAs spécifiques à chaque espèce dans des contextes de co-culture orientent vers certains critères pour le choix de l'outil d'alignement :

- Sensibilité et précision à aligner de courtes séquences : Étant donné la petite taille des smRNAs, l'outil d'alignement doit être optimisé pour aligner de très courtes séquences de manière précise.
- Capacité à gérer des séquences multimappées : Les smRNAs peuvent avoir des séquences qui s'alignent sur plusieurs régions du génome, rendant essentielle la capacité de l'outil à traiter efficacement les lectures multimappées.
- Performance et efficacité : Le traitement rapide des données est crucial, surtout lorsqu'on travaille avec un grand nombre d'échantillons typique des analyses à grande échelle.

En considérant ces critères, BWA-MEM2[12] semble être un choix judicieux pour ce projet. BWA-MEM2 est une version améliorée de BWA-MEM, connue pour son efficacité dans l'alignement de séquences courtes

et longues. Ses avantages sont particulièrement pertinents ici :

- Haute performance : BWA-MEM2 est optimisé pour les architectures de processeur modernes, offrant une vitesse d'alignement supérieure, ce qui est bénéfique pour traiter rapidement de grands jeux de données.
- Précision avec les courtes lectures : Bien que BWA-MEM (et par extension, BWA-MEM2) soit conçu pour être polyvalent, sa capacité à aligner précisément de courtes séquences en fait un choix adapté pour l'analyse de smRNAs.
- Gestion des lectures multimappées : BWA-MEM2 offre des options pour ajuster la manière dont les lectures multimappées sont traitées, permettant une analyse fine qui est essentielle pour distinguer les smRNAs provenant de différentes sources dans les co-cultures.

D'autre part, Bowtie2 [13] 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. Sorties attendues :

- Fichiers d'alignement indiquant où chaque read s'aligne sur le génome de référence
- Format de sortie : Fichiers SAM ou BAM.

En résumé, BWA-MEM2 est recommandé pour sa performance globale et sa capacité à traiter efficacement des séquences courtes et complexes comme celles rencontrées dans ce projet.

2.4 Identification des petits ARN produits

L'étape cruciale suivant l'alignement des données de séquençage de petits ARN (smRNA-seq) est l'identification et l'annotation des microARNs (miARNs) présents dans chaque échantillon, provenant de diverses situations de co-culture. Pour cela, deux outils principaux sont utilisés : miRDeep2[14] et miRBase [15], chacun ayant un rôle spécifique dans le processus d'analyse.

- Utilisation de miRDeep2 : miRDeep2 est un outil de bioinformatique conçu spécialement pour l'identification et la prédiction de nouveaux miARNs à partir des données de séquençage de smRNA. Il fonctionne en intégrant les séquences de smRNA alignées avec des informations sur la structure secondaire des précurseurs de miARN pour prédire de nouveaux miARNs et évaluer leur probabilité. Cette capacité à découvrir de nouveaux miARNs le rend particulièrement utile pour les recherches exploratoires où l'identification de miARNs inconnus peut fournir des insights novateurs sur les mécanismes moléculaires sous-jacents à l'interaction entre les espèces dans les co-cultures.
- Rôle de miRBase : En contraste, miRBase sert de base de données de référence exhaustive pour les séquences de miARNs déjà connues et annotées. Cet outil est crucial pour l'annotation des séquences de miARN détectées dans les échantillons de séquençage, permettant de relier les séquences observées à des fonctions et des rôles biologiques établis. L'utilisation de miRBase est essentielle pour distinguer les miARNs nouvellement découverts des séquences déjà documentées, facilitant ainsi la détermination de leur potentiel biologique et fonctionnel.
- Approche combinée : Pour une analyse complète des données de smRNA-seq issues de co-cultures, une approche combinée qui utilise à la fois miRDeep2 pour la découverte de nouveaux miARNs et miRBase pour l'annotation des miARNs connus est recommandée. Cette stratégie permet de maximiser la couverture analytique, offrant à la fois la possibilité de découvrir de nouvelles entités régulatrices et de relier les miARNs détectés à des connaissances biologiques existantes. Ce processus d'identification et d'annotation est fondamental pour comprendre les mécanismes moléculaires par lesquels les petits ARN facilitent la communication entre les espèces dans les co-cultures, ouvrant la voie à des stratégies de prévention des contaminations par des mycotoxines dans les céréales.

Sorties attendues :

- Liste des miARN identifiés dans chaque échantillon
- Format de sortie : Rapport ou fichier texte.

2.5 Quantification des petits ARN

L'étape de quantification de l'expression des petits ARN repose sur l'évaluation des niveaux d'expression des séquences identifiées, en s'appuyant sur des outils spécialisés tels que HTSeq[16] ou featureCounts [17]. Ces outils analysent les données de séquençage d'ARN à petite échelle pour estimer l'abondance des ARN

dans les échantillons.

L'analyse inclura notamment la quantification des ARN ribosomaux et la détermination de leur distribution et abondance relative dans les échantillons.

— Utilisation de HTSeq : HTSeq est un outil développé en Python, spécialement conçu pour traiter les données issues de séquençage à haut débit, y compris les données de séquençage de petits ARN. Il fonctionne en prenant les lectures séquencées alignées (au format BAM) et les compare à un fichier d'annotation de référence pour attribuer chaque lecture à une caractéristique génomique spécifique, telle que les gènes ou les petits loci d'ARN.
— Fonctionnalités de featureCounts : D'autre part, featureCounts, qui fait partie intégrante du package Subread, est également utilisé pour cartographier et quantifier les lectures de séquences. Comme HTSeq, featureCounts utilise des lectures alignées et un fichier d'annotation pour assigner les lectures à des caractéristiques génomiques définies. L'outil génère une matrice de comptage où chaque ligne représente une caractéristique génomique (par exemple, un petit ARN) et chaque colonne un échantillon, affichant le nombre de lectures mappées à chaque caractéristique pour chaque échantillon. Une fonctionnalité clé de featureCounts est sa capacité à effectuer le comptage de lectures spécifique à un brin, une propriété essentielle pour la quantification précise des petits ARN, surtout en considération de la transcription antisens.
— Choix entre HTSeq et featureCounts : Pour ce projet, où la distinction précise entre les petits ARN issus des co-cultures est critique, featureCounts est privilégié en raison de sa capacité à effectuer des quantifications spécifiques à un brin. Cette fonctionnalité est particulièrement pertinente pour les petits ARN, dont l'orientation peut être cruciale pour leur fonction biologique et leur identification correcte. La précision dans la quantification de l'expression des petits ARN est fondamentale pour analyser leur rôle dans la communication inter-espèces.

Sorties attendues :
— Matrice de comptage des lectures de séquences attribuées à chaque caractéristique génomique dans chaque échantillon
— Format de sortie : Matrice de comptage (par exemple, format texte ou tableur).

2.6 Normalisation

L'objectif principal de la normalisation est d'ajuster les différences de profondeur de séquençage et les biais de composition entre les échantillons. Ceci est essentiel pour une analyse précise en aval, en particulier lors de la comparaison des niveaux d'expression entre échantillons ou conditions. EdgeR[18] et DESeq2[19] proposent tous deux des méthodes de normalisation largement utilisées dans l'analyse de séquençage d'ARN. Ces outils utilisent des modèles statistiques pour tenir compte des effets de profondeur de lecture et de composition d'ARN.

— EdgeR : Utilise une méthode basée sur l'échelle des tailles de bibliothèque normalisées. La fonction calcNormFactors dans EdgeR calcule des facteurs d'échelle pour chaque bibliothèque qui sont ensuite utilisés pour ajuster les comptages de lecture. DESeq2 : Fournit une méthode de normalisation des facteurs de taille qui est calculée en utilisant la méthode de médiane des ratios. Cela permet d'ajuster simultanément la profondeur de séquençage et la composition d'ARN.
— Sorties attendues :les données lues de séquençage réelles représentées au format BAM ne sont pas modifiées en termes de normalisation. La normalisation affecte généralement la façon dont vous interprétez les données (comme l'ajustement du nombre de lectures dans l'analyse), et non les lectures brutes elles-mêmes.

2.7 Analyse différentielle

Identification des petits ARN différentiellement exprimés entre les différentes conditions de co-culture à l'aide d'outils comme DESeq2 [19] ou edgeR[18].

— Utilisation de DESeq2 et edgeR DESeq2 et edgeR sont des packages de bioinformatique conçus spécifiquement pour analyser les variations d'expression génique ou, dans ce cas, des petits ARN, à partir de données de séquençage d'ARN. Ces outils appliquent des modèles statistiques pour identifier les caractéristiques génomiques—ici, les petits ARN—qui montrent des changements d'expression significatifs

entre les conditions expérimentales étudiées. DESeq2 utilise un modèle basé sur les négatives binomiales pour estimer la variance et la moyenne des données d'expression, permettant une comparaison précise entre les groupes. EdgeR, de son côté, emploie également une approche basée sur les négatives binomiales, avec une emphase sur la normalisation des tailles d'échantillons et une estimation robuste de la dispersion pour identifier les différences d'expression.
— Vérification de la validité de l'approche : L'adéquation de DESeq2 ou edgeR pour l'analyse différentielle des petits ARN dans le contexte des co-cultures de céréales et Fusarium dépend de plusieurs facteurs, notamment la normalisation appropriée des données et la précision dans l'estimation de la dispersion. Ces facteurs sont cruciaux étant donné la complexité des interactions et le potentiel de variation entre les échantillons.

Sorties attendues :
— Liste des miARN différentiellement exprimés avec leur niveau de significativité statistique
— Format de sortie : Tableau ou fichier texte.

2.8 Visualisation des résultats

Matplotlib, en tant que bibliothèque de visualisation de données pour Python, offre une flexibilité et une variété d'options pour présenter ces données de manière informative et visuellement attrayante. Voici quelques types de visualisations spécifiquement adaptées à l'analyse de séquençage de petits ARN qui pourraient être envisagées :
— Histogrammes de distribution
— Graphiques en boîte (Boxplots)
— Graphiques de dispersion (Scatter plots)
— Heatmaps
— Diagrammes de Volcano

Sorties attendues :
— Graphiques représentant les résultats de l'analyse de séquençage
— Format de sortie : Graphiques (par exemple, fichiers image ou intégration dans un rapport).

Le pipeline sera livré sous la forme de scripts Python, éventuellement accompagnés de fichiers de configuration ou de documentation supplémentaire. Ces scripts seront conçus pour être exécutés de manière séquentielle ou modulaire, selon les besoins de l'utilisateur.

En ce qui concerne le dépôt, GitHub est souvent utilisé comme plateforme populaire pour héberger des projets de développement logiciel. Nous pourrions créer un référentiel GitHub dédié pour stocker le code source du pipeline, ainsi que toute documentation supplémentaire ou des exemples d'utilisation. Cela permettra aux utilisateurs de télécharger facilement le pipeline, de suivre les mises à jour et de contribuer au développement si nécessaire.

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.
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.
— Sortir un fichier BigWig compatible avec les plates-formes de navigateur de génome fonctionnelles existantes comme JBrowse pour garantir une visualisation adéquate.

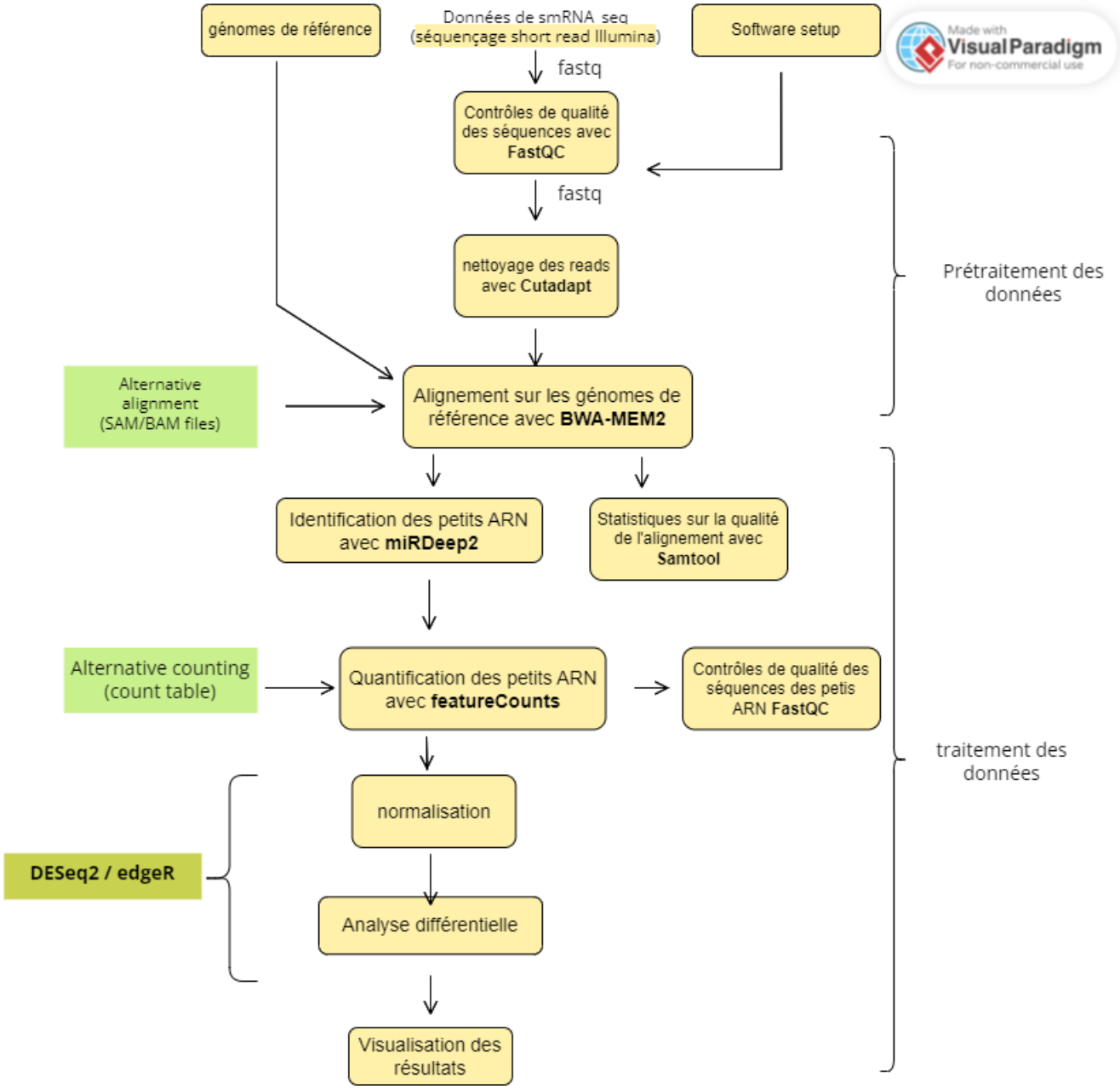
3.4 Ajouts optionnels

— 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

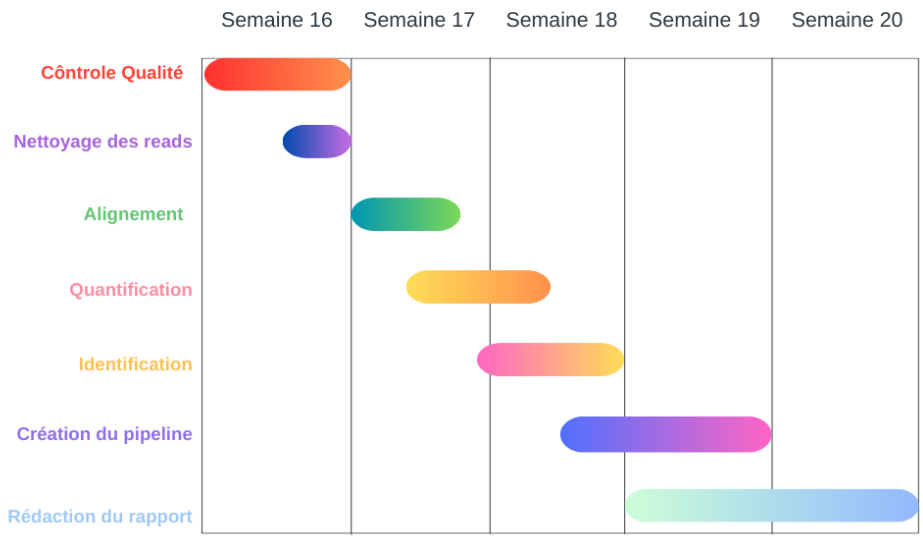


FIGURE 4.1 – Diagramme de Gantt qui nous servira de base pour l'organisation de notre travail

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] Keith A Seifert, Takayuki Aoki, R P Baayen, D Brayford, L W Burgess, S Chulze, W Gams, D Geiser, J de Gruyter, J F Leslie, A Logrieco, W F O Marasas, H I Nirenberg, K O'Donnell, J Rheeder, G J Samuels, B A Summerell, U Thrane, and C Waalwijk. The name fusarium moniliforme should no longer be used. *Mycological Research*, 107(6) :643–644, 2003.

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

[7] 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.

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

[9] 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.

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

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

[12] Heng Li. Bwa-mem2 : Faster and more accurate read alignment to large reference genomes. Accès en ligne, Year. Disponible sur : https ://github.com/bwa-mem2/bwa-mem2.

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

[14] 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.

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

[16] 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.

[17] 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.

[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.

[19] 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.