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
DGE	Analyse d'Expression Différentielle
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
Σ_{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 Environnement 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 par Anthony Boureux, enseignant-chercheur. L'équipe est en collaboration étroite avec des services cliniques et des plateformes hospitalières, ce qui favorise évidemment l'innovation et la résolution de questions liées au champ médical. Ainsi, 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). Les différentes missions qui m'ont étaient demandées s'inscrivent dans le champ de la transcriptomique, et tout particulièrement 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 dans la SLA , 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 [3], et la SLA en fait partie : elle touche un individu sur 20,000 en Europe [5], et sa prévalence mondiale varie de 1,57 à 11,8 pour 100,000 selon les pays, de l'Iran aux États-Unis [2]. C'est une maladie neurodégénérative causée par une atteinte du motoneurone central au niveau du cortex cérébral **(figure 1)**, et par une dégénérescence progressive des fonctions musculaires. Cette pathologie est très handicapante, tant sur le plan physique que social. 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 pertinent d'intégrer une approche transcriptomique afin d'augmenter le rendement diagnostique des formes génétiques de la maladie et de mieux en comprendre les mécanismes. Notons que les gènes responsables de la SLA sont globalement bien documentés. À ce jour, une quarantaine de gènes ont été identifiés et associés à la maladie. Dans 90% des cas, leur implication est directe dans les formes familiales. Dans les 10% restants, correspondant aux formes sporadiques, l'implication est plus indirecte, via des

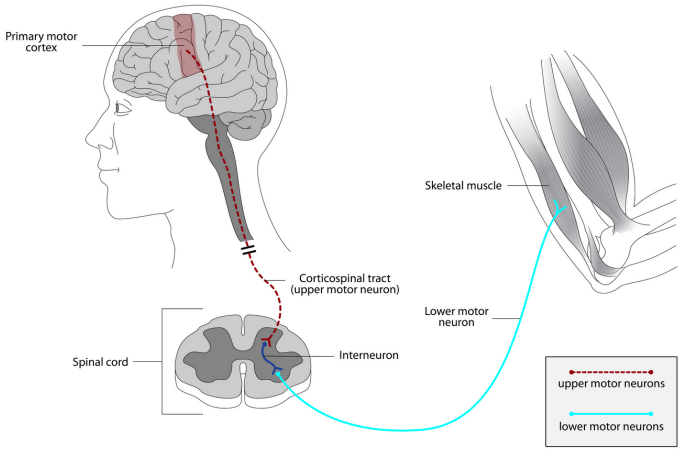

Figure 1 : *Atteinte neuronale dans la SLA(P.Wicks, 2024)*

perturbations de processus cellulaires clés tels que l'homéostasie de l'ARN, le transport axonal ou l'autophagie. Enfin, on observe une forte hétérogénéité génétique, perceptible à travers l'implication de gènes aux fonctions parfois très différentes., mais qui convergent toujours vers une dégénérescence neuronale. Les plus fréquemment impliqués sont SOD1, TARDBP, FUS et C9ORF72 [3], majoritaires dans la maladie ils constituent le socle des recherches génétiques pour tenter la mise au point de thérapies ciblées.

1.3 Modéliser le signal biologique en données exploitables

Supposons maintenant que l'on cherche à établir un profil d'expression génique pour notre quarantaine de gènes. Il faut alors s'interroger sur le support de lecture, c'est-à-dire le nombre de fois qu'une région d'ADN a été lue (ou comptée) au cours du séquençage. On s'attend logiquement à avoir une densité de distribution de ces lectures, qui peut être intuitivement interprétée comme un reflet du niveau d'expression de chaque gène étudié. En pratique, toutefois, un certain nombre de variables — d'origine biologique ou expérimentale — influencent cette expression théorique, introduisant une dispersion dans les données brutes. A travers le schéma ci-dessous je présente un nombre non exhaustif de biais qui peuvent faussées l'expression génétique :

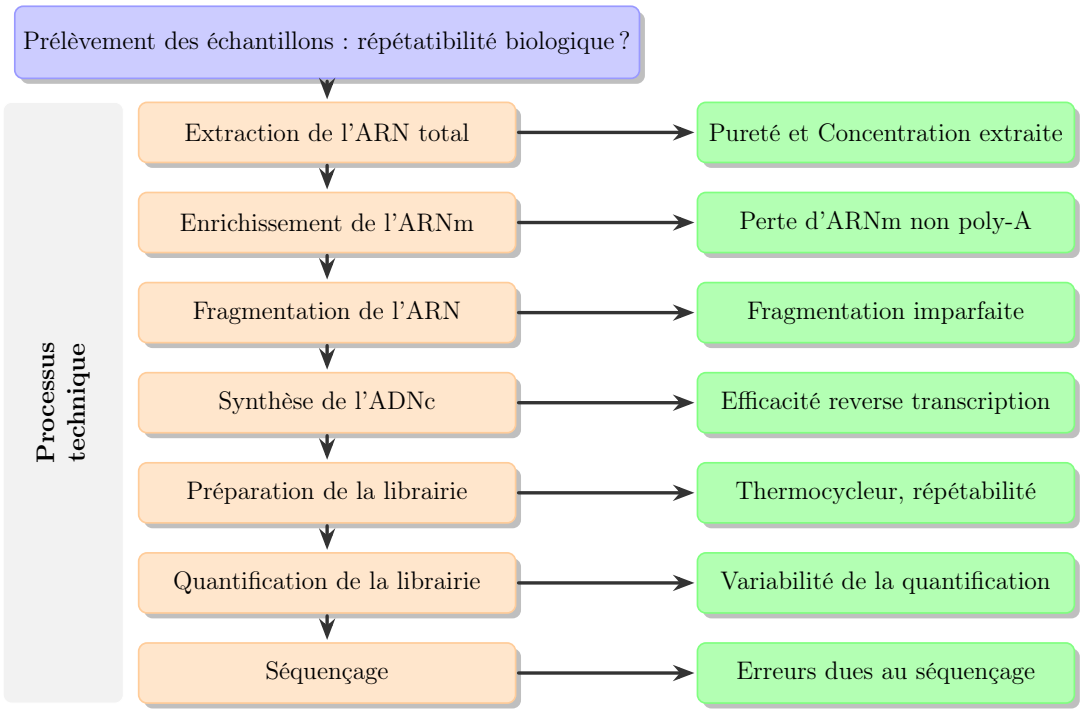

Figure 2 : *Schéma décisionnel pour normaliser les données RNASeq*

1.4 Etat de l'art

doivent elles être intégrées à l'analyse finale ? Cette question vas être pertinente dans une étude quantitative comme ici, puisqu'il s'agit, en autre, de caractériser une eventuelle haplo-insuffisance ou surexpression. Nous allons comparer les différentes stratégies.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-

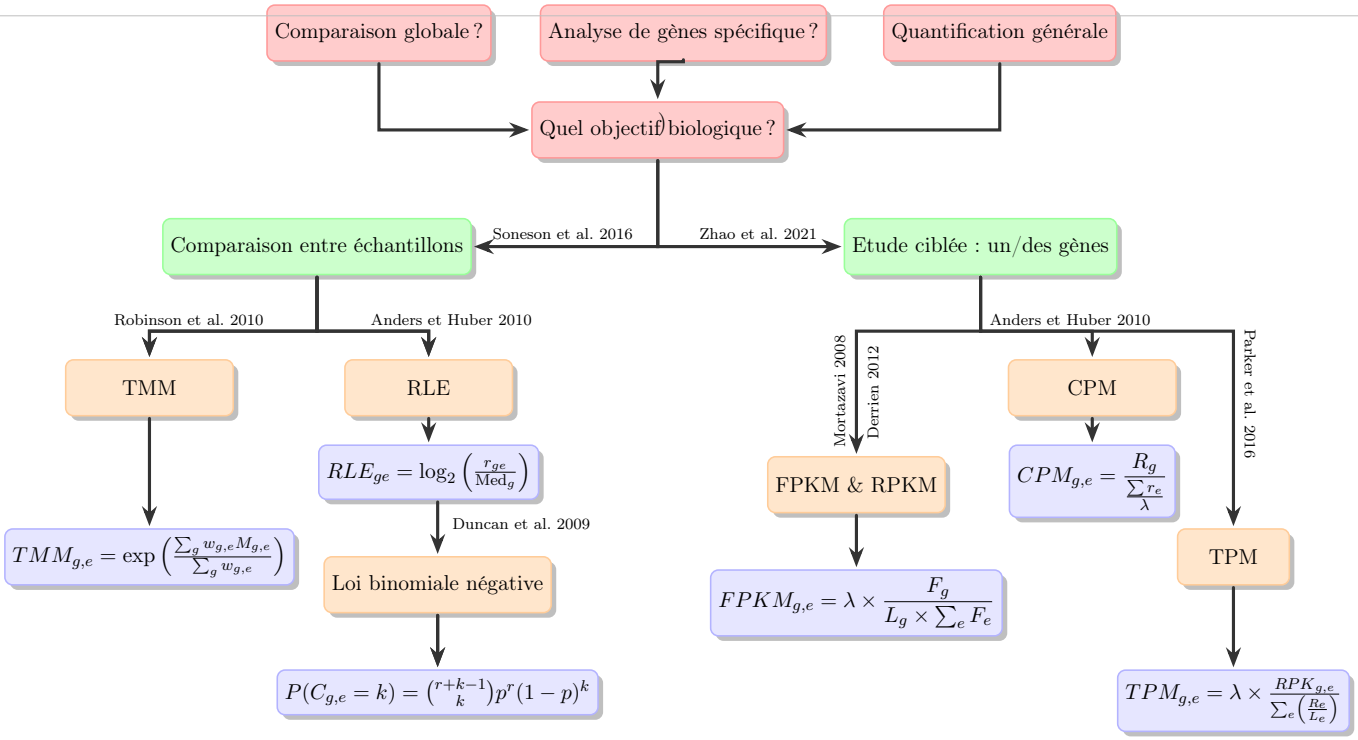

Figure 3 : *Schéma décisionnel pour normaliser les données RNASeq*

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* en 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 de synthèse ci-dessous :

Après avoir explorer les différentes méthodes de normalisation, en mettant particulièrement l'accent sur TMM, RLE et LBN, qui apportent une valeur ajoutée dans le contexte du RNA-Seq ciblé. Cette analyse repose sur plusieurs critères de performance clés en génétique, tels que les comparaisons inter-individuelles et la robustesse face aux valeurs extrêmes. L'objectif final etait d'évaluer l'approche la plus adaptée d'un panel de gènes ciblés .

1.5 Performance à faible profondeur et robustesse aux valeurs extrêmes

Un des principaux critères d'évaluation est la robustesse aux valeurs extrêmes. Lorsqu'on cherche à établir un profil d'expression pour une maladie rare, comme la SLA, il est essentiel de déterminer si les valeurs de comptage aux extrémités de la distribution doivent être intégrées à l'analyse finale. Cette question peut être pertinente dans une étude quantitative visant à caractériser une éventuelle haplo-insuffisance ou surexpression. Nous allons comparer les différentes stratégies.

Les méthodes dites *"Total Count"*, comme CPM, bien qu'utilisées pour leur simplicité, sont particulièrement sensibles aux valeurs extrêmes[4]. Elles considèrent peu d'informations et ne permettent pas une évaluation efficace de la dispersion des données. Dans un contexte où l'expression génétique varie, elles se révèlent inadéquates. De même, FPKM et TPM présentent des limitations similaires. Ces méthodes sont souvent critiquées pour leur manque de reproductibilité et leur rigueur scientifique. Par exemple, **robinson_scaling_2010** (2010) ont montré que la division par la longueur du gène amplifie l'effet des valeurs aberrantes, surtout pour les gènes courts ou faiblement exprimés[anders_differential_2010]. Les travaux de **zhao_tpm_2021** (2021) confirment également cette limite. Quant à la méthode "Upper Quartile", cela est plus nuancé, car elle élimine les données du quartile inférieur[4], donc partiellement les valeurs extrêmes, mais elle reste peu adaptée à notre contexte où certains gènes peuvent être faiblement ou fortement exprimés par rapport à ce qui est attendu biologiquement.

A l'inverse, TMM, qui prend en compte la composition de l'ARN au moment de la normalisation, offre une meilleure résistance aux biais causés par les valeurs extrêmes[robinson_scaling_2010]. C'est aussi le cas pour RLE, grâce à l'utilisation de la médiane comme paramètre de position, qui donne une tendance centrale. Enfin, LBN repose sur l'hypothèse que les gènes ne sont pas différentiellement exprimés entre les conditions, ce qui renforce sa pertinence pour analyser les extrémités de la distribution. Ces trois approches partagent une philosophie commune : estimer le facteur de normalisation à partir d'un ensemble de gènes supposés stables.

1.6 Compromis entre facilité d'interprétation et la pertinence biologique

Après avoir fait un panorama des différentes stratégies mathématiques pour traiter les comptages bruts, il est évident que la diversité des méthodes peut paraître déroutante. Finalement, ce qui va être essentiel pour le biologiste, c'est de lui proposer une approche qui repose sur un compromis entre facilité d'interprétation et pertinence médicale. Le biologiste doit être en mesure de comprendre les corrections pour les relier aux décisions diagnostiques. Les méthodes conceptuellement simples comme CPM et UQ sont éliminées des options disponibles pour le périmètre de notre étude, car nous avons montré que, avec des échantillons présentant des variations biologiques importantes, la normalisation conduira à des conclusions erronées. Nous avons vu que les métriques comme TPM et FPKM sont des approches moins avancées que TMM, RLE et LBN, qui sont également plus simples à comprendre et à implémenter que les autres, mais elles manquent de correction statistique pour les comparaisons inter-échantillons.

En conclusion, la piste à creuser dans le périmètre de notre étude pencherait vers TMM ou RLE, et venir modéliser avec la LBN la surdispersion. Dans la littérature, il est régulièrement fait référence que la solution adéquate réside dans l'application d'une double correction[robinson_scaling_2010]. Comme nous l'avons souligné dans la méthodologie de la LBN, une fois que les biais techniques sont absorbés (par la correction RLE ou TMM), la méthode LBN peut être appliquée pour l'analyse de l'expression différentielle. Cette stratégie répondra à la fois à la problématique de la surdispersion et

permettra d'obtenir une estimation précise de la variabilité inter-individuelle. Enfin, il est intéressant de mentionner que la taille de l'échantillon statistique (nombre de librairies) est directement corrélée à la puissance de la correction LBN. Comme souvent en statistique : plus les données seront consistantes plus la correction le sera également. Une piste de travail pour s'adapter aux contraintes de temps et de coût au laboratoire et pouvoir appliquer de manière crédible ces corrections serait donc de compiler (TMM ou RLE avec LBN) les données des diagnostics des librairies successives, semaine après semaine, pour augmenter la taille des données. Bien évidemment, cela doit se faire sous contrôle d'analyses statistiques pour valider une telle approche. C'est une option à envisager pour fournir un modèle probabiliste puissant pour l'analyse d'expression malgrès les contraintes organisationnelles (8 échantillons par expérience).

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

1.7 Le problème

Avant tout, rappelons que nous travaillons sur une maladie constitutionnelle. Par conséquent, il y a une très faible probabilité d'analyser des échantillons distincts pour un même patient à un intervalle de temps raisonnable, ce qui a peu de sens d'un point de vue de la génétique constitutionnelle, sauf dans des cas particuliers comme les études familiales. La notion de variabilité intra-individuelle est une qualité à étudier sur une matrice biologique identique. De manière générale, au laboratoire de génétique, il s'agit du sang, dans un milieu de transport adapté (tube STRECK pour l'ARN). Si ces conditions ne sont pas respectées, cela pourrait plutôt refléter une hétérogénéité biologique inhérente à une régulation de l'expression liée à la nature même du tissu, aux conditions de transport, à la variabilité de l'environnement analytique, avant même de songer à absorber les différents biais. Il est donc fondamental d'avoir une planification expérimentale aussi standardisée que possible et d'un référencement complet et précis des biais biologiques connus et d'en évaluer leur impact. Évaluer les biais biologiques et techniques est délicat. Comme nous l'avons vu plus haut, ils peuvent être confondus (préparation de la librairie, effets de lots, longueur des gènes, etc.). La littérature aborde cet aspect sous des angles parfois différents, et les conclusions sont attribuées à des facteurs techniques et/ou biologiques, car il est difficile de faire la distinction. Chaque expérience étant unique, chaque prélèvement l'est aussi, notamment par ses délais et ses modalités d'acheminement. Pour capturer au mieux ces sources de variabilité, et à travers les différentes études comparatives, j'ai constaté que les méthodes les méthodes conventionnelles présentées en **figure 1** ne sont pas ou peu adapté au RNASeq ciblé pour faire de DGE. Ceci s'explique par leurs modalités de calcul, qui intègrent la dispersion et la comparaison inter-échantillons à l'échelle du transcriptome [4]. TMM, en utilisant une moyenne tronquée des ratios de lectures pour calculer les facteurs de normalisation, va corriger efficacement les biais liés à la taille de la librairie (profondeur de séquençage notamment) et à la

composition de l'ARN (régions riches en GC)[1].

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

5 Bibliographie compilée

[1] Zachary B. Abrams et al. "A protocol to evaluate RNA sequencing normalization methods". In : *BMC Bioinformatics* 20.24 (déc. 2019), p. 679. ISSN : 1471-2105. DOI : 10.1186/s12859-019-3247-x. URL : https://doi.org/10.1186/s12859-019-3247-x (visité le 16/08/2024).

[2] 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).

[3] 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).

[4] M.-A. Dillies et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". en. In : *Briefings in Bioinformatics* 14.6 (nov. 2013), p. 671-683. ISSN : 1467-5463, 1477-4054. DOI : 10.1093/bib/bbs046. URL : https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs046 (visité le 16/08/2024).

[5] 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).