liser les liens

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 collabore avec des services cliniques et des plateformes hospitalières, ce qui favorise la résolution de problématiques liées au champ médical. Ainsi, j'ai eu l'occasion de contribuer à un projet de recherche translationnelle que Bio2M mène en partenariat 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 été confiées s'inscrivent dans le champ de la transcriptomique, et plus particulièrement dans le cadre de l'analyse d'expression génique appliquée à la SLA, l'idée étant d'initier une stratégie permettant de détecter une expression différentielle à l'échelle de certains gènes d'intérêt dans la SLA, avec des contraintes à la fois techniques et biologiques, pour *in fine* tenter d'identifier la causalité génétique chez les patients.

1.2 Contexte biologique

Une maladie rare se définit par une prévalence[1] de 0,05 % dans la population générale. Quatre-vingts pour cent de ces maladies sont d'origine génétique [2], et la SLA en fait partie : elle touche un individu sur 20 000 en Europe [3], et sa prevalence 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 **(figure 1)**, conduisant à 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 cliniques. 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, on observe

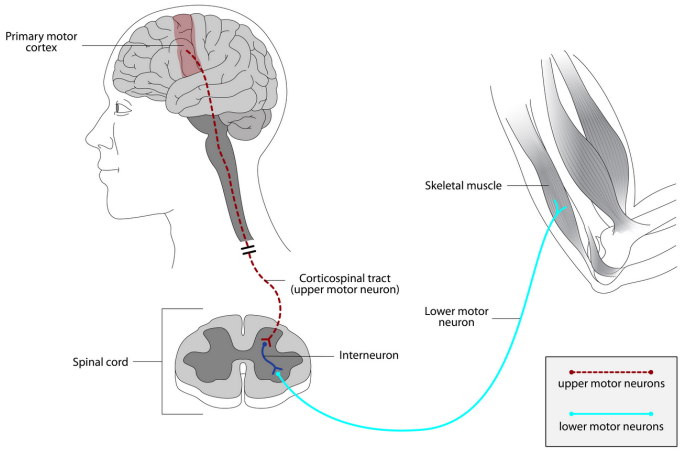

Figure 1 :
Atteinte neuronale dans la SLA(P.Wicks, 2024)

des formes dites sporadiques, où la causalité génétique est cette fois indirecte, via des perturbations de processus cellulaires clés tels que l'homéostasie[2] de l'ARN, le Transport axonal[3] ou l'autophagie[4]. On observe également une forte hétérogénéité génétique dans cette maladie, perceptible à travers l'implication de gènes aux fonctions parfois très différentes, mais qui convergent toujours vers une dégénérescence neuronale. Les gènes les plus fréquemment impliqués sont *SOD1*, *TARDBP*, *FUS* et *C9ORF72* [2]. Majoritaires dans la maladie, ils constituent le socle des recherches génétiques pour tenter la mise au point de thérapies ciblées et faire progresser la compréhension ainsi que le diagnostic de cette pathologie.

1.3 L'apport du RNA-Seq ciblé dans le périmètre de notre analyse

D'un point de vue biologique, *stricto sensu*, on sait que l'étape de transcription est fondatrice de la diversité protéomique ; elle constitue, de ce fait, une source majeure d'anomalies génétiques. À l'issue de ce mécanisme, un gène peut exprimer plusieurs isoformes, dont certaines peuvent avoir un impact pathologique. Tout l'objet de ce travail est de proposer une méthode pertinente pour détecter des différences d'expression de tel ou tel gènes, susceptibles d'aider le biologiste à établir un lien avec la maladie, notamment à travers une haploinsuffisance[5] ou, au contraire, une surexpression.

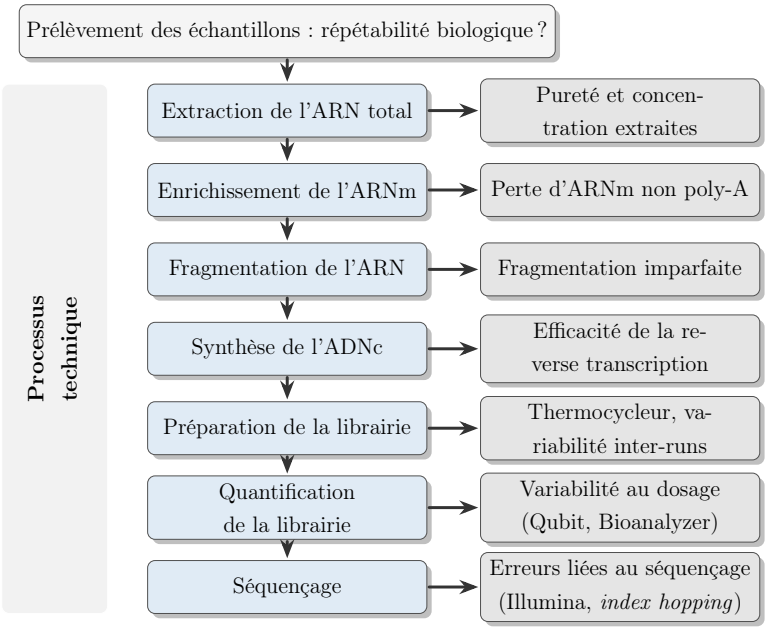

Figure 2 :
Processus expérimental du RNA-Seq et biais potentiels

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 de chacun de ces gènes, 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 observer une distribution des lectures, interprétable comme le reflet du niveau d'expression de chaque gène étudié. Dans les faits, si l'on prend les données brutes, on s'en éloigne considérablement : un certain nombre de variables — biologiques (rythme circadien ?) ou expérimentales (moment du prélèvement ?) — influencent cette distribution théorique, introduisant une dispersion non négligeable dans le jeu de données. De plus, l'analyse bioinformatique peut, elle aussi, fausser le profil d'expression, notamment lors des choix effectués à l'étape d'alignement — nous le verrons. À travers la **figure 2**, j'illustre quelques biais, non exhaustifs, susceptibles d'affecter l'expression génique mesurée.On comprend à travers cette illustration, qu'à chaque étape technique du protocole RNA-Seq, on introduit de la variabilité, parfois de manière systématique entre les expériences, parfois de manière aléatoire.

Toute la stratégie de quantification de l'expression génique consiste à limiter ces biais, ou le cas échéant, à les intégrer dans l'analyse — afin d'obtenir des résultats fidèles à la réalité biologique du patient, mais aussi reproductibles dans des conditions de routine diagnostique. L'importanct est d'effectuer un certain nombre de vérifications préalables sur les données, notamment via des statistiques descriptives, ainsi que des corrections appropriées, comme la normalisation. J'ai eu l'occasion d'effectuer un travail préliminaire qui consiste à établir un état des lieux des méthodes conventionnelles de normalisation, rappelées dans la **figure 3**. Qui m'a permis d' étayer ma compréhension de cette étape d'un pipeline bioinformatique mais surtout d'en tirer une conclusion, elle ne sont pas les plus adaptés pour travailler sur des gènes ciblés, notamment par le manque de consistance des données de comptage.(**a**jout ref travail bibliographique)

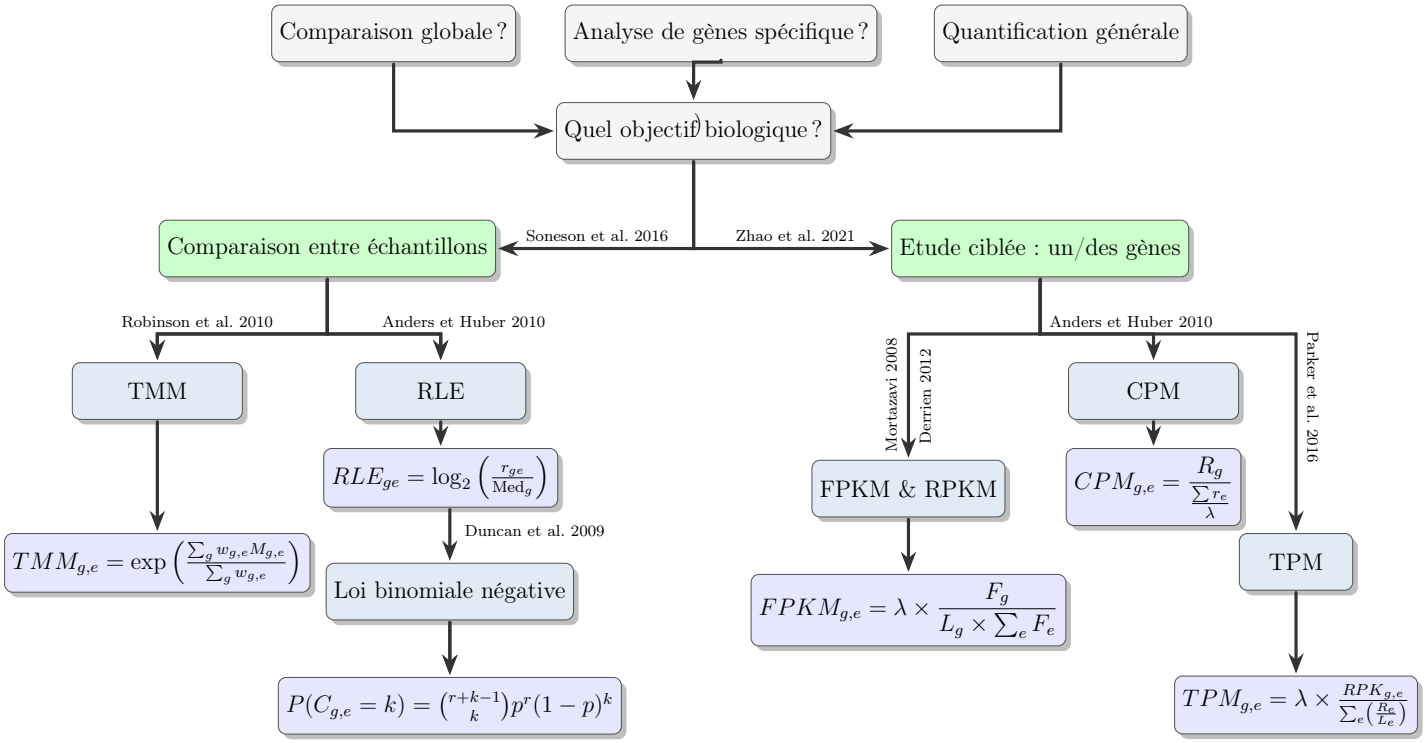

Figure 3 : Schéma décisionnel pour normaliser les données RNASeq

Le périmètre de mon travail étant définit, il reste une question que l'on peut légitimement se pose à laquelle je n'ai pas répondu dans cette section : pourquoi avoir une approche de RNA-Seq ciblé plutôt qu'une analyse transcriptomique globale ? A première vue, en se concentrant sur un panel restreint de gènes d'intérets (ceux impliqués dans la SLA) j'améliore la sensibilité de détection et surtout je réduit les coûts expérimentaux et de traitements bioinformatiques. C'est d'autant plus adapté au diagnostic, puisqu'on aura des résultats plus rapides et immédiatement exploitables. Du reste, je limite l'analyse aux seuls gènes ciblés, excluant la détection d'événements potentiellements pertinents, comme des isoformes rares ou des altérations affectant d'autres régions du transcriptome. C'est donc un compromis qui est fait, car le séquençage transcriptomique global fournit une vision d'ensemble, mais au prix d'une complexité analytique accrue, d'un coût plus élevé et d'un bruit de fond plus important. Il n'ya pas ici une stratégie meilleure que l'autre tout dépend du contexte

médicale et de la pertinence clinique et contraintes techniques. Dans le contexte des maladies rares comme la SLA, où les gènes en cause sont bien caractérisés, le RNA-Seq ciblé constitue une option pragmatique et efficiente.

Partant de ce constat, mon travail est d'évaluer la qualité et la reproductibilité des données issues du protocole de RNA-Seq ciblé appliqués aux échantillons. J'ai tout d'abord analysé les métriques d'alignement générées par deux outils largement utilisés dans ce domaine, STAR et CRAC, afin de déterminer si la variabilité expérimentale pouvait s'expliquer en tout ou partie par des biais liés à l'étape d'alignement. Je propose ensuite une méthode pour quantifier l'expression fondée sur une approche par par k-mers, indépendamment de cette étape d'alignement. Cette démarche vise à contourner certaines limites inhérentes aux méthodes classiques présentées en **figure 3** dans le cadre du RNA-Seq ciblé tout en conservant une résolution suffisante pour la détection de l'haploinsuffisance.

Ainsi, les sections suivantes décrivent la méthodologie adoptée pour contrôler la qualité des données, comparer plusieurs stratégies d'alignement, puis amorcer une réflexion sur les approches de quantification adaptées à l'étude que j'ai en charge de mener.

2 Matériels & Méthodes

Le « *mapping* » dans un pipeline bioinformatique est une étape charnière, et dans notre cas d'étude, on peut raisonnablement penser que cette étape est susceptible de biaiser le signal biologique réel. Pour évaluer la phase d'alignement, j'ai sélectionné deux outils que je vous présente ici : `STAR` et `CRAC`.

2.1 Présentation conceptuelle de STAR et CRAC

Ainsi, j'ai cherché à évaluer l'impact de cette étape en générant les fichiers `BAM` pour l'intégralité des patients concernés à l'aide de deux outils distincts : `STAR` et `CRAC`, fondés sur des stratégies algorithmiques différentes. La majeure partie des commandes, ainsi que leurs paramètres, sont consultables dans un `Makefile` disponible sur le `Git`. Comme évoqué dans la partie introductive, nous travaillons ici sur un jeu de données limité à 72 patients. La génération des fichiers `BAM` s'effectue à partir de deux fichiers `FASTQ`, obtenus via la technologie Illumina™ , puisque les données ont été produites en mode séquençage pairé[6].

2.2 Génération des fichiers d'alignement au format BAM

Dans l'absolu, ce qui nous intéresse, c'est le fichier consignant les métriques d'alignement générées durant l'exécution de l'outil, et non directement le fichier `BAM` lui-même. (Il s'agit d'un fichier de « log » pour `STAR` et d'un fichier « summary » pour `CRAC`.) J'ai mis en place un pipeline reproductible à cette fin. Pour chaque outil, j'ai conçu quelques cibles en `Bash` permettant de lancer les alignements en série à partir des paires de fichiers `FASTQ`.

```
1 SAMPLES = $(shell \
2    for R1 in $(FASTQ_DIR)/*_1.fastq.gz; do \
```

```
      R2= echo $$R1 | sed 's/_1.fastq.gz/_2.fastq.gz/'; \
      if [ -f $$R2 ]; then \
         basename $$R1 _1.fastq.gz; \
      fi; \
   done)
```

Code 1 : Construction de la variable `SAMPLES` pour traiter séquentiellement les `fastq`

Dans les différents code que je vous présente ci-dessous, je cherche à stocker les parties sujettes à dans des variables, pour rendre le plus générique l'éxécution, considérer le répertoire de stockage des dépendances (l'index par exemple) et des fichiers de sortie sont dépendantes de l'organisation de l'utilisateur sur sa machine d'éxécution :

```
# Détection des échantillons par présence des fichiers fastq _1 et _2
BAMS_STAR = $(addprefix $(OUTBAM_STAR)/, $(addsuffix .bam, $(SAMPLES)))

Star_Paire: $(BAMS_STAR)
$(OUTBAM_STAR)/%.bam:
  @mkdir -p $(OUTBAM_STAR) $(OUTLOG_STAR) $(TMPDIR)/$*;
  @echo ">>Lancement de l'alignement de l'échantillon $* avec STAR";
  STAR --runThreadN $(THREADS)
    --genomeDir $(REF_STAR)
     --readFilesIn $(FASTQ_DIR)/$*_1.fastq.gz $(FASTQ_DIR)/$*_2.fastq.gz
     --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate
     --outFileNamePrefix $(TMPDIR)/$*/;
  @mv $(TMPDIR)/$*/Aligned.sortedByCoord.out.bam $(OUTBAM_STAR)/$*.bam;
  @mv $(TMPDIR)/$*/Log.final.out $(OUTLOG_STAR)/$*.Log.final.out;
```

Code 2 : Cible `Star_Paire` pour générer les `BAM` avec STAR

Concernant l'utilisation de CRAC, une étape supplémentaire est nécessaire car nativement l'aligneur génère une sortie SAM, j'ai rédigé une cible supplémentaire pour convertir les `SAM` en `BAM` de l'utilitaire `Samtools`, on peut ajouter également pour vérifier la vacuité de la sortie, pour ne relancer le traitement que pour les `fastq` dont le `SAM` n'a pas été généré.

```
CRAC_SAMS = $(addprefix output/crac/bam/, $(addsuffix .sam, $(SAMPLES)))
Crac_Paire: $(CRAC_SAMS)

# Règle pour chaque .sam
output/crac/bam/%.sam: $(FASTQ_DIR)/%_1.fastq.gz $(FASTQ_DIR)/%_2.fastq.gz
  @echo "Vérification de l'échantillon : $*"
  @if [ ! -f output/crac/bam/$*.bam ]; then
     echo "Lancement de l'alignement de l'échantillon : $* avec Crac";
     mkdir -p output/crac/summary output/crac/log output/crac/bam; \
     gunzip -c $(FASTQ_DIR)/$*_1.fastq.gz > $(FASTQ_DIR)/$*_1.fastq;\
     gunzip -c $(FASTQ_DIR)/$*_2.fastq.gz > $(FASTQ_DIR)/$*_2.fastq; \
     crac --nb-tags-info-stored 10000 --bam --stranded -i $(REF_CRAC) -k
        $(KMER_CRAC)
```

```
      --summary output/crac/summary/$*.summary
       --nb-threads $(THREADS) -r $(FASTQ_DIR)/$*_1.fastq
          $(FASTQ_DIR)/$*_2.fastq -o output/crac/bam/$*.sam
       2> output/crac/log/$*_crac.log;
else
   echo "Fichier déjà aligné : $@";
fi
```

Code 3 : Cible `Crac_Paire` pour générer les `BAM` avec CRAC

Structure de traitement par run (batch de 8 échantillons).

Mise en place d'un pipeline reproductible.

2.2.1 Analyse de la variabilité expérimental et levée d'ambiguité sur l'alignement

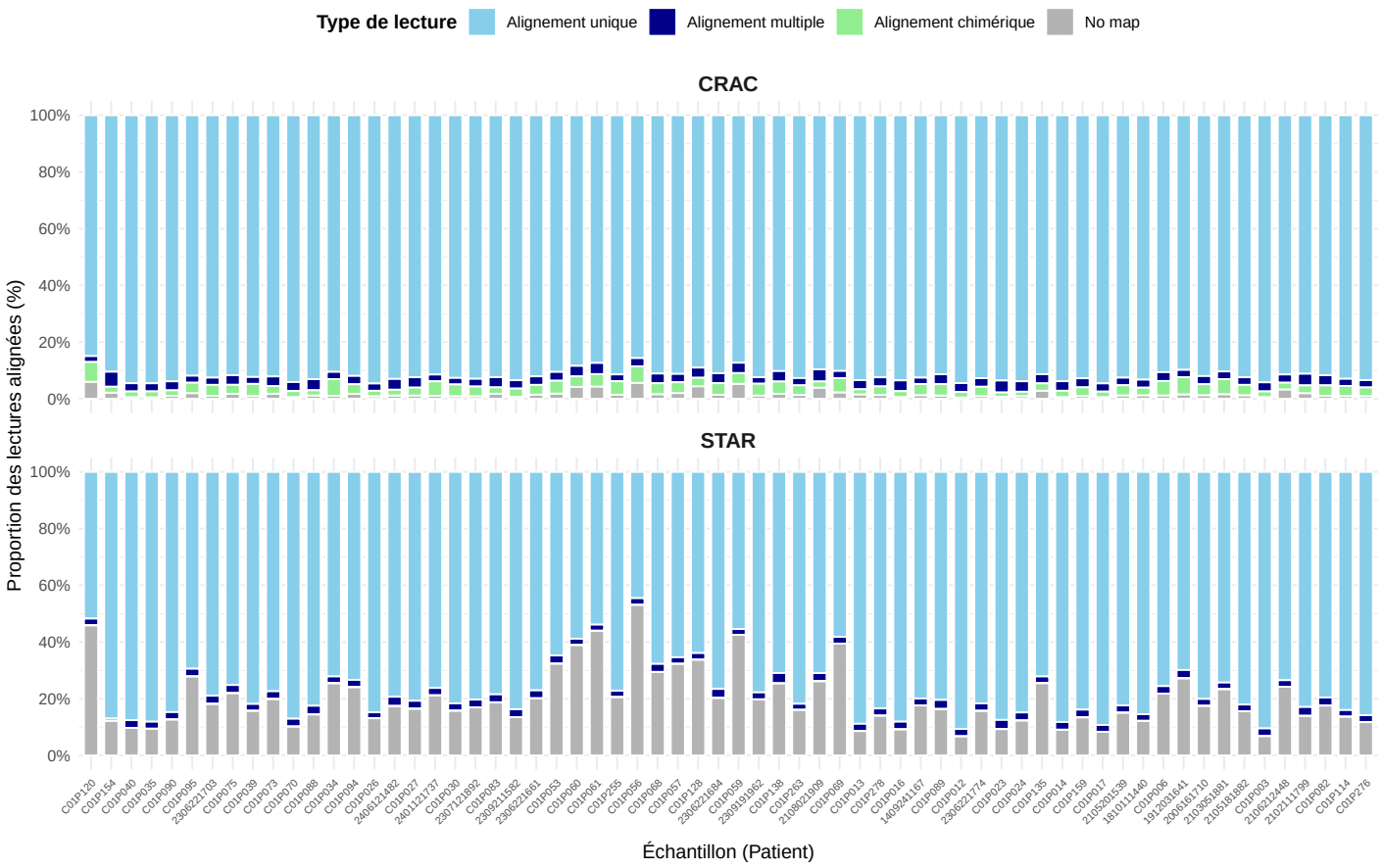

Figure 4 : Comparaison des profils d'alignement entre STAR et CRAC

.... explications des deux approches algorithmiques a rajouter.

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.

Afin de comparer les performances respectives des deux outils d'alignement (STAR et CRAC) en termes de profondeur de lecture totale par patient, nous avons tout d'abord extrait les totaux de lectures alignées pour chacun d'eux. Une étape préalable essentielle consiste à évaluer la normalité des distributions de ces profondeurs d'alignement à l'aide du test de Shapiro-Wilk. Dans les deux cas (STAR et CRAC), les p-values obtenues sont inférieures à 0,05, ce qui indique un écart significatif à la normalité. Dès lors, l'hypothèse d'une distribution gaussienne est rejetée, ce qui justifie le recours à un test non paramétrique.

Nous avons donc utilisé le test de Wilcoxon pour données appariées afin de comparer les profondeurs d'alignement entre STAR et CRAC pour un même ensemble de patients. Ce test est particulièrement adapté dans ce contexte, car il ne repose pas sur l'hypothèse de normalité des données, et il tient compte de la structure appariée des échantillons (comparaison patient par patient).

Le test de Wilcoxon a retourné une p-value extrêmement faible ($p = 3,63710^{-13}$), ce qui indique une différence significative entre les profondeurs d'alignement fournies par STAR et CRAC. Autrement dit, l'outil CRAC génère de manière systématique un nombre de lectures alignées significativement plus élevé que STAR pour les mêmes échantillons, ce qui suggère une sensibilité accrue ou une stratégie d'alignement plus permissive.

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.

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.

2.2.2 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.

2.2.3 Qualité et spécificité de l'alignement

3 Resultats

4 Discussion

Glossaire

autophagie L'autophagie est un processus cellulaire de recyclage qui dégrade et élimine les composants cellulaires endommagés, contribuant à l'homéostasie et à la protection contre le stress cellulaire. p. 5

haploinsuffisance Incapacité d'une seule copie fonctionnelle d'un gène à produire une quantité suffisante de produit génique (ARN ou protéine) pour assurer une fonction biologique normale, entraînant ainsi un phénotype pathologique. p. 5

homéostasie Ensemble des mécanismes qui régulent la production, la maturation, le transport et la dégradation pour maintenir un équilibre fonctionnel dans la cellule. p. 5

prévalence Proportion d'individus dans une population donnée présentant une caractéristique (généralement une maladie) à un instant donné ou sur une période donnée. p. 4

séquençage pairé Méthode de séquençage où les deux extrémités d'un fragment d'ADN sont séquencées indépendamment, permettant d'obtenir deux lectures (lectures appariées) qui facilitent l'alignement et la détection des variants, notamment dans les régions complexes ou répétées. p. 7

Transport axonal Le transport axonal permet le déplacement des organites, protéines et ARN le long de l'axone en assurant la communication et la survie neuronale sur de longue distance. p. 5

5 Bibliographie compilée

[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).