

Analyse de données RNA-Seq (champignons *Fusarium*)

Université de Bordeaux - Master Bio-informatique

Alani Maroa, Khodja Linda, Ouandaogo Djemilatou, Piat Lucien

May 23, 2024

Superviseur: Marie Beurton-Aimar

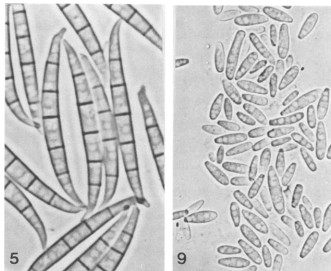
Clients: Dumetz Fabien, Ponts Nadia

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

- **Contexte de recherche** : Comprendre les mécanismes moléculaires chez les espèces de *Fusarium*.
- **Importance** : Espèces de *Fusarium* en agriculture, contamination des cultures.

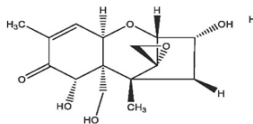
Comment peut-on aider l'INRAE dans sa recherche sur la sécurité alimentaire?

- **Objectif** : Analyser les séquences de petits ARN dans différents scénarios de culture de Fusarium.



(5) : Macroconidies de *F. graminearum* (950X), (9) :
Microconidies de *F. verticillioides* (*F. moniliforme*) (1000X).
Ils sont des champignons qui font partie du MetaFusarium
sp.[6, 13, 10].

- **Contexte biologique du genre *Fusarium* :**
Pathogènes affectant les cultures, produisant des mycotoxines.
- **Changement d'échelle, le Meta-*Fusarium* sp. :** Révision du paradigme “un pathogène - une maladie” [11, 9].



Deoxynivalenol = DON (MW=296)

Formule du Désoxynivalénol, une molécule de la famille des B-trichothécènes [6].

fox-MIR-1g

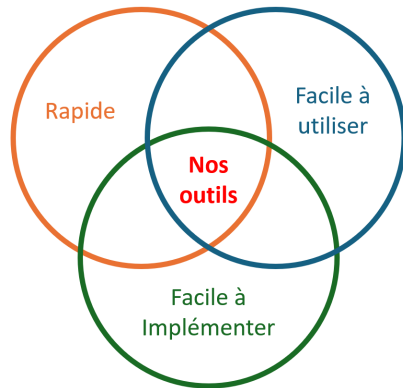


Structure secondaire d'un précurseurs miARNs déjà mis en évidence chez les champignons du genre *Fusarium* [5].

- **Communication au sein de Meta-Fusarium sp. :** Petits ARN comme les smRNAs et miRNAs.
- **La production des toxines :** Nombreuses et très stables, de la famille des B-trichothécènes.

Taches classiques, nombreux outils disponibles :

- **Outils de contrôle de qualité** : FastQC [1, 2], RSeQC.
- **Outils d'alignement** : STAR, HISAT2, TopHat, BWA, BWA_MEM2 [14, 7, 4].
- **Identification de miRNA** : miRDeep2, ShortStack [3].
- **Quantification et analyse d'expression différentielle** : DESeq2 [8], edgeR [12], Cuffdiff.



Analyse - Présentation des données

Confrontation	Name	i7 RUDI	index_i7	i5 RUDI	index_i5
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA349	C1	i7RUDI-481	GTTTCGACAAT	i5RUDI-481	AAGAGGAGAT
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA349	C2	i7RUDI-482	TGGTAGGTGG	i5RUDI-482	CCATGAGTCG
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA812	C3	i7RUDI-483	GTAACCGATC	i5RUDI-483	TGCGATACGC
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA812	C4	i7RUDI-484	CACCTCACCA	i5RUDI-484	GTTCTCCATA
<i>F. graminearum</i> INRA812 / <i>F. graminearum</i> INRA812	C5	i7RUDI-485	CCTGATTGTT	i5RUDI-485	CCTTGGAGCT
<i>F. graminearum</i> INRA812 / <i>F. graminearum</i> INRA812	C6	i7RUDI-486	TGCACACCAG	i5RUDI-486	AGACGGTTGG
<i>F. graminearum</i> INRA349 / <i>F. verticillioides</i> INRA63	C9	i7RUDI-489	AGCCTGTATT	i5RUDI-489	TAGCATCGAT
<i>F. graminearum</i> INRA349 / <i>F. verticillioides</i> INRA63	C10	i7RUDI-490	GAAGGCAACG	i5RUDI-490	CGTATCTGCG
<i>F. verticillioides</i> INRA63 / <i>F. verticillioides</i> INRA63	C17	i7RUDI-497	TTCAATCGCT	i5RUDI-497	AATTGCGCAT
<i>F. verticillioides</i> INRA63 / <i>F. verticillioides</i> INRA63	C18	i7RUDI-498	TTGGCCAATG	i5RUDI-498	TTAATCCTCG
<i>F. graminearum</i> INRA156 / <i>F. graminearum</i> INRA156	C23	i7RUDI-503	GCATAAGGCG	i5RUDI-503	TCCTGTCAAC
<i>F. graminearum</i> INRA156 / <i>F. graminearum</i> INRA156	C24	i7RUDI-504	AGGAGGCGTA	i5RUDI-504	CACGCTGTCA
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA156	C25	i7RUDI-505	CGTACTCATT	i5RUDI-505	GTACCTTGT
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA156	C26	i7RUDI-506	TAAGCGCGCT	i5RUDI-506	TACCGTGTGT
<i>F. graminearum</i> INRA349 / <i>F. graminearum</i> INRA156	C27	i7RUDI-507	AGACTACTTG	i5RUDI-507	AGGTGTTACG
<i>F. graminearum</i> INRA349	C28	i7RUDI-508	TACGCACTGC	i5RUDI-508	CTAGGTTGAC
<i>F. graminearum</i> INRA812	C29	i7RUDI-509	GCCTTACAA	i5RUDI-509	GCCTAGATTA
<i>F. graminearum</i> INRA812	C30	i7RUDI-510	ATTCGGATCT	i5RUDI-510	TATCACTGG
<i>F. graminearum</i> INRA156	C31	i7RUDI-511	CGGTAGCCT	i5RUDI-511	TTGGAATGGT
<i>F. graminearum</i> INRA156	C32	i7RUDI-512	CCTCCTCTTG	i5RUDI-512	GACAATAACG
<i>F. verticillioides</i> INRA63	C39	i7RUDI-519	TCCTCCGTCA	i5RUDI-519	GCAGGCTTAA
<i>F. verticillioides</i> INRA63	C40	i7RUDI-520	GTATGTCGCT	i5RUDI-520	CGAGTACAGG

Description du jeu de données

- 14 lots de co-culture, 8 lots de monoculture.
- Espèces utilisées : *Fusarium graminearum* et *Fusarium verticillioides*.
- Format FASTQ compressé.
- Illumina Short Read.

Liste des demandes du client vis-à-vis du programme :

- **Besoins fonctionnels** : Acquisition de données, prétraitement, normalisation.
- **Besoins d'analyse** : Alignement, identification de miRNA, quantification.
- **Besoins de visualisation** : Profils d'expression, analyse comparative.

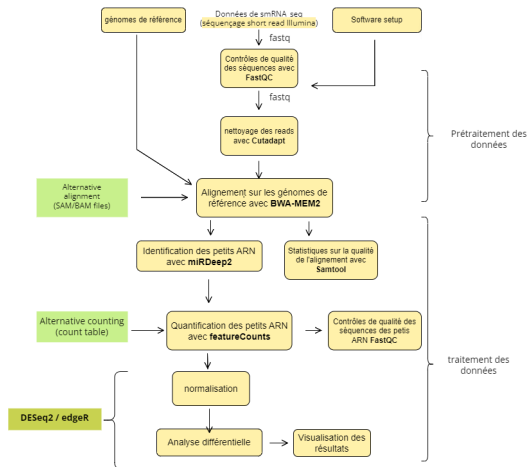
Contraintes :

- Utilisation d'outils open-source.
- Réalisation dans un cadre pseudo-restreint.
- Production de fichiers BigWig.

Le choix du format de pipeline :

- Garantit une flexibilité maximale
- Architecture modulaire et possibilités d'adaptation futures
- Permet de lancer des scripts dans des langages différents

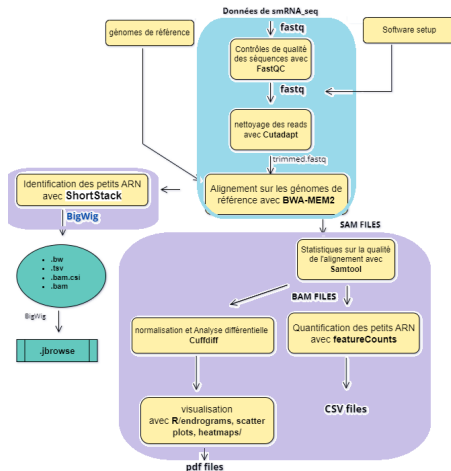
A droite : Le pipeline que nous avons initialement prévu



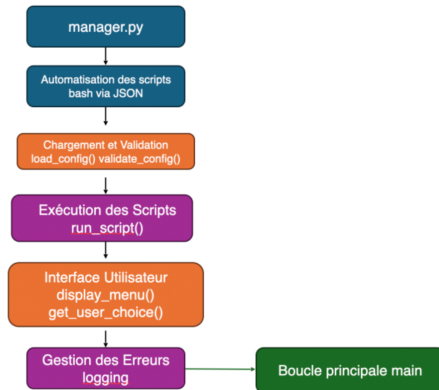
Le pipeline final :

- Outils plus simples d'utilisation
- Minimise la conversion des formats d'input/output
- Rapidité accrue

A droite : Le pipeline final après révisions



- **Le fichier JSON** Permet une manipulation externe des scripts
- **Le script python manager.py :**
 - Centralise et automatise l'exécution des scripts.
 - Réduit les erreurs manuelles
 - Augmente l'efficacité
 - Facilite le suivi et le contrôle (logging)
 - Permet une modularité et flexibilité

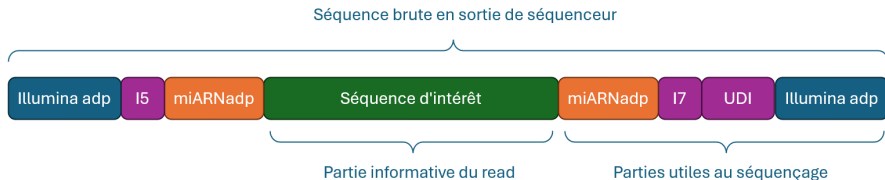


Effectué en amont des autres traitements, le contrôle qualité rapide mais crucial :

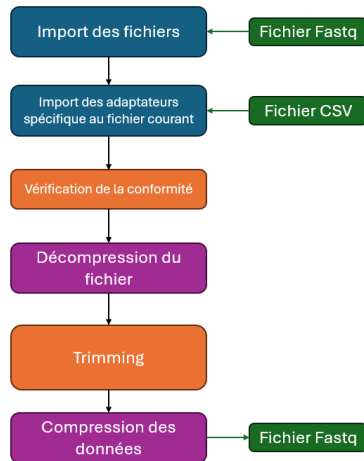
- **Pourquoi :**
 - Afin d'avoir une vision globale de la qualité des données
 - Pour éliminer d'office les potentielles erreurs
 - S'assurer que le séquençage se soit bien passé
- **L'outil Fastqc :** Le plus classique et efficace pour ce genre de traitement. Outil rapide qui produit des sorties imagées en HTML.
- **Le script :** Écrit en Bash, boucle sur tous les fichiers et génère un rapport pour chacun d'entre eux. Dans le terminal, s'affiche l'état d'avancement en temps réel.

Mise en œuvre - Le nettoyage des séquences

- **Pourquoi** : Afin de ne garder que les données exploitables et maintenir un contexte biologique cohérent.
- **Nombreux traitements nécessaires** : Retrait des séquences de mauvaise qualité, longueur inadéquate et retrait des adaptateurs.
- **Nombreux adaptateurs** : Introduit au cours du processus expérimental sont parfois spécifiques à l'échantillon.



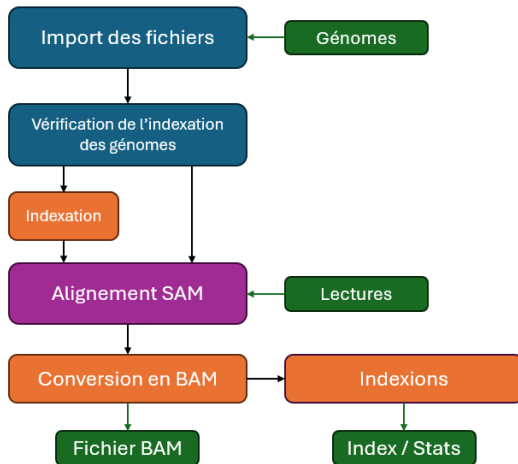
- **L'outil Cutadapt** : Spécialisé dans le retrait des adaptateurs, utilise la transformée de Burrows-Wheeler
- **Les paramètres de l'outil** : Taille min/max, qualité minimale et séquences à retirer.
- **Le script** : Écrit en Bash, ajout dynamique des adaptateurs à partir d'un fichier CSV. Vérification du format IUPAC, décompression, trimming et compression.



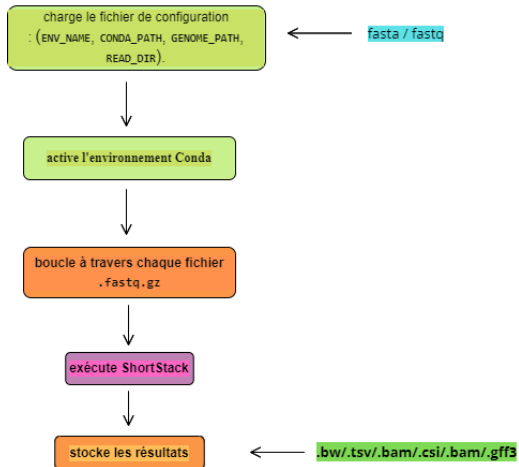
Une fois les jeux de données prêts, nous passons à l'alignement :

- **Pourquoi :** Avoir une vision globale de la position des séquences sur le génome
- **L'outil BWA-MEM2 :** Aligne les fichiers de reads avec les génomes de référence et produit des fichiers SAM
 - Utilise la transformée de Burrows Wheeler
 - Implémente le parallélisme
 - Comporte une phase de seeding.

- **SAMtools:** Traite les fichiers SAM produits par alignement.sh
 - Les convertit en fichiers BAM
 - Les trie
 - Les indexe
 - Et génère les statistiques d'index.



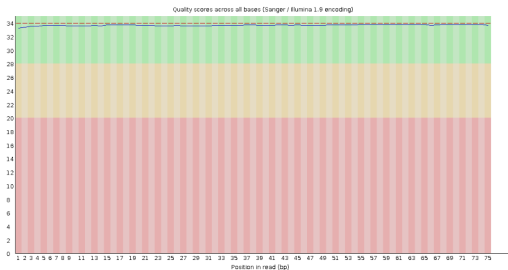
- **L'outil ShortStack** : Identification et la quantification des petits ARN.
- **Les paramètres de l'outil** : `-genomefile` / `-readfile` / `-outdir`.
- **Le script**: Ecrit en Bash, charge les paramètres, active l'environnement Conda, puis exécute ShortStack pour chaque read, en stockant les résultats dans des répertoires de sortie organisés par timestamp.



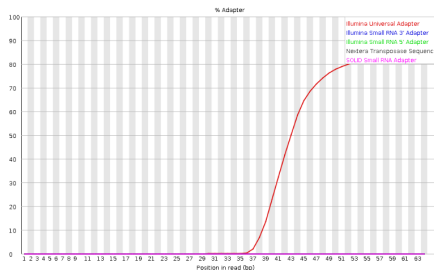
- **Cuffdiff** : Comparer les expressions géniques entre différentes conditions et produire des fichiers de résultats les gènes différentiellement exprimés.
- **Le script R** : Analyser les données de comptage des ARN à partir des fichiers générés par cuffdiff.sh et produire une dataframe pour des analyses supplémentaires.

- **Résultats du contrôle de qualité :** Résumé de la qualité initiale des données.

✓ Per base sequence quality

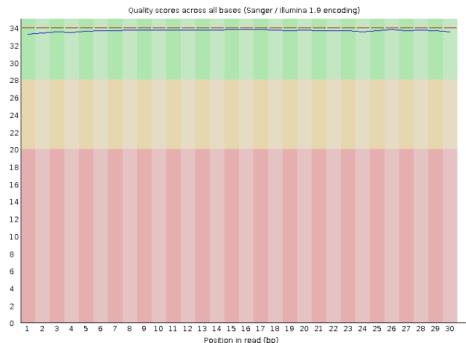


✗ Adapter Content

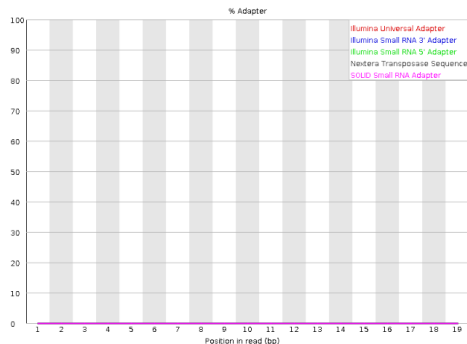


- **Résultats du trimming** : Contrôle qualité après de trimming.

✓ Per base sequence quality



✓ Adapter Content



- **Résultats de l'alignement** : Données extraites du fichier "idxstats.txt".

Reference	Length	Mapped Reads	Unmapped Reads
HG970332	11760891	83343	0
HG970333	8997558	56096	0
HG970334	7792947	38217	0
HG970335	9395062	43437	0
HG970330	5846	0	0
HG970331	95638	7100	0
* 0	0	1030357	

- **Résultats de miRNA** : miRNAs identifiés et quantifiés.

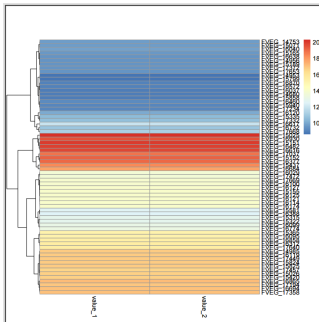
Cluster ID	Chromosome	Début	Fin	Longueur	Reads Alignés	Séquence Majeure
Cluster_1	CM000578.1	50417	50831	415	1	ACUCCCACUGAGGCG
Cluster_2	CM000578.1	72304	72718	415	1	UUUACAGCGCAGAU
Cluster_11	CM000578.1	473255	473879	625	3118	GAAUGGCUCAGUGAGGCGUC
Cluster_14	CM000578.1	548257	548731	475	23851	UCUUCCGUAGUAUAGUGGUC

- **Analyse d'expression différentielle : Résultats clés.**

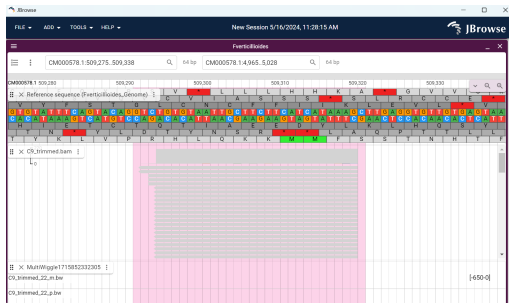
7728	FVEG_15024-t26_1	FVEG_15024	-	CM000583.1:1468234-1471239	condition1	condition2	NOTEST	0	0	0	0	1	1	no
7729	FVEG_15025-t26_1	FVEG_15025	-	CM000583.1:1467872-1467956	condition1	condition2	OK	1.29659e+06	1.29659e+06	0	0	1	1	r
7730	FVEG_15026-t26_1	FVEG_15026	-	CM000583.1:1467243-1467325	condition1	condition2	OK	109573	109573	0	0	1	1	no
7731	FVEG_15027-t26_1	FVEG_15027	-	CM000583.1:1453632-1453704	condition1	condition2	NOTEST	0.154869	0.154869	0	0	1	1	1
7732	FVEG_15028-t26_1	FVEG_15028	-	CM000583.1:1450813-1451407	condition1	condition2	NOTEST	0	0	0	0	1	1	no

Résultats et Discussion - Quelques visualisations

- **Visualisations** : Heatmaps, dendrogrammes, fichiers BigWig



Heatmap



BigWig

Difficultés rencontrées:

- Remplacement d'outils complexes: miRDeep2 par ShortStack, DESeq2 et edgeR par Cuffdiff, impactant le calendrier.
- Temps de calcul long, nécessitant l'utilisation de sous-ensembles de données.
- Manque de visualisations essentielles dû à des défis techniques et au manque de temps.















Perspectives d'amélioration:

- Optimiser outils et workflows pour simplicité, compatibilité et efficacité.
- Améliorer l'utilisation des ressources computationnelles.
- Développer des visualisations dynamiques et informatives pour une meilleure interprétation et communication des résultats.

Grâce à ce projet, nous avons pu créer un pipeline de NGS complet :

- **Le pipeline :**
 - Rapide et facile d'utilisation
 - Outils récents et de pointe
- **Travail en groupe :** De grande envergure.
- **Contributions à la compréhension des interactions des Fusarium.**
 - Faire tourner le programme sur toutes les données dans un cluster de calcul
 - Cibler les gènes concernés

Références – Merci pour votre écoute !

-  S. Andrews.
FastQC: A quality control tool for high throughput sequence data.
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, 2010.
-  S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett.
FastQC.
Babraham Institute, Jan. 2012.
-  M. J. Axtell.
Shortstack: Comprehensive annotation and quantification of small rna genes.
RNA, 19(6):740–751, Apr. 2013.
-  bwa mem2.
bwa-mem2 github page/readme, 2022.
-  R. Chen, N. Jiang, Q. Jiang, X. Sun, Y. Wang, H. Zhang, et al.
Exploring microrna-like small rnas in the filamentous fungus fusarium oxysporum.
PLoS ONE, 9, 2014.
-  M. A. Gab-Allah, Y. Getachew Lijalem, H. Yu, S. Lee, S.-Y. Baek, J. Han, et al.
Development of a certified reference material for the accurate determination of type b trichothecenes in corn.
Food Chemistry, 404, 2023.
-  H. 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>.
-  M. I. Love, W. Huber, and S. Anders.
Moderated estimation of fold change and dispersion for rna-seq data with deseq2.
Genome biology, 15(12):550, 2014.
-  C. MycSA.
Le projet anr 2022-2026 teamtox, 2023.
-  P. E. Nelson, M. C. Dignani, and E. J. Anaissie.
Taxonomy, biology, and clinical aspects of fusarium species.
Clinical microbiology reviews, 1994.
-  N. Ponts, L. Couedelo, L. Pinson-Gadais, M.-N. Verdal-Bonnin, C. Barreau, and F. Richard-Forget.
Fusarium response to oxidative stress by h2o2 is trichothecene chemotype-dependent.
FEMS Microbiology Letters, 293:255–262, 2009.
-  M. D. Robinson, D. J. McCarthy, and G. K. Smyth.
edgeR: a bioconductor package for differential expression analysis of digital gene expression data.
Bioinformatics, 26(1):139–140, 2010.
-  K. A. Seifert, T. 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.
-  M. Vasimuddin, S. Misra, H. Li, and S. Aluru.
Efficient architecture-aware acceleration of bwa-mem for multicore systems.
In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, May 2019.