



Master Bioinformatique
Université de Montpellier

HAU803I : Rapport de stage de M1

---

# Titre à choisir

---

***Etudiant :***
Mickael Coquerelle

***Professeur :***
Anthony Boureux

**Résumé**

a remplir

# Liste des abréviations

**Acronymes**

| | |
|---|---|
| **ADN** | Acide DésoxyRiboNucléique |
| **ARN** | Acide RiboNucléique |
| **API** | Application Programming Interface |
| **CVS** | Concurrent Versions System |
| **FP** | Faux Positifs |
| **FN** | Faux Négatifs |
| **KB** | KiloBase |
| **SLA** | Sclérose Latérale Amyotrophique |
| **RCS** | Révision Control System |
| **SHD** | Séquençage Haut Débit |
| **SNP** | Polymorphisme nucléotidique unique |
| **SIF** | Singularity Image Format |
| **UML** | Language de modélisation unifié |
| **VP** | Vrai Positifs |
| **VN** | Vrai Négatifs |

**Symboles**

| | |
|---|---|
| $\mathcal{A}_{\mathrm{x}}$ | Alphabet de x |
| $\Sigma_{\mathrm{x}}$ | Somme de x |
| $Q_P$ | Score de qualité Phred |
| $Q_A$ | Score de qualité Phred encodé en ASCII |
| $S$ | Séquence biologique |
| $P$ | Motif recherché |
| $S_e$ | Sensibilité |
| $S_p$ | Spécificité |
| $\mathcal{T}$ | Texte |
| $w$ | Mot |
| $\mathcal{O}()$ | Notation de Landau |
| $\mathcal{SA}_x$ | Table des suffixes de x |

# Table des matières

# 1 Introduction

## 1.1 Positionnement du stage

Ce travail est la synthèse de mon stage de première année de Master, durant lequel j'ai intégré l'équipe de la professeure Thérèse Commes du laboratoire Bio2M, rattaché à l'Institut national de la santé et de la recherche médicale (INSERM). J'ai eu la chance d'être accompagné dans mon apprentissage et mes travaux par Anthony Boureux, enseignant-chercheur. Ce laboratoire a la particularité d'être en interaction étroite avec des services cliniques et des plateformes hospitalières, ce qui favorise évidemment l'innovation et la résolution de questions liées au champ médical. C'est dans ce contexte, j'ai eu l'occasion de contribuer à un projet de recherche translationnelle que Bio2M mène avec le CHU de Nîmes. Ce travail est en lien direct avec des enjeux diagnostiques, puisqu'il concerne une maladie neurodégénérative : la sclérose latérale amyotrophique (SLA). La problématique de mon stage s'inscrit dans le champ de la transcriptomique, et plus précisément dans le cadre de l'analyse d'expression génique appliquée à la SLA, l'idée étant d'initier, proposé, une stratégie pour détecter une expression différentielle à l'échelle de certains gènes d'intérêt, avec des contraintes à la fois technique et biologiques, nous le verrons.

## 1.2 Contexte biologique

Une maladie rare se définit par une prévalence de 0,05 % dans la population générale. Quatre-vingts pour cent de ces maladies sont d'origine génétique [2]. La SLA en fait partie : elle touche un individu sur 20 000 en Europe [3], et sa prévalence mondiale varie de 1,57 à 11,8 pour 100 000, selon les pays, de l'Iran aux États-Unis [1]. C'est une maladie neurodégénérative causée par une atteinte du motoneurone central au niveau du cortex cérébral, et par une dégénérescence progressive des fonctions musculaires. Cette pathologie est très handicapante, tant physiquement que socialement.

En raison de sa gravité et des conséquences dévastatrices pour les patients et leur entourage, elle constitue un domaine de recherche de premier plan pour les généticiens. C'est pourquoi il est intéressant d'intégrer une approche transcriptomique dans le diagnostique des formes génétique de la maladie pour mieux comprendre la pathologie.

Le périmètre de ce travail préliminaire était de faire un état des lieux des méthodes conventionnelles de normalisation, pour *in fine* détecter une éventuelle expression génétique différentielle à partir des données issues du séquençage RNA-Seq. En effet, la transcription, étant fondatrice de la diversité protéomique, est par nature source de nombreuses anomalies génétiques. À l'issue de ce mécanisme, un gène peut exprimer plusieurs isoformes dont l'impact peut être pathologique. Identifier de telles anomalies implique la création d'un protocole de séquençage spécifique et d'un pipeline d'analyse RNA-Seq. L'objectif, à terme, est de déployer ce *workflow* intégrant une dimension transcriptomique pour le diagnostique de la SLA.

Durant mon premier semestre, j'ai eu l'occasion d'effectuer un travail de recherche bibliographique

pour étayer ma compréhension des différentes approches permettant de mettre en œuvre cette étape. La synthèse de ce travail préliminaire peut s'apprécier à travers la figure ci-dessous :

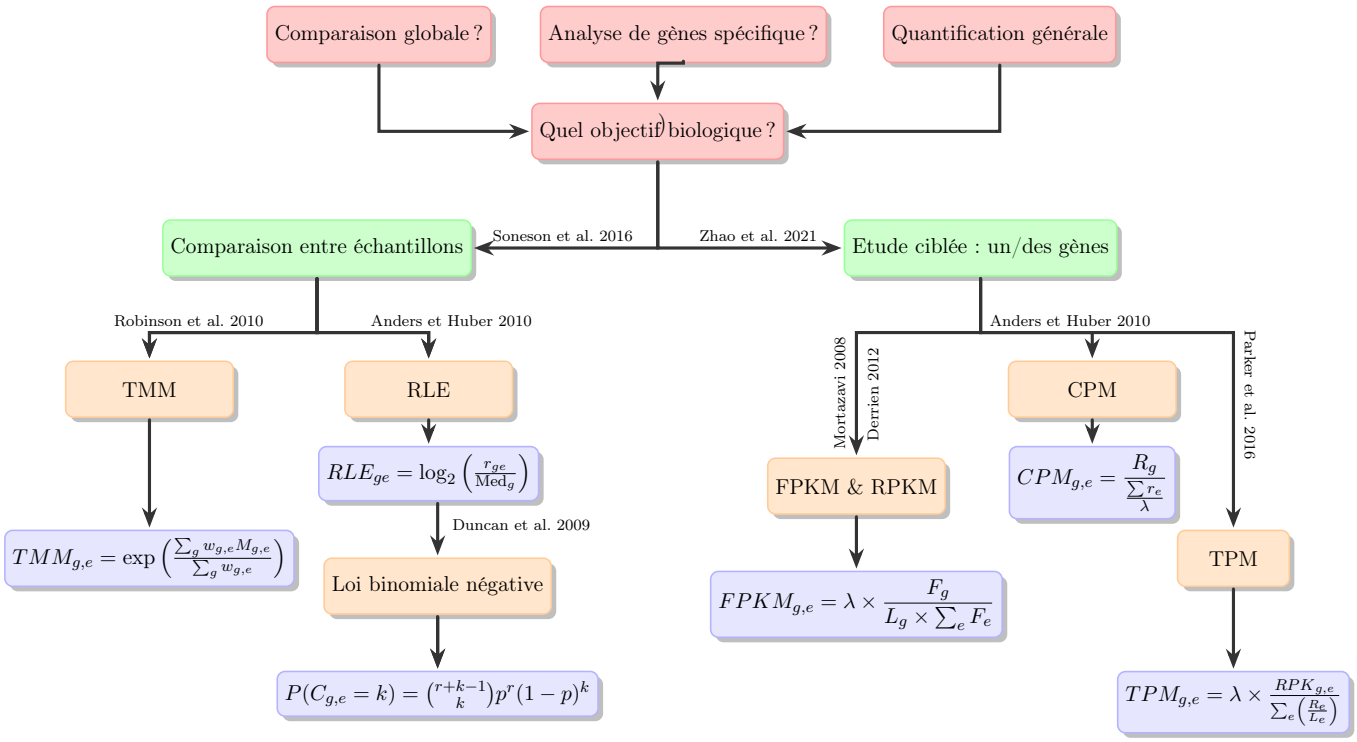


**Figure 1 :** *Schéma décisionnel pour normaliser les données RNASeq*

doit être apprécié en considérant que chaque méthode a ses avantages et inconvénients, et le choix sera fait sur un certain nombre de critères, en tenant compte des spécificités des données et de l'objectif(s) biologique(s). Ce sujet sur la normalisation des données RNA-Seq dans le cadre du panel de gènes ciblés nous a permis de mettre en lumière la complexité de cette étape de normalisation des données RNA-Seq et la nécessité d'une collaboration étroite entre biologistes et bioinformaticiens pour faire parler efficacement les résultats expérimentaux

## 1.3 Contexte expérimental et contraintes

Présentation du jeu de données : 72 échantillons de transcriptomique ciblée (56 gènes SLA). Organisation en runs de séquençage (n = 8 par run). Limites méthodologique, absence de gènes de référence stables, peu de données pour une normalisation robuste. Constats initiaux : forte variabilité de profondeur de lecture inter-run, faible reproductibilité.

# 2 Matériels & Méthodes

## 2.1 Génération des fichiers d'alignement

Stratégie de génération automatisée des fichiers BAM à partir des paires FASTQ (1 & 2).
Structure de traitement par run (batch de 8 échantillons).
Mise en place d'un pipeline reproductible.

## 2.2 Analyse de la variabilité expérimentale

Statistiques descriptives sur la profondeur de lecture, par run et par échantillon.
Mise en évidence de la non-reproductibilité : inter-run vs. intra-run.
Visualisations pour illustrer la disparité de couverture.

## 2.3 Alignement : levée d'ambiguïté sur la variabilité

Présentation de STAR et CRAC :
Méthodologie, hypothèses sous-jacentes, différences de traitement.
Alignement des mêmes échantillons avec les deux outils.
Comparaison des métriques de mapping : taux d'alignement unique/multiple/non-aligné.
Conclusion : validation que l'étape d'alignement n'est pas responsable de la variabilité observée.

## 2.4 Vers une approche alternative de quantification

Justification du besoin de s'affranchir des biais d'alignement.
Introduction à la quantification par k-mer : promesse de neutralité méthodologique.
Transition vers une stratégie plus robuste de détection différentielle.

### 2.4.1 Qualité et spécificité de l'alignement

L'étape de « *mapping* » dans un processus d'analyse est une étape charnière et pour notre cas d'étude on peut raisonnablement penser que cette étape peut biaiser le signal biologique réel,

Pour évaluer si les deux outils d'alignement, STAR et CRAC, diffèrent significativement dans leur capacité à aligner les lectures uniques, j'ai réalisé une analyse statistique sur les proportions relatives des lectures uniques par patient.J'ai d'abord effectué un test de normalité de Shapiro-Wilk sur les données de proportions de lectures uniques pour chaque outil afin de vérifier l'hypothèse de normalité nécessaire pour utiliser un test paramétrique. Les résultats ont validé cette hypothèse, j'ai donc pu appliquer un test t de Student apparié pour comparer les moyennes des proportions

de lectures uniques entre STAR et CRAC. Le test t n'a pas montré de différence statistiquement significative ($p > 0.05$) entre les deux outils, ce qui suggère que, globalement, leur capacité à aligner des lectures uniques est comparable dans mon jeu de données.

Cependant, je tiens à souligner que ces résultats quantitatifs ne reflètent pas nécessairement les différences qualitatives dans les approches d'alignement de STAR et CRAC. En effet, ces deux outils reposent sur des algorithmes différents, avec des stratégies distinctes pour gérer les lectures multiples, les lectures chimériques, et les jonctions d'épissage. Ces différences peuvent influencer la localisation précise des alignements, la sensibilité à certains types de variants, ainsi que la qualité globale des données alignées. Par conséquent, même si la proportion de lectures uniques est comparable, j'interprète ces résultats avec prudence. Pour les analyses de quantification ultérieures, je recommande d'intégrer une étude plus approfondie des alignements, incluant une inspection visuelle des lectures alignées et une analyse des différences dans les catégories de lectures multiples et chimériques. Cela me permettra de m'assurer que la variabilité introduite par les différences d'algorithmes n'impacte pas les conclusions biologiques que je tirerai des données.

# 3 Resultats

# 4 Discussion


[1] C. Wolfson et al. "Global Prevalence and Incidence of Amyotrophic Lateral Sclerosis : A Systematic Review". en. In : *Neurology* 101.6 (août 2023). ISSN : 0028-3878, 1526-632X. DOI : 10.1212/WNL.0000000000207474. URL : https://www.neurology.org/doi/10.1212/WNL.0000000000207474 (visité le 13/11/2024).

[2] Centre Constitutif SLA de Tours. *Protocole National de Diagnostic de et de Soins (PNDS) Filière FILSLAN*. Argumentaire scientifique. TOURS : Centre de référence SLA TOURS, nov. 2020, p. 6. URL : https://www.has-sante.fr/upload/docs/application/pdf/2021-12/pnds_argumentaire_sla_genetique_2020.final.pdf (visité le 28/03/2024).

[3] Orphanet. *Prévalence des maladies rares : Données bibliographiques*. Rapp. tech. 2. Nov. 2023, p. 14. URL : https://www.orpha.net/pdfs/orphacom/cahiers/docs/FR/Prevalence_des_maladies_rares_par_prevalence_decroissante_ou_cas.pdf (visité le 28/03/2024).