Université de Montréal

Caractérisation de l'étiologie génétique de patients atteints de différentes maladies neuromusculaires par l'intégration de données omiques

Par

Valérie Triassi

Département de biochimie et médecine moléculaire, Faculté de médecine

Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maîtrise (M.Sc.) en bio-informatique

Décembre 2022

Université de Montréal

Faculté de médecine

Ce mémoire intitulé

Caractérisation de l'étiologie génétique de patients atteints de différentes maladies neuromusculaires par l'intégration de données omiques

Présenté par

Valérie Triassi

A été évalué par un jury composé des personnes suivantes

Julie Hussin

Président-rapporteur

Martine Tétreault

Directeur de recherche

Despoina Manousaki

Membre du jury

Résumé

Les progrès des technologies de séquençage ont joué un rôle important dans le diagnostic moléculaire des maladies rares, telles que les myopathies et les dystrophies musculaires. Cependant, plusieurs patients atteints de maladies neuromusculaires restent sans diagnostic. Ceci est dû à la grande hétérogénéité clinique et génétique ainsi qu'au caractère hautement polymorphique des gènes associés à ces troubles. L'interprétation des données génétiques est un grand défi et les tests génétiques aboutissent souvent à l'identification de variants de signification inconnue (VUSs). Plusieurs de ces variants peuvent perturber l'épissage normal de l'ARN ou affecter l'expression des gènes. À cet égard, nous proposons une approche bio-informatique.

Notre objectif est de mettre en place un pipeline identifiant et caractérisant des variants d'intérêt dans un contexte pathologique. Afin de déterminer si les variants ont un impact fonctionnel, notre pipeline se concentre sur l'épissage alternatif ainsi que sur l'intégration des données génomiques et transcriptomiques. Nous émettons l'hypothèse qu'une partie des patients sans diagnostic pour leur maladie neuromusculaire s'explique par des variants introniques jouant un rôle régulateur ou affectant l'épissage et l'abondance de l'ARNm. Cette approche multi-omique permet de déterminer si les variants ont un impact fonctionnel.

Pour ce faire, nous avons réalisé un séquençage de l'ARN et de l'ADN à partir de biopsies musculaires de quatre patients. Les données ont été alignées et annotées avec différents scores de pathogénicité. Les événements d'épissage sont analysés par SpliceAI et rMATS. L'analyse des gènes différentiellement exprimés a été réalisée par l'outil LPEseq. Les CNVs et les expansions de répétitions ont été analysés avec CNVkit et ExpansionHunter. Plusieurs scripts maison filtrent et intègrent les données ARN et ADN. Pour l'instant, un total de huit variants pathogéniques sont proposés pour nos patients, mais des investigations supplémentaires sont nécessaires.

Les variants recherchés sont rares et nécessitent donc un pipeline cohérent et efficace. Ce projet apportera un bénéfice significatif pour les patients en leur permettant d'obtenir un diagnostic et ainsi d'avoir accès à un meilleur suivi clinique.

Mots-clés : Neuromusculaire, Traits complexes, Variants rares, Bio-informatique, Épissage alternatif, RNA-Seq, Exome-Seq, WGS, Maladies Mendélienne.

Abstract

Advances in sequencing technologies have played an important role in the molecular diagnosis of rare diseases, such as myopathies and muscular dystrophies. However, several patients with these neuromuscular diseases remain undiagnosed. This is due to the great clinical and genetic heterogeneity as well as the highly polymorphic nature of the genes associated with myopathies and muscular dystrophies. The interpretation of genetic data is a great challenge and genetic testing often results in the identification of variants of uncertain significances (VUSs). Many of these variants can disrupt normal RNA splicing or affect gene expression. In this regard, we propose a bioinformatics approach.

Our aim is to put in place a pipeline identifying and characterizing variants of interest in a pathological context. To determine if the variants have a functional impact, our pipeline focuses on alternative splicing as well as the integration of genomic and transcriptomic data. We hypothesize that a portion of patients without a diagnosis for their neuromuscular disorder is explained by intronic variants having a regulatory role or affecting mRNA splicing and abundance. This multi-omics approach will make it possible to determine whether the variants have a functional impact.

To do so, we performed RNA and DNA sequencing using muscle biopsies from four patients. Data was aligned and annotated with different pathogenicity scores. Splicing events are analyzed by SpliceAI and rMATS. The analysis of the differentially expressed genes was carried out by the LPEseq tool. CNVs and repeat expansion were analyzed with CNVkit and ExpansionHunter. Several in-house scripts filter and integrate RNA and DNA data. For now, a total of eight pathogenic variants are proposed for our patients, but further investigations are needed.

The variants sought are rare and therefore require a coherent and efficient pipeline to facilitate their characterization. This project will have a significant benefit for patients by allowing them to obtain a diagnosis and thus have access to better clinical follow-up.

Keywords : Neuromuscular, Complex traits, Rare variants, Bioinformatics, Alternative splicing, RNA-Seq, Exome-Seq, WGS, Mendelian diseases.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ADN : Acide désoxyribonucléique

AF : Fréquence allélique

ARN : Acide ribonucléique

ARNm : ARN messager

ARNmt : ARN mitochondrial

BWA : Alignement Burrows-Wheeler

CNV : Variabilité du nombre de copies

ES : Séquençage de l'exome (Exome-Seq)

HISAT2 : Indexation hiérarchique pour l'alignement épissé des transcriptions version 2.1.0

IGV : Visualiseur de génomique intégrative

Indels : Insertions et délétions

IncLevelDiff : Niveau d'inclusion différent

Log2 : Logarithme en base 2

L2FC : Logarithme en base 2 Fold Change

LRS : Séquençage de longues lectures

NGS : Séquençage de nouvelle génération

MNV : Variant multinucléotidique

QC : Contrôle de qualité

RNAseq : Séquençage de l'ARN

SNV : Variant mononucléotide

SR : Protéines riches en sérine/arginine

STR : Courte répétition en tandem

VCF : Format d'appel de variant

VUS: Variant de signification inconnu

WGS : Séquençage du génome complet

Remerciements

J’aimerais remercier infiniment ma directrice de projet et mon mentor, Dre Martine Tétreault. J'ai débuté dans votre laboratoire lors d’une pandémie, vous avez vraiment réussi à m’encadrer de manière exceptionnelle. Merci de m’avoir motivée durant mon stage à poursuivre mes études aux cycles supérieurs, ce fut un réel plaisir d’approfondir mes connaissances. Merci pour votre patience et confiance envers moi, en tant qu’étudiant il arrive parfois des moments plus difficiles et vous étiez toujours disponible pour moi afin de me remettre sur le bon chemin. Merci d’être comme vous êtes, avec vos encouragements, votre gentillesse et pour le partage de vos connaissances. Merci également à Dr Martin Smith pour son parrainage et ses bons conseils.

J’aimerais maintenant remercier mes parents et ma sœur Olivia pour leur support et la stabilité qu'ils m’ont fournis lors de mes études. Je n’oublierai jamais cette expérience avec ma mère qui était devenue ma partenaire de télétravail pendant la pandémie. Merci à mon conjoint Mathieu, je suis très chanceuse d'avoir eu à mes côtés une personne qui a su m'aider à gérer mes inquiétudes et anxiétés, me motiver et m'inspirer. Merci.

Merci également à mes deux chères amies Kristina Atanasova et Marjorie Labrecque. Vous avez été avec moi depuis le début de mon baccalauréat, nous avons vécu les mêmes expériences, les mêmes examens difficiles, les mêmes projets complexes. Je n'oublierai jamais nos fous rires et notre complicité durant notre parcours universitaire.

Merci aux membres du laboratoire de Dre Tétreault au CRCHUM, malgré la pandémie nous avons formé de belles relations créant une atmosphère professionnelle, chaleureuse et agréable.

Merci au Département de Biochimie, programme de Bio-informatique, et la Faculté des Études Supérieures et Postdoctorales. Finalement, merci aux membres de ce jury qui ont accepté de donner de leur temps pour évaluer ce travail.

Chapitre 1 – Introduction

Les maladies musculaires touchent un grand nombre d'individus. Certains peuvent être nés avec la maladie et d'autres l'acquièrent au courant de leur vie (1). Malgré plusieurs investigations par des cliniciens et généticiens certains patients restent sans diagnostic. Bien que leur phénotype ressemble à celui de patients déjà diagnostiqués, du point de vue moléculaire ils présentent une forme plus rare de la maladie. Afin d'élucider le mystère derrière leur pathologie, une approche plus personnalisée est requise.

L'étiologie génétique de la condition de ces patients atteints de maladies neuromusculaires rares reste inconnue malgré de nombreuses investigations. Pour les dystrophies musculaires et les myopathies, il est très fréquent que le phénotype associé à un gène spécifique soit hétérogène, ce qui rend le diagnostic moléculaire plus difficile. Les différentes mutations trouvées dans un même gène provoquent une grande diversité de symptômes pathologiques entre les patients même au sein d'une même famille. Dans le cadre de ce projet on se concentre sur les maladies neuromusculaires rares, ces mutations sont plus difficiles à identifier avec les approches bio-informatiques standards. Un pipeline personnalisé pour ce type de données est important et nécessaire pour définir l'étiologie de ces patients. Avec un diagnostic, les patients peuvent bénéficier d'un conseil génétique et avoir accès à une prise en charge clinique plus personnalisée. De plus, la découverte d'un nouveau traitement ou d'une nouvelle approche thérapeutique repose sur la connaissance des variations génétiques, ainsi de trouver la cause moléculaire ouvrira des opportunités pour ces patients.

Le présent mémoire propose une approche bio-informatique intégrant le transcriptome et le génome de quatre patients suspectés d'être atteints d'une maladie neuromusculaire rare. Ce premier chapitre est divisé en deux parties principales. La première partie introduit le contexte clinique puis la seconde introduit le contexte génomique et bio-informatique de ce projet. Les outils bio-informatiques seront aussi introduits à travers la deuxième section de ce chapitre.

Les patients de ce projet ne sont pas apparentés, n'ont pas le même âge ou les mêmes symptômes. Ils ont chacun un phénotype différent donc on ne recherche pas un évènement génétique commun à ces patients. Des analyses indépendantes utilisant l'approche proposée sont effectuées afin de trouver un diagnostic pour chacun. Pour cela, il faut comprendre les différentes formes des maladies neuromusculaires.

1.1 Introduction du contexte clinique: Les maladies neuromusculaires rares

Les maladies neuromusculaires affectent généralement le système nerveux périphérique comprenant les muscles, les jonctions neuromusculaires, les motoneurones et les neurones sensoriels (Figure 1) (2). Ce sont tous des composants essentiels du système nerveux qui permettent la communication entre les nerfs du cerveau et la moelle épinière avec tout le reste des membres et des organes du corps. Lorsqu'un marqueur génétique provoque un dysfonctionnement de ses voies de communication, le patient atteint peut rencontrer des difficultés à effectuer des tâches physiques. La perte de communication causée par le variant peut entraîner une détérioration musculaire (atrophie).

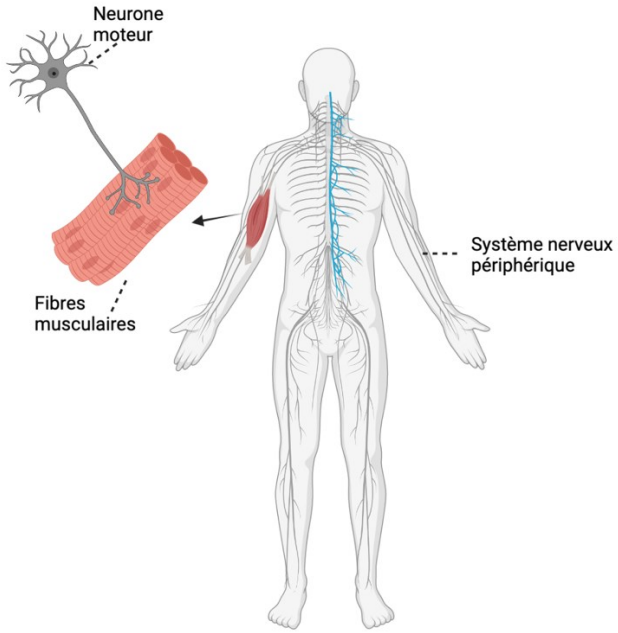

Figure 1. – Les principaux composants essentiels affectés par les maladies neuromusculaires. Cette image a été faite avec BioRender.

Souvent, ces maladies sont dégénératives, signifiant que l'état du patient ne fera qu'empirer avec le temps. De plus, ces maladies peuvent se manifester plus tard dans la vie, mais à d'autres moments, elles peuvent apparaître à la naissance (congénitale) ou dans la petite enfance. Les maladies neuromusculaires peuvent provenir d'un variant *de novo* (acquise) ou être héréditaires. Les modes de transmission possible sont résumés dans la figure 2: autosomique récessive, autosomique dominante, et lié à l'X (3). Aussi, le variant peut provenir de l'ADN mitochondrial (3). Dans certains cas, une maladie neuromusculaire peut résulter d'un trouble du système immunitaire (4).

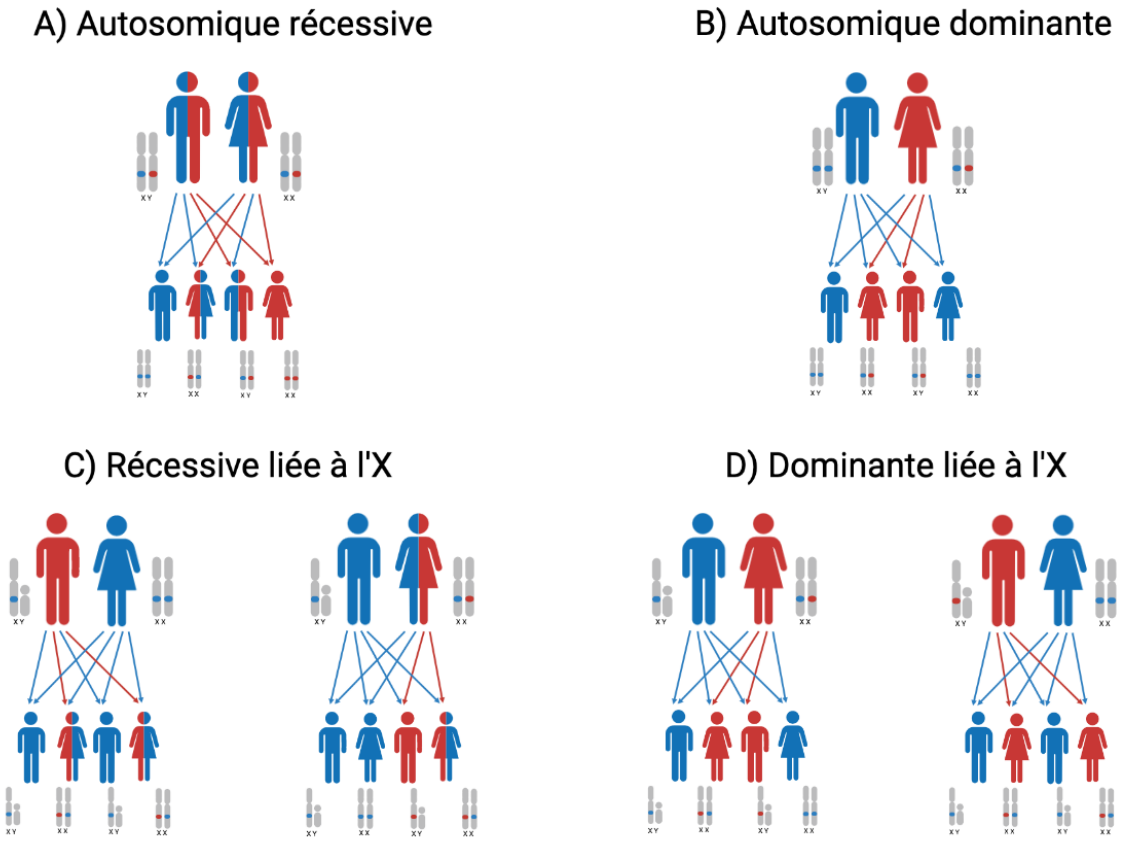

Figure 2. – Les différents types de transmissions impliquées chez les maladies neuromusculaires.

Le code de couleur de cette figure représente si l'individu est non affecté (bleu) , porteur de la mutation (bleu et rouge) et affecté (rouge). A) Description d'une transmission autosomique récessive B) Description d'une transmission autosomique dominante. C) Description d'une transmission récessive liée au chromosome X. D) Description d'une transmission dominante liée au chromosome X. Cette image a été faite avec BioRender.

Présentement, il n'y a pas de traitements curatifs pour les maladies neuromusculaires, de plus, il n'est pas possible de renverser les effets des maladies. Les approches thérapeutiques actuellement disponibles peuvent ralentir la détérioration de l'état du patient et aider à soulager les symptômes du patient. Il est reconnu qu'en général une intervention précoce peut grandement améliorer la qualité de vie du patient. Pour accélérer le temps de diagnostic, une approche plus spécialisée et personnalisée pour ces types de maladies neuromusculaires rares est nécessaire.

1.1.1 Les Myopathies

Les myopathies désignent un large éventail de maladies neuromusculaires affectant principalement les muscles squelettiques. La gravité des symptômes varie selon la personne et la maladie. Certains peuvent présenter des symptômes potentiellement mortels et d'autres peuvent avoir une forme plus bénigne de la maladie. Néanmoins, sans diagnostic, il est difficile d'identifier le gène ou les voies dysfonctionnelles, laissant ainsi le patient sans possibilité d'une intervention médicale personnalisée.

Les myopathies peuvent être classées en deux groupes principaux en fonction de la manière dont la maladie apparaît chez un patient. Lorsqu'il y a des antécédents familiaux, la maladie est très probablement héritée des parents, ce qui fait que le patient est né avec une mutation génétique. Les patients peuvent également avoir hérité d'un gène défectueux provoquant des erreurs métaboliques telles que des défauts enzymatiques. Les types héréditaires de myopathies comprennent les myopathies mitochondriales ; myopathies congénitales; myopathies métaboliques; myotonie et canalopathies. (5) Pour les myopathies acquises, il s'agit des myopathies inflammatoires ; myopathies infectieuses ; les myopathies endocriniennes ; myopathies à médiation électrolytique ; myopathies médicamenteuses et toxiques. Par ailleurs, les types de myopathies acquises sont causés par de nombreux facteurs qui influencent le risque de développer ces troubles tels que certains médicaments ; les toxines; les infections par des virus ou des bactéries; l'inflammation; les minéraux; les électrolytes; et les irrégularités hormonales (5).

1.1.2 Les dystrophies musculaires

Une autre forme de maladies neuromusculaires est les dystrophies musculaires. Ces maladies ressemblent aux myopathies à bien des égards, mais elles regroupent spécifiquement des maladies où la faiblesse musculaire et l'atrophie musculaire sont les principaux symptômes observés pour le patient. Une distinction importante entre les deux repose sur la biopsie musculaire. Pour les dystrophies musculaires, une anomalie est observée dans les éléments structurels du muscle alors que les myopathies ont des anomalies dans les fibres musculaires contractiles (6). De nombreux gènes sont impliqués dans les dystrophies musculaires ce qui rend difficile l'identification du gène responsable dans certains cas. La mutation à l'origine de la pathologie peut être héritée par le chromosome X (liée à l'X), de manière autosomique dominante ou récessive (7).

On observe que dans les cas les plus courants de ce type de maladie, il s'agit souvent du long gène Dystrophine située sur le chromosome X, expliquant ainsi pourquoi les dystrophies musculaires sont plus souvent présentes chez les hommes en raison de leur copie unique de leur chromosome X (hémizygote) (8, 9). Les dystrophies musculaires regroupent de nombreux sous-types de la maladie, certains plus communs que d'autres. Puis, différentes délétions ou mutations de gènes sont la cause des divers types, engendrant plusieurs défauts enzymatiques ou métaboliques (10). Par exemple, les dystrophinopathies telles que la dystrophie musculaire de Duchenne et Becker, les dystrophies musculaires myotoniques, et les dystrophies musculaires congénitales ont des variations génétiques spécifiques provoquant différentes représentations de dystrophies musculaires. Chacun de ces sous-types ont des biomarqueurs connus et, dans certains cas, le patient peut apparaître symptomatique pour la maladie, mais n'a pas le marqueur génétique connu, ce qui rend difficile le diagnostic du patient. Une approche analytique comportant de multiples étapes est nécessaire pour explorer les génomes de ces patients plus complexes.

1.1.3 Les myopathies mitochondriales

Les muscles ont constamment besoin d'énergie pour pouvoir effectuer la contraction et la relaxation musculaire. Le muscle est un organe qui contient beaucoup de mitochondries puisque

sa fonction principale est de fournir de l'énergie à l'organe en synthétisant l'ATP, ce processus est connu sous le nom de phosphorylation oxydative (11). En ayant une mutation génétique altérant directement ou indirectement la fonction des mitochondries, cela peut conduire à une classe de troubles musculaires connue sous le nom de myopathie mitochondriale. Ces types de maladies peuvent entraîner une faiblesse musculaire, des problèmes de coordination, des problèmes d'équilibre et d'autres symptômes où le muscle est requis pour accomplir la tâche. Les mutations provoquant un dysfonctionnement du gène peuvent apparaître de manière sporadique (aléatoire) ou elles peuvent être héritées soit du génome nucléaire, soit du génome mitochondrial. Étant donné que les mitochondries sont présentes dans presque toutes les cellules du corps humain, le génome mitochondrial affecté peut entraîner un dysfonctionnement d'autres organes, ce qui rend le diagnostic très complexe et multisystémique (figure 3) (12).

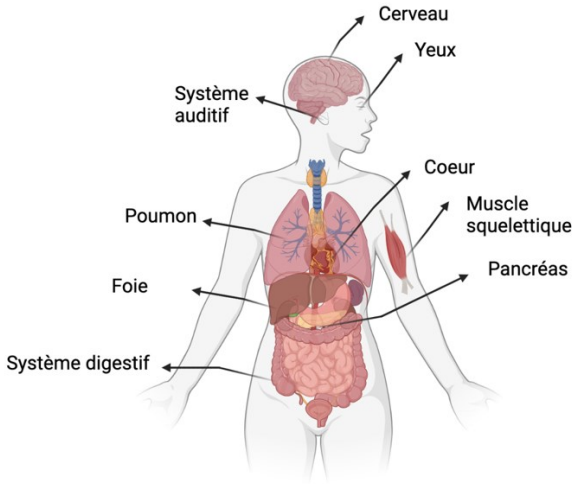

Figure 3. – Les différents organes affectés par des maladies mitochondriales.

Cette image a été faite avec BioRender.

Comme expliqué dans les sections précédentes pour ces types de maladies, il n'existe actuellement aucun traitement disponible pour les patients atteints de myopathies mitochondriales. Certaines cibles de traitement existent, mais reposent sur des suppléments de trois substances naturelles dans les réactions ATP : la créatine ; la carnitine et le coenzyme Q30 (13). Ces suppléments n'ont pas d'effets secondaires majeurs comparés à l'utilisation de

médicaments. Ils aident le patient en comblant le déficit de substance, causé par la maladie, dans les processus ATP au niveau des mitochondries.

1.2 Introduction du contexte biologique et méthodologique : exploration des données omiques

Dans la section précédente, nous nous sommes concentrés sur les notions cliniques qui affectent potentiellement les patients de cette étude. Afin d'effectuer de bonnes analyses, il est nécessaire de se familiariser avec les différents types de maladies neuromusculaires pour mieux comprendre les patients à l'étude. Pour investiguer plus en profondeur ce qui se passe au niveau moléculaire et de mettre en place une approche bio-informatique, il faut explorer plusieurs sujets traités dans les prochaines sections.

Le terme omique regroupe plusieurs disciplines technologiques de la biologie telles que la génomique, la transcriptomique, la protéomique et la métabolomique (14). Celles-ci permettent une investigation de l'ADN, l'ARN, les protéines et la biochimie des métabolites afin de tenter de comprendre la cause moléculaire du phénotype du patient. Sachant que plus de 80% des maladies rares ont un patrimoine génétique complexe, l'exploration de différentes omiques permet de mieux comprendre les mécanismes complexes des maladies rares (15). Dans le cadre de ce projet, la génomique et la transcriptomique sont principalement analysées. Dans le chapitre de la discussion, nous allons étudier les pistes de recherches possibles en lien avec la protéomique et métabolomique.

Le séquençage de deuxième génération et la bio-informatique ont fait grandement progresser le monde de la science médicale permettant ainsi l'étude du génome humain. Aujourd'hui, plusieurs méthodes sont utilisées pour tenter de diagnostiquer un patient tel que les panels de gènes, séquençage du génome complet (WGS), séquençage de l'exome (ES) et séquençage de l'ARN (RNAseq) (figure 4). Ces techniques seront décrites dans les prochaines sous-sections, voici une brève description de celles-ci. Les panels de gènes analysent la région du gène présélectionné. Le WGS permet l'analyse du génome complet. L'ES permet l'analyse des

régions codantes pour des protéines, soit les exons présents dans le génome. Puis le RNA-Seq (protocole sélection poly-A) permet de séquencer le brin d'ARNm, soit le transcriptome. Chacune des technologies présentées produit des séquences de lectures ("reads" en anglais) correspondant à des paires de bases déduites d'un fragment d'ADN. Les chevauchements des lectures permettent de reconstruire la séquence génomique séquencée (soit celle du patient à l'étude dans notre cas) à partir d'un génome de référence lors de l'étape de l'alignement.

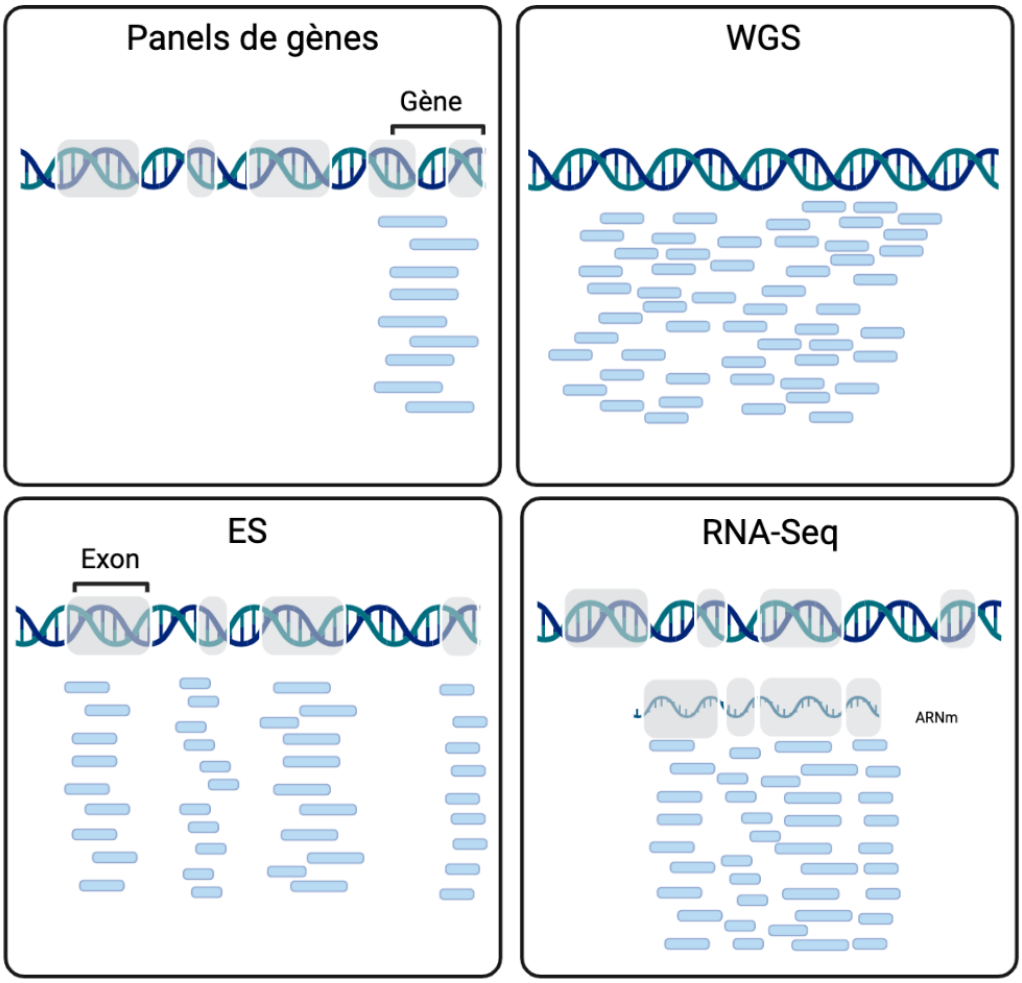

Figure 4. – Les différentes techniques des séquençages génomiques pour poser un diagnostic.

Chaque section de cette figure décrit une technique différente. En haut à gauche on retrouve les panels de gènes. En haut à droite on retrouve la technique du WGS. En bas à gauche on retrouve la technique du ES. Puis en bas à droite on retrouve la technique du RNA-Seq. Les exons sont représentés par un rectangle gris pâle. Les séquences de lectures sont représentées par des bâtonnets bleu pâle. Cette image a été faite avec BioRender.

1.2.1 Le séquençage du génome complet (WGS) et de l'exome (ES)

Le génome représente l'ADN contenu dans une cellule de notre corps, une copie de celle-ci est retrouvée dans presque toutes nos cellules. Approximativement 99.9% du génome est identique d'une personne à l'autre, mais ce qui nous intéresse en recherche médicale afin de prédire, prévenir, diagnostiquer et traiter une maladie est le 0.1% de variations de l'ADN (16). Le génome regroupe tout le matériel génétique dans 23 paires de longues molécules d'ADN étroitement enroulées: les chromosomes. Ceux-ci contrôlent la fonction et la composition de chaque cellule. L'ADN est formé par quatre bases nucléotidiques (Adénine, Thymine, Cytosine, Guanine). Le génome inclut les régions codantes pour des gènes (exon) et non codantes (intron) (17). Les gènes sont les régions codantes qui sont transcrites par l'ARN polymérase et traduites en protéine (18). On retrouve de l'ADN dans le noyau (ADN nucléaire) et mitochondrie (ADN mitochondrial) (19).

Afin d'effectuer un ES, l'ADN est extrait de l'échantillon (musculaire dans le cadre de ce projet); séparé en différents fragments (reads) afin de construire une librairie; ajout des adaptateurs aux fragments d'ADN; enrichissement des exons par hybridation en phase aqueuse; les fragments hybrides (fragments d'ADN + séquence complémentaire aux adaptateurs) sont conservés et amplifiés afin de procéder au séquençage (20). Le WGS est similaire à l'ES, mais certaines étapes diffèrent. Afin d'effectuer un WGS, l'ADN génomique est extrait; ADN est fragmenté de manière aléatoire; sélection de taille du fragment par électrophorèse; amplification et séquençage à haut débit (21).

Les maladies neuromusculaires sont difficiles à diagnostiquer, car elles impliquent parfois de grands gènes qui contiennent de nombreux polymorphismes. Cela signifie que la maladie peut être acquise par plusieurs variants de faible pénétrance ou une combinaison de facteurs génétiques et environnementaux. Plusieurs études au courant des dernières années ont bien montré comment WGS et ES permettent la détection de variants (22). Une étude en 2020 a montré qu'en utilisant ES, ils étaient capables d'identifier un nouveau variant homozygote du gène *PREPL* sur le chromosome 2 c.1940G>A (p.Arg647Gln) hérité des deux parents porteurs pour

un patient présentant des symptômes de syndrome myasthénique 22 (23). Dans une étude que nous avons menée, nous avons également utilisé le séquençage de l'exome pour identifier un variant dans le gène *KIF4A* présent chez une famille de quatre individus (Article en révision) (24). Malgré le coût plus élevé du WGS, c'est en fait une approche qui est très bien documentée pour montrer son pouvoir de diagnostic. Lorsque possible, avoir accès aux données de WGS permet de faire plusieurs analyses différentes dont l'identification de variants, l'analyse des CNVs, des expansions de répétitions et des gènes de fusions (25-27). Ces concepts seront décrits dans les sous-sections suivantes. Dans une étude, ils ont utilisé les données provenant du WGS afin d'identifier des perturbations de la transcription pour des variants non codants des gènes de *ARPC1B*, *GATA1*, *LRBA* et *MPL* (27). Une autre étude démontre que le WGS détecte des variants pathogéniques manqués par les méthodes de diagnostic de NGS actuelles pour la maladie rétinienne héréditaire (28). Puis, une troisième étude démontre que WGS a identifié des variants chez 41 % des individus, soit 18 nouveaux diagnostics comprenant des variants structurels non exoniques et non détectables avec ES (29). De telles études supportent que le WGS est une technique de diagnostique puissante ainsi permettant la caractérisation de l'étiologie de patients atteints de maladies rares.

Ces approches ont permis d'identifier les gènes responsables de nombreuses maladies génétiques rares au fil des années et sont encore fréquemment utilisées dans de nombreuses études à ce jour. Il est tout de même important de garder à l'esprit que selon le type d'analyses, d'autres approches comme le RNA-Seq peuvent parfois être un meilleur choix.

Avant tout, il faut brièvement présenter comment les variants seront identifiés à partir de données de séquençage. Les variants ressortis dans ce mémoire sont identifiés par deux outils, GATK (30) et VarDict (31). Ces deux outils servent à identifier les variations nucléotidiques, soit les variants, dans les séquences alignées avec un génome de référence. Ils peuvent identifier des variants mononucléotidiques (SNV), variants multinucléotidiques (MNV) et des insertions et délétions de nucléotides (InDels) (31). Il existe plusieurs outils pour détecter les variants, chacun ayant leur propre algorithme. La performance de ceux-ci peut même varier dépendant du type

de données analysées. Plusieurs utilisent GATK, on pourrait même dire qu'il est considéré comme l'outil de référence pour la détection de variants. Par contre, il est reporté par nos expériences *in silico* du laboratoire et par d'autres, que l'outil identifie des variants artefactuels (32). Sur ce, nous avons donc choisi d'utiliser également l'outil VarDict, car il semble performer mieux (32).Une description de la méthode d'utilisation de ceux-ci sera décrite plus loin dans le chapitre 3.

1.2.1.1 Les CNVs

La variation du nombre de copies (CNV) fait référence à la diversité du nombre de copies d'un gène présent dans le génome par rapport à un génome de référence. Ces biomarqueurs ont déjà été associés à des maladies mendéliennes et sporadiques telles que la maladie de Parkinson ; atrophie musculaire spinale; ataxie spinocérébelleuse de type 20 (33). Une perte ou un gain de segments génomiques peut survenir à partir des CNVs. Ceux-ci peuvent causer des variations structurelles génomiques telles que des délétions, des duplications, des insertions, des translocations et inversions déséquilibrées, entraînant possiblement une perte ou un gain de segments génomiques. La formation de ces biomarqueurs implique des réarrangements génomiques ; recombinaison homologue non allélique (NAHR) et non homologue "end-joining" (NHEJ), et rétrotransposition (22). Bien que les approches NGS soient excellentes pour détecter les CNVs, cela reste un défi en raison des lectures courtes et des biais de contenu de nucléotides GC (Guanine Cytosine) (34). Plusieurs outils de détection des CNVs ont été développés à la fois pour l'ES et le WGS. Par contre, le WGS tend à fournir une identification plus précise en raison d'une couverture plus uniforme et de la possibilité d'identifier les points de rupture localisés dans les régions introniques (35). Les CNVs ne sont pas facilement bien interprétés, c'est pourquoi la combinaison de données omiques peut aider à comprendre les processus menant aux CNVs pathogéniques.

L'analyse des CNVs peut considérablement augmenter le rendement du diagnostic dans les troubles musculaires (36). Un outil fréquemment utilisé pour ce genre d'analyse est CNVkit (37). Nous avons choisi d'utiliser cet outil basé sur la performance reportée par d'autre ainsi que notre propre expérience avec l'outil dans le laboratoire (38). Il utilise à la fois des bacs ("bins" en anglais) de lectures cibles et de lectures hors cible (position génomique entre les régions ciblées) pour calculer les ratios de copie logarithmique en base 2 (Log2) à travers le génome. L'outil infère

les CNVs par un algorithme implémenté évaluant la profondeur de la couverture c'est-à-dire le nombre de lectures dont la séquence (BAM) est alignée à un bin des régions cibles et hors cibles de référence, ensuite normalise le nombre obtenu par la longueur de ce bin (39). Pour obtenir les régions cibles et hors cible, l'outil implémente un pipeline qui fonctionne en tandem avec des outils d'appel de SNP et indels utilisant des modèles de profondeur de lecture pour estimer le nombre de copies sur l'intégralité de chaque chromosome. La visualisation se fait à partir des ratios et segments Log2. L'outil produit deux types de fichiers utiles pour l'analyse, soit Ratios Log2 au niveau du bac (.cnr) et Ratios Log2 segmentés (.cns). Dans le fichier .cnr, il contient pour chaque bac le chromosome; la position du début et fin de l'exon; le gène auquel appartient l'exon; le Log2 qui infère le CNV; le score du poids. Puisque tous les bacs ne fournissent pas la même quantité d'informations, le score de poids se sert d'une référence. S'il s'agit d'un bac particulièrement bruyant dans la référence, il supposera que ce bac sera également bruyant dans l'échantillon à l'étude et, par conséquent, attribuera à ce bac un poids inférieur. Dans le fichier .cns, le score de sonde est ajouté aux autres scores présentés précédemment, indiquant le nombre de bacs couverts par le segment. Il est possible d'extraire ces résultats en format VCF, ce qui est plus facile pour l'analyse. Les seuils recommandés selon la documentation de l'outil pour les valeurs de Log2 sont décrits dans le tableau I, soit le nombre de copies selon le rapport de Log2. Donc par exemple, toutes les valeurs de Log2 entre le seuil de 0.3 et de 0.7 sont assignées un nombre de copies de 3.

Tableau I. – Les seuils de Log2 décrivant le nombre de copies du variant pour l'outil CNVkit

Log2	*Nombre de copies*
-1.1	0
-0.4	1
0.3	2
0.7	3
...	...

Le nombre de copies (CN) décrit s'il est question d'une délétion; d'une perte; d'un neutre; d'un gain; d'une amplification multiple. Un CN = 2 est interprété comme neutre, il n'y a pas de perte ou gain de matériel génétique. Un CN = 0 signifie qu'il y a une perte complète sur les deux allèles, donc une délétion d'information génétique. Un CN = 1 signifie qu'une seule copie de l'allèle est retrouvée ainsi on interprète ceci comme une perte d'information génétique. Pour CN = 3, il y a un gain d'information génétique dans la région génomique concernée et CN = 4 est interprété comme une amplification multiple. Une représentation visuelle de ceci se retrouve dans la figure 5 ci-dessous.

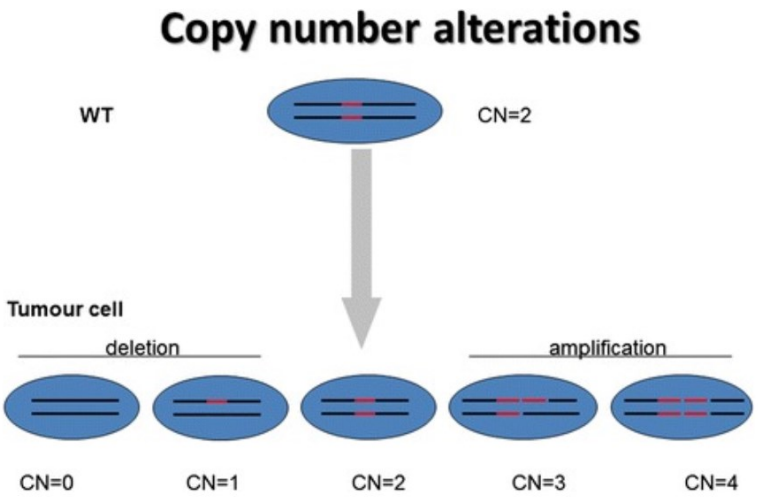

Figure 5. – Description des différents types de CNVs étudiés (40).

En bas à gauche on observe les CN de délétion et en bas à droite les CN d'amplification. Les paires d'allèles sont décrites par les deux lignes parallèles, les gènes par des sections rouges. CN =2 en haut de la figure représente une copie diploïde neutre du gène (WT: "Wild Type" en anglais).

1.2.1.2 Les expansions de répétitions

De façon intéressante, notre corps est capable de se protéger des mutations sporadiques en réparant les séquences d'ADN endommagées. Pourtant, ce même processus semble être impliqué dans de nombreuses maladies génétiques, soit les expansions de répétitions. Un trouble de répétition des trinucléotides est causé par une expansion anormale de trinucléotides en tandem instable dans les gènes. Il a également été montré récemment que des expansions répétées de tétra-, penta-, hexa- et mêmes dodécanucléotides ont été associées à des maladies (41). Les conséquences de l'expansion dépendent de l'emplacement de la répétition dans le gène

affecté, de la taille de la répétition, du nombre de répétitions présentes dans l'allèle et de la séquence de la répétition (42). La taille des expansions répétées peut varier d'une génération à l'autre lorsqu'elles sont héréditaires et d'un individu à l'autre, expliquant peut-être la large variété phénotypique observée dans ces maladies. Alors que la majorité de ces expansions répétées pathogéniques sont héritées de manière autosomique dominante, certaines sont autosomiques récessives ou récessives liées à l'X (41). Une corrélation génotype-phénotype entre la longueur des répétitions et la sévérité et l'apparition de la maladie est démontrée pour plusieurs de ces maladies, par exemple les maladies CAG/polyglutamine, comme pour la DM1, l'ataxie de Friedreich, le syndrome de l'X fragile, le syndrome d'ataxie du tremblement de l'X fragile et plusieurs ataxies spinocérébelleuses (41). Un outil intéressant et couramment utilisé pour faire ce genre d'analyse est ExpansionHunter (43-45).

ExpansionHunter (45) estime, par une recherche ciblée dans le BAM/CRAM, la taille des répétitions pour les lectures qui s'étendent, flanquent et sont entièrement contenues dans chaque répétition. L'outil utilise une liste de variants prédéfinis de loci ciblés, puis extrait du fichier BAM/CRAM les lectures pertinentes de chacun et les réaligne à l'aide d'un modèle basé sur des graphes représentant la structure du locus (45). Il est possible à l'aide du fichier en sortie d'avoir une description du nombre de répétitions d'unité pour l'allèle. Par exemple, un STR de 9 signifie que l'allèle comprend une répétition de 9 unités. En utilisant la littérature, il est possible de vérifier pour chacun des locis cibles : le seuil sain; le seuil de porteur de prémutation; le seuil pathogénique. Dans le contexte des expansions de répétions, une prémutation est une taille de répétition qui se retrouvent entre le seuil normal et le seuil pathogénique. Les prémutations ont souvent une pénétrance incomplète avec certains individus présentant des symptômes et d’autres non (41). Avec les méthodes NGS, il est difficile d'identifier ces événements en raison de la faible longueur des lectures de séquençage. Les analyses bio-informatiques utilisant les données WGS sont désormais capable de détecter une expansion anormale, mais fonctionne mal dans son estimation de la taille et de la séquence d'expansion ainsi qu'en présence de grandes expansions répétées (46). Par conséquent, nous optons pour l'utilisation d'une approche de séquençage à lecture longue pour aider à identifier et valider ces expansions répétées.

1.2.3 Le séquençage d'ARN

Le transcriptome contient les transcrits d'un tissu ou d'une cellule spécifique à un moment précis. Souvent, dans les maladies neuromusculaires, une biopsie musculaire est effectuée pour avoir accès directement aux tissus affectés contenant des cellules spécifiques d'intérêt pour ce type d'analyse. L'accès à ces données permet une meilleure compréhension des gènes exprimés et des voies impliquées dans le tissu affecté.

Les méthodes de séquençage de nouvelle génération sont fréquemment utilisées pour ce type d'analyse, telles que RNA-Seq (47). Sans entrer dans les détails du processus de préparation des échantillons pour le RNAseq, la première étape consiste à isoler l'ARN. La qualité et la quantification sont des facteurs importants à prendre en compte pour ce procédé afin de disposer d'une librairie riche (48). Plusieurs protocoles de librairie existent pour le séquençage de l'ARN et chacun d'eux a ses propres avantages et inconvénients. Le protocole choisi dépend vraiment du projet et de l'hypothèse abordée. L'approche la plus générale, et celle utilisée dans cette étude est le protocole de sélection polyA. Il utilise la queue 3' polyadénylée afin d'isoler et de quantifier précisément l'ARN codant (48). Pour d'autres types d'analyses où l'ARN non codant et le pré-ARNm non modifié post-transcription sont au centre des préoccupations, des librairies appauvries en ribosomes ("ribosome depleted" en anglais) peuvent être utilisées. De plus, pour les petits ARN non codants impliqués dans la régulation de l'expression génique après la transcription, le protocole de sélection de taille pour enrichir en microARN peut être utilisé (48).

Des études, se servant du RNA-Seq, ont démontré l'utilité de cette technique dans un contexte clinique. Dans une étude ayant une cohorte de 25 cas exome/panels de gènes de maladies neuromusculaires non résolus, les chercheurs ont déterminé avec RNA-Seq chez 36% (9/25) de ces cas l'explication génétique (49). Un aperçu de la littérature décrit qu'une augmentation de 15% des diagnostics reportés à date est grâce au RNA-Seq (50). Cette technique s'avère un outil de diagnostic efficace, permettant l'identification des variants; l'épissage aberrant; l'expression aberrante et fusion de gènes. Avec cette approche, le transcriptome des tissus affectés de nos patients peut être analysé.

1.2.3.1 L'épissage alternatif

L'épissage alternatif est un processus essentiel dans la régulation de l'expression des gènes, permettant à un gène de coder plusieurs protéines différentes. Il joue un rôle important dans les maladies neuromusculaires et dans de nombreuses autres maladies. L'épissage alternatif se produit dans plus de 90% des gènes et le muscle squelettique a l'un des taux les plus élevés d'épissage alternatif spécifique aux tissus (51). Ceci s'explique par le fait que les protéines musculaires sont plus volumineuses. Leurs transcrits contiennent de nombreux exons, des structures unitaires répétitives et un épissage alternatif étendu (51). En bref, la façon que ce processus fonctionne, est que la séquence d'ADN subit une transcription en ARN prémessager, puis par des événements de coupure d'intron et de ligature d'exon, elle est traduite en plusieurs protéines à partir d'un seul gène (52). Cinq types différents d'événements d'épissage alternatif existent (figure 6): site d'épissage alternatif en 5' (A5SS); alternative 3' (A3SS); les exons mutuellement exclus (MXE); rétention d'intron (RI); saut d'exon (SE). Plusieurs facteurs influencent les exons inclus dans l'ARNm mature, notamment les signaux d'épissage constitutifs ou alternatifs, un exon plus court ou une conservation de séquence supérieure entourant les exons alternatifs. Les exons de l'ARNm mature sont définis en fonction des interactions entre les éléments agissant en cis et en trans. L'épissage en trans signifie que deux exons de gènes différents sont joints pour former un brin d'ARNm, alors qu'en cis il s'agit d'un épissage de deux exons sur le même brin d'ARN d'un gène. Les éléments activateurs, agissant en cis liés par des facteurs positifs tels que les protéines riches en sérine/arginine (SR), sont reconnus par des protéines activatrices. Les éléments silencieux, agissant en trans, sont liés par des protéines répressives telles que les ribonucléoprotéines nucléaires hétérogènes (52). L'épissage alternatif peut provoquer des protéines défectueuses, tronquées, manquantes ou trop nombreuses dans un complexe d'interactions. Lors de l'alignement de séquence, les transcrits épissés sont représentés par deux séquences de lectures distinctes qui sont liées par une jonction d'épissage. Lorsqu'il s'agit d'épissage alternatif cryptique, on retrouve soit de nouvelles jonctions entre deux exons ou une perte/gain de jonctions (53) . Dans un contexte pathologique, une variation génétique peut créer ou abolir des sites d'épissages et ainsi causer un épissage alternatif cryptique. Bien que ce processus soit associé à de nombreux troubles, les thérapies de modulation

d'épissage peuvent être utilisées pour des traitements potentiels de maladies neuromusculaires génétiques rares. Ces approches manipulent le résultat du processus d'épissage et permettent de restaurer le fonctionnement normal du gène défectueux qui est généralement causé par un événement d'épissage anormal. Certaines études basées sur la myopathie de Duchenne, l'amyotrophie spinale et d'autres dystrophinopathies utilisent cette médecine génétique personnalisée dans des essais cliniques (54, 55).

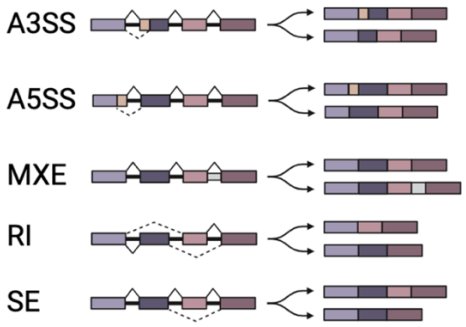

Figure 6. – Les différents types d'épissage alternatif.

Cinq types différents d'événements d'épissage alternatif présentés: site d'épissage alternatif en 5' (A5SS); alternative 3' (A3SS); les exons mutuellement exclus (MXE); rétention d'intron (RI); saut d'exon (SE). Cette image a été faite avec BioRender.

Deux outils seront utilisés pour étudier les évènements d'épissage alternatif cryptique: SpliceAI (56) et rMATS (57). Le choix des outils se base sur une analyse comparative que nous avons effectuée afin d'utiliser l'outil qui fonctionne le mieux pour nos types de recherche. De plus, la littérature contient de nombreuses études d'analyses comparatives (“Benchmarking" en anglais) soutenant le fait que SpliceAI est un excellent outil pour étudier l'épissage (58-61). L'outil SpliceAI est un réseau neuronal (profondeur de 32 couches) qui prédit avec précision si chaque position est un donneur, un accepteur d'épissage ou aucun à partir d'une séquence de transcription pré-ARNm provoquant un épissage cryptique. Le réseau neuronal apprend à partir de la séquence primaire en évaluant 10 000 nucléotides de séquence flanquante (séquence donnée en entrée) pour prédire la fonction d'épissage de chaque position dans le pré-ARN messager (56). Les forces qui sont ressorties dans la littérature sur celui-ci sont qu'il permet de réduire le temps d'exécution, le coût de diagnostic et les performances sont très spécifiques (61). De plus, SpliceAI est capable d'identifier les mutations non codantes pouvant influencer l'épissage

alternatif comprises dans le séquençage, il peut reconnaitre des déterminants nucléotidiques d'épissage sur de très grandes distances génomiques et finalement c'est un réseau d'apprentissage en profondeur lui permettant d'apprendre à prédire les positionnements des variants causant de l'épissage cryptique (56, 58). Ses faiblesses sont qu'il ne reflète pas complètement le comportement du spliceosome, les variants introniques profonds sont difficiles à prédire, il utilise que des séquences primaires pour l'apprentissage du réseau et a de la difficulté à différencier les jonctions d'épissage qui varie d'un tissu à l'autre (56). En somme, l'outil produit tout de même quelques erreurs de faux positifs, mais a une meilleure performance comparativement à plusieurs autres. SpliceAI utilise quatre scores et le maximum de chaque score est appelé le score delta du variant (DS). Le score delta d'un variant va de 0 à 1 et peut être interprété comme la probabilité que le variant modifie l'épissage.

Tableau II. – La description des quatre scores delta (DS) de SpliceAI

Scores	***Définition***
DS_AG	Gain accepteur
DS_AL	Perte accepteur
DS_DG	Gain donneur
DS_DL	Perte donneur

Le prochain outil est rMATS, c'est un outil de calcul pour détecter les événements d'épissage alternatif différentiel à partir de données RNA-Seq sortant une liste de jonction exon-intron d'épissage potentiel. Ses forces sont qu'il modélise les niveaux d'inclusion d'exon, détecte de nouveaux évènements d'épissage non annotés et il y a une bonne puissance statistique et vitesse computationnelle. Aussi, il détecte plus de gènes provenant de l'épissage différentiel, de la stabilité de la précision et de la détection des validations qPCR et il analyse tous les types majeurs d'épissage alternatif (62, 63). Ses faiblesses sont qu'il analyse l'épissage que sur des lectures de la même longueur, il n'est pas le plus rapide, il utilise plus de mémoire que les autres et les valeurs aberrantes causent une baisse significative dans la performance (62, 63). Le modèle statistique de rMATS calcule la valeur P et le taux de fausses découvertes, selon lesquelles la différence du rapport isoforme d'un gène entre deux conditions dépasse un seuil défini (IncLevelDiff).

Pour nos analyses, les deux outils (SpliceAI et rMATS) seront utilisés pour optimiser la détection d'évènements d'épissage alternatif potentiellement pathogéniques. Comme expliqué, l'outil SpliceAI associe des scores de prédiction à un variant. Il prédit s'il y a possibilité d'un épissage cryptique à ce changement nucléotidique. D'autre part, rMATS prédit au niveau des jonctions d'épissage s'ils sont potentiellement cryptiques. Il serait intéressant qu'un variant identifié par SpliceAI soit à proximité d'une jonction d'épissage possiblement cryptique identifiée par rMATS. Cet évènement serait ciblé par les deux outils, malgré leur différence algorithmique, supportant davantage l'hypothèse qu'il soit un évènement pathogénique.

1.2.3.2 Les gènes différentiellement exprimés (DEGs)

L'étude de la transcription des gènes en protéines et l'analyse de l'expression des gènes est une étape essentielle pour la recherche de variant pathogénique et la compréhension de la variation phénotypique entre patients (64). Il est estimé qu'entre 9-30% des variants pathogéniques affectent l'expression des gènes (65). Le génome contient des gènes qui codent pour une protéine unique et de nombreux processus biochimiques doivent se produire pour les générer. Si une protéine est anormalement exprimée, cela peut entraîner un dysfonctionnement du processus, provoquant ainsi le trouble. Pour les troubles neuromusculaires, étant donné que l'expression des gènes et l'isoforme de l'ARNm varient beaucoup d'un tissu à l'autre, l'accès au tissu affecté peut aider à expliquer les variations génétiques (66). De plus, si un variant est soupçonné de causer la maladie, ce type d'analyse peut être une indication que le variant affecte l'expression du gène contribuant au trouble. Alors que d'autres études utilisent l'expression génique globale du tissu pour décider par où commencer la recherche de variants candidats (67). Dans l'ensemble, l'utilisation des données de transcriptome obtenues par RNA-Seq est une excellente représentation de l'activité génique, et de nombreux outils sont disponibles pour ce type d'investigation (68).

Pour effectuer ce genre d'analyse, il est statistiquement plus pertinent d'utiliser des répliques biologiques/techniques. Comme les échantillons des patients proviennent de biopsie musculaire, ceci n'est pas possible pour notre projet. Donc, il est important de noter que les résultats de la recherche de gène différentiellement exprimé sont utilisés de manière

exploratoire. Un outil très populaire à utiliser normalement pour ce genre d'analyse est DESeq2 (69). DESeq2 utilise un modèle linéaire binomial négatif pour chaque gène (paramètre de coefficient et dispersion) et utilise le test de Wald pour les tests de signification (70). Une étape de normalisation des comptes est effectuée afin d'éliminer les comptes de gènes sous le seuil moyen (71). Autrement, un outil créé pour des jeux de données sans réplique serait plus approprié pour ce projet. Un tel outil serait LPEseq (72), il teste l'expression différentielle basée sur l'erreur groupée locale. L'outil détermine d'abord les bacs d'intensité et évalue la distribution des erreurs groupée locale. Pour les cas non répliqués: il supprime les valeurs aberrantes; effectue un ajustement de régression non paramétrique puis une estimation de variance pour prédire les gènes différentiellement exprimés (72). Malgré le fonctionnement de l'outil pour des expériences sans réplique, les conclusions qui en sont tirées peuvent être limitées et doivent être interprétées avec des précautions supplémentaires (fausses valeurs aberrantes signalées). Plusieurs valeurs sont fournies par l'outil d'analyse, le score z.stats est interprété comme valeur de Logarithme en base 2 Fold Change (L2FC). Celle-ci s'agit du logarithme en base 2 du ratio entre deux conditions (condition A/condition B), soit le "Fold change" en anglais. Par exemple, un Fold change de 2 signifie que la condition A est deux fois plus grande que condition B , donc ayant un L2FC de 1 (log2(2)=1). Les scores "mean.x" et "mean.y" soustrait indique le changement de L2FC, ce qui signifie que si la différence est de k, le changement de Fold doit être de 2 exposants k (2 ^ k).

1.2.3.3 Les gènes de fusions

RNA-Seq permet d'étudier les fusions de gènes qui sont un processus important à explorer pour la médecine de précision (73). Les fusions de gènes sont formées de deux parties de gènes différentes. Cela peut se produire à partir d'un réarrangement structurel du chromosome tel qu'une translocation, où deux chromosomes non appariés (non homologues) se réarrangent ensemble: inversion; suppressions et insertions. Un réarrangement structurel peut également se produire en raison de l'épissage en trans et en cis du pré-ARNm (74). Dans la littérature, le consensus est que les fusions de gènes sont peu impliquées dans les maladies neuromusculaires, mais plus chez les patients cancéreux.

1.2.4 Le génome mitochondrial

Les maladies neuromusculaires comme les myopathies mitochondriales peuvent survenir à partir d'un défaut de l'ADN mitochondrial (ADNmt). L'ADNmt est beaucoup plus petit que le génome nucléaire (16 569 pb contre 3,3 milliards de pb) codant pour 37 gènes et repose beaucoup sur des protéines codées au niveau nucléaire (75). L'ADNmt est hérité de la lignée maternelle et l'ADNmt paternel est dégradé lors de la fécondation. Le concept d'hétéroplasmie est important ici (figure 7), car chaque cellule contient différents types de copies d'ADNmt, ce qui la rend plus à risque de mutations. Un ratio élevé de génome muté par rapport aux types sauvages ("Wild Type" en anglais) provoque l'apparition de maladies et de nombreuses approches thérapeutiques visent à réduire le ratio d'ADNmt mutant (75). De plus, les mutations de l'ADNmt se sont avérées avoir un effet cumulé avec le temps et ont été associées à de nombreux troubles du vieillissement tels que la maladie de Parkinson, la maladie d'Alzheimer, le cancer, etc... (75). Pour étudier les mitochondries, le NGS séquence avec précision l'ADNmt. L'impact de l'altération du génome mitochondrial est discuté dans une section précédente et est associé à de nombreuses maladies neuromusculaires. Comme ces types de troubles sont souvent multisystémiques, combiner l'exploration de l'ADNmt à l'analyse de l'ADN nucléaire peut améliorer le diagnostic de ces maladies plus complexes.

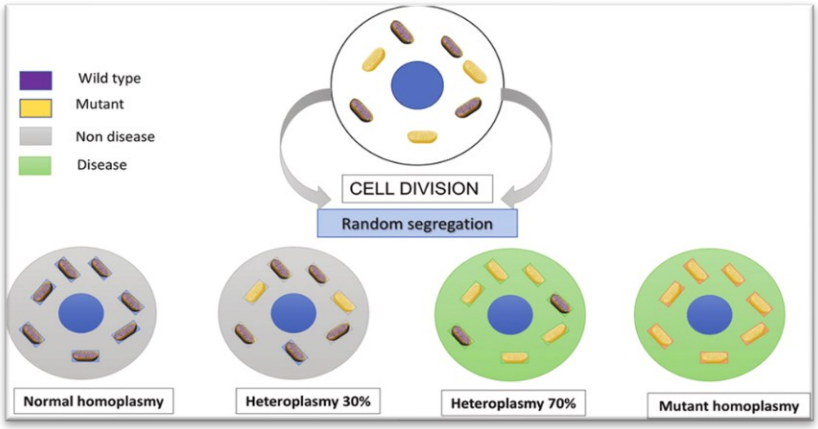

Figure 7. – La description du concept d'hétéroplasmie (76).

L'hétéroplasmie lors de la division cellulaire. Chaque cellule contient des mitochondries. Les cellules en grises sont saines et vertes sont affectées. Les mitochondries mauves sont saines (WT: "Wild Type" en anglais) et jaune sont affectées.

Afin de plonger dans l'information mitochondriale de ces patients, nous avons utilisé MToolBox (77). C'est un pipeline hautement automatisé pour l'annotation de l'hétéroplasmie et l'analyse de priorisation des variants mitochondriaux humains dans le séquençage à haut débit. Nous avons réussi à exécuter cet outil sur des données RNAseq et DNAseq. Les principales étapes du flux de travail MToolBox sont la cartographie ("mapping" en anglais) des lectures; le filtrage des séquences nucléaires mitochondriales; traitement post-mappage; assemblage du génome; prédiction des haplogroupes (groupe de mutations sur un chromosome); annotation des variants. Il n'est pas documenté à ce jour que l'outil fonctionne avec le RNAseq. Pour notre projet, cela nous permet de comparer entre les deux données omiques. L'outil fournit une estimation de la variabilité du site des nucléotides et des acides aminés via les programmes SiteVar et MitVarProt (la variabilité varie de 0 à 1). L'outil utilise la base de données HmtDB contenant des séquences du génome mitochondrial humain annotées avec des données de population et de variabilité (78). Les sites dont la valeur de variabilité est proche de 1 présentent un allèle dont l'état est commun à tout un sous-arbre de l'arbre entier représentatif des génomes stockés dans la base de données HmtDB ou pourraient être soumis à des mutations récurrentes. De faibles valeurs de variabilité peuvent suggérer un nouveau variant définissant l'haplogroupe ou indiquer un score de mutation lié à une maladie rare MitVarProt.

1.2.3 Séquençage de longues lectures (LRS)

Les méthodes précédentes décrites sont des techniques de lecture courte produisant des lectures de 100 à 600 bases. Le séquençage à lecture longue (LRS) peut générer des lectures d'au moins 10 kb. Le séquençage à lecture courte est précis et largement utilisé pour les études de découverte de variants. Le séquençage de certains événements dans de courts fragments complexifie la reconstruction et le comptage de la séquence, manquant souvent des événements importants. L'utilisation de LRS peut affiner l'assemblage *de novo*, la cartographie, l'identification des isoformes de transcription et la détection des variants structurels (79). Bien que le focus de ce projet ne soit pas centré sur le LRS, un objectif du projet était d'inclure cette approche pour valider les évènements trouvés avec les méthodes WGS, ES et RNA-Seq. De nombreuses études se concentrent uniquement sur l'utilisation du LRS et cela est compréhensible en raison des nombreux avantages de l'utilisation du LRS. Certains variants sont contenus dans des régions

difficiles à séquencer. Comme pour les expansions répétées, il est difficile de confirmer qu'elles sont bien présentes et impliquées dans l'apparition de la maladie. Les expansions répétées peuvent dépasser le nombre de bases dans une courte lecture (50-150 pb), donc la répétition n'est pas directement observée. Cette approche peut également être utilisée pour valider ou découvrir des événements d'épissage aberrants suspectés et des rapports d'isoformes (80). C'est pourquoi l'utilisation de LRS peut augmenter la confiance de diagnostic des nouvelles découvertes liées à des expansions répétées (22, 81). Aussi, nous avons mené une étude utilisant le LRS pour l'évaluation fonctionnelle d'un variant VUS sur le gène MLIP, ainsi que l'annotation de transcrit spécifique aux tissus (82). Les VUSs sont des changements d'un nucléotide (variants) qui n'ont pas été préalablement associés à une maladie et dont l'impact fonctionnel n'a pas été démontré, puis le LRS permet l'évaluation fonctionnelle de ceux-ci. De manière intéressante, cette technique nous permet également de déterminer si des variants sont en *cis* ou en *trans* et donc confirmer un diagnostic de maladie récessive (hétérozygote composé) en absence de matériels génétiques des parents (83).

Chapitre 2 – L'intégration des données omiques

Ce second chapitre introduit le principe d'intégration des données omiques. Premièrement on se concentre sur les contraintes de la génomique clinique. Une revue plus critique de la littérature pour les différentes approches de séquençage est abordée au début de ce chapitre. Ensuite, une description de ce qui peut être analysé avec les données d'ADN et d'ARN combinés.

2.1 Les défis de la génomique clinique

Avec l'avancement de la technologie, l'univers de la génomique clinique a grandement avancé. Les investigations génétiques deviennent de plus en plus accessibles étant donné que le prix des techniques de séquençage continue à diminuer. Malgré cela, il y a toujours des défis qui prolongent l'odyssée diagnostique du patient. L'idée générale d'intégrer les données omiques est de combiner les points forts de chaque approche pour une analyse complète et optimale de l'information génétique des patients. Les limitations des approches de panels de gènes, de WGS, ES et RNAseq seront également discutées dans les sections suivantes de ce chapitre.

2.1.1 Les panels de gènes

Le diagnostic des patients présentant une maladie neuromusculaire rare est complexe et nécessite une approche minutieuse. Les panels de gènes ne sont pas l'approche la plus optimale dans ce cas-ci. Les panels de gènes se sont montrés utiles pour le diagnostic de plusieurs maladies dont les variants pathogéniques étaient déjà connus. C'est une méthode à faible coût très utile pour rapidement diagnostiquer un patient. Cependant, l'analyse de quelques centaines de gènes rend cette approche restrictive et non optimale pour l'étude de maladie rare. De plus, une étude dans la littérature décrit que cette approche échoue à 64% pour établir un diagnostic comparativement à une approche utilisant du séquençage de l'exome (84). Dans le cas de maladie rare, cette approche ne parvient pas à identifier les variants causaux de la maladie. Elle utilise des listes de gènes prédéterminés et ne couvre que les exons ce qui est moins de 2% du génome complet. De nouvelles associations gènes-pathologies sont fréquemment identifiées et donc les panels de gènes ne permettent pas de revisiter les données en incluant ces nouvelles

découvertes. Les patients ont besoin d'une investigation plus profonde, plus large et moins biaisée afin de découvrir ce qui cause la pathologie au niveau moléculaire.

2.1.2 Les forces et faiblesses du ES, du WGS et du RNA-Seq

La technique de l'ES séquence tous les exons contenus dans l'échantillon d'ADN donné. Cela signifie que tous les introns et autres régions non codantes ne sont pas inclus dans les données séquencées. Mais dans certaines recherches, il est démontré que cette affirmation n'est pas entièrement vraie. En fait, le séquençage de l'exome entier peut générer un séquençage de haute qualité des introns et des régions régulatrices non codantes puisque certaines de ces régions régulatrices ont été incluses dans le processus de capture (85). L'ES se concentre sur les régions génomiques codantes, ce qui en fait un outil important dans le diagnostic génétique des troubles mendéliens. C'est une approche efficace, plus rapide et moins coûteuse que WGS. Toutefois, le WGS permet une couverture plus complète du génome. Maintenant, la question qu'il faut se poser est de déterminer lequel est la meilleure technique pour détecter les mutations. La réponse dépend principalement du type d'étude menée. Chaque approche a ses propres avantages et ses inconvénients. Si l'étude s'intéresse aux régions régulatrices des génomes, avoir l'information en dehors des exons obtenus avec WGS est plus utile dans ce cas. En revanche, si l'on soupçonne qu'un variant codant est à l'origine du trouble affectant une protéine, il serait plus logique d'utiliser l'ES. Cette approche contient les exons codants pour la protéine et la mutation se situe très probablement dans ces régions. En d'autres termes, ES est une approche de séquençage ciblée limitée dans l'interprétation des VUS, l'identification des réarrangements structurels, les CNVs et les expansions répétées (86). Il est important de noter que plusieurs variants introniques sont manqués par WGS et par ES. De nombreux variants introniques sont étiquetés comme des VUSs, faisant du RNA-Seq un outil important dans le diagnostic génétique en termes d'annotation et d'interprétation de VUS.

Les techniques d'ES et de WGS sont utilisées depuis longtemps pour identifier les variants pathogéniques. Ces approches ne sont pas parfaites, elles peuvent malgré tout manquer un variant pathogénique. Une combinaison aux données transcriptomiques peut optimiser la recherche et la priorisation des variants pathogéniques(87). Les avantages de l'utilisation de RNA-

Seq incluent une faible saturation des signaux ; pas d'hybridation; haute précision et sensibilité; l'épissage alternatif est détecté ainsi que les SNP et les ARN non codants (47). L'inconvénient de l'utilisation de la méthode RNA-Seq est qu'elle nécessite un stockage de données élevé et qu'un certain ensemble de compétences est nécessaire pour préparer le protocole de la librairie, le séquençage et l'analyse des données (47). En utilisant des approches plus traditionnelles comme le WGS et l'ES, ces approches échouent souvent à identifier les variants pathogéniques dans 25 à 50 % du temps (66). La raison étant que pour ES, les variants positionnés dans de grandes délétions ou dans la région intronique sont souvent manqués, car les données de séquençage sont enrichies pour les régions exoniques. Pour le WGS, la caractérisation des variants n'est pas disponible ou est limitée en raison du manque de connaissances sur les fonctions biologiques introniques. Les variants à faible impact comme celles situées dans les UTRs et les variants synonymes ne sont pas prioritaires, et les variants à fort impact comme nonsynonyme; STOP (nonsense); site d'épissage; décalage de cadre; les grandes délétions ne sont souvent pas associées à une maladie (88). De plus, on estime que 30 % des variants à l'origine de maladies se trouvent dans les régions non codantes (88). Sachant cela, les informations contenues dans le transcriptome du patient peuvent potentiellement conduire à la découverte du variant pathogénique.

2.2 Combinaison des données omiques

Individuellement, les techniques de séquençage discutées dans la section précédente ont chacune démontrées leur potentiel de diagnostic, mais certaines contraintes de ces approches font échouer les analyses de recherche de variant pathogénique (tableau III). Les panels de gènes utilisent une liste prédéterminée, dans le cas des maladies rares où nous recherchons un nouveau gène impliqué dans la maladie, nous manquons complètement cela en utilisant cette méthode. Pour l'ES, il ne couvre que les exons, soit moins de 2 % du génome humain (quelques exons sont également manqués par cette méthode). Pour WGS, il existe une interprétation limitée des données génomiques. Puis, au niveau des limites de l'approche RNA-Seq, l'annotation précise des séquences et l'interprétation des données peuvent être difficiles sur le plan informatique. Ces limites combinées à d'autres défis signifient qu'environ 25 % des patients n'ont

pas de diagnostic moléculaire. En combinant les données omiques, on peut augmenter le rendement diagnostique des maladies neuromusculaires rares et complexes.

Tableau III. – Résumé des limitations des différentes techniques de séquençage.

Limites des approches ADN
Panels de gènes Utilise une liste prédéterminée pour un gène connu Couvre uniquement les exons
Séquençage de l'exome (ES) Exons <2 % du génome humain Quelques exons manqués Variants introniques manqués
Séquençage du génome entier (WGS) Interprétation limitée des données génomiques
Limites des approches ARN (RNA-Seq)
Séquençage de l'ARN (RNA-Seq) L'annotation précise et l'interprétation des données Petits transcrits Transcriptions qui se chevauchent

Afin de mieux comprendre l'impact clinique et fonctionnel des variants, une combinaison de ES, WGS et RNA-Seq augmenteraient le rendement diagnostique des maladies neuromusculaires complexes et rares. Cette intégration peut donner une représentation plus large de la complexité biologique des variants pathogéniques (89). Pour améliorer l'interprétation de ces données, RNA-Seq peut aider à comprendre les perturbations transcriptionnelles des changements génétiques. RNA-Seq a déjà été utilisé pour observer l'effet de variants pathogéniques, manqués par les approches traditionnelles. Un exemple de cela pour les maladies neuromusculaires est la découverte d'un événement d'inclusion d'intron dans *COL6A1* utilisant le RNA-Seq pour la dystrophie sévère liée au collagène VI (66). Cela a été observé chez quatre patients qui avaient précédemment été déclarés négatifs avec les tests de suppression/duplication, ES et WGS. Un fait intéressant est que l'inclusion intronique (pseudo-exon) était spécifique au tissu musculaire et représente maintenant la mutation la plus commune

associée aux myopathies-ColVI. Le séquençage de l'ADN a par la suite identifié un variant intronique responsable de cet épissage alternatif. Ce cas montre comment la combinaison du transcriptome avec des informations sur le génome entier et l'exome peut rendre possible le diagnostic de maladies complexes. Comme décrit précédemment, chaque approche a ses propres avantages et inconvénients et leur combinaison optimiserait la recherche de variants pathogéniques (66, 90). Il peut également être intéressant d'intégrer des données omiques pour la détection des débalancements alléliques et pour aider à comprendre l'impact de variants au niveau transcriptomique.

2.2.1 Les débalancements alléliques

Pour améliorer le rendement diagnostic, la combinaison des données d'exome et de transcriptome est effectuée permettant l'analyse de débalancements alléliques. Le débalancement allélique fait référence aux différences observées dans le niveau d'expression des gènes dans différents allèles par un écart d'un rapport normal de 1: 1. Cela peut être causé par une instabilité génomique dans les régions régulatrices générant une perturbation aléatoire et une fusion des chromosomes du génome (91). Un débalancement allélique peut indiquer que l'allèle concerné d'un gène spécifique est dégradé, entraînant une perte de fonction du gène. Par exemple, il a été montré que le débalancement de l'expression allélique était associé à une forme congénitale de dystrophie musculaire. RNA-Seq a identifié une mutation *POMT2* homozygote c.1502A>C (p.Glu501Ala), qui n'a entraîné aucune réduction de l'expression du gène (90). Ce variant homozygote identifiée dans les données d'exome et le transcriptome a montré que les gènes adjacents avaient également une expression monoallélique, mais dans l'ADN maternel, elle a été identifiée comme hétérozygote (90).

2.2.2 Impact des variants introniques sur l'ARN

Pendant longtemps, les régions introniques étaient considérées comme non pertinentes dans la recherche de variants pathogéniques. Toutefois, il y a de plus en plus de preuves que cette affirmation est fausse. Les régions introniques jouent un rôle important dans la régulation des gènes. Lors de la transcription, les introns sont séparés de la séquence génomique, mais ils ne sont pas complètement éliminés. En fait, les informations restantes peuvent affecter l'épissage

alternatif si le variant intronique est contenu dans les sites donneurs et accepteurs. Une variant dans cette région peut très bien conduire à une altération des événements d'épissage normaux. Ceci peut conduire à un épissage aberrant provoquant un changement de l'expression du gène affecté et contribuer aux phénotypes de la maladie. On estime qu'environ 25 % des mutations non-sens et faux-sens connues altèrent les activateurs/silencieux d'épissage exonique et peuvent potentiellement entraîner un épissage mal régulé (92). Cette étude de l'impact sur les variants introniques est très pertinente lorsqu'une maladie est connue pour être liée à un gène long, car le gène est plus à risque de variants pathogéniques. Il a été démontré que ces mutations introniques profondes dans des gènes longs provoquent des troubles neuromusculaires tels que la dystrophie musculaire de Duchenne (49).

2.3 Hypothèses et objectifs

Notre hypothèse est que l'intégration des données omiques permettra de mieux interpréter les VUSs et d'identifier les variants introniques ayant un impact sur l'ARN chez les patients recrutés. L'objectif principal de cette étude est de définir l'étiologie génétique de patients atteints d'une maladie neuromusculaire sans diagnostic moléculaire, malgré plusieurs investigations cliniques et génétiques. Pour ce faire, on propose une approche bio-informatique combinant des résultats provenant de multiple exploration faite au niveau de l'ARN et ADN. Les explorations au niveau de l'ARN sont d'identifier des variants, d'investiguer l'épissage alternatif et les gènes différentiellement exprimés. Au niveau de l'ADN, on veut analyser les CNVs, les expansions de répétitions et identifier des variants. En combinant les données ADN et ARN, on veut analyser s'il y a des débalancements alléliques et déterminer l'impact fonctionnel des variants. Puisqu'on estime que 15 à 60 % des mutations provoquent un épissage alternatif, la combinaison de ces deux approches peut permettre d'identifier des variants et de déterminer un impact fonctionnel au niveau de l'ARNm. Cette intégration de données permettra de minimiser les listes de variants potentiellement pathogéniques pour chaque patient.

Le matériel génétique de nos patients provient de biopsie musculaire afin d'effectuer le séquençage d'ADN et d'ARN. Essentiellement, pour chaque patient une analyse personnalisée de leur génotype sera effectuée afin d'identifier des variants causaux et de donner des pistes sur le mécanisme pathophysiologique. Aussi, en ayant accès aux tissus affectés, nous avons également

la possibilité d'identifier des événements spécifiques au muscle. Ce pipeline facilitera grandement l'exploration génomique dans le cas de maladies neuromusculaires rares issues de troubles hétérogènes.

Chapitre 3 – Établissement et application clinique de l'approche bio-informatique

L'approche bio-informatique a été conçue dans le but de personnaliser la recherche de variant pathogénique pour des patients recrutés présentant des formes variées de maladies neuromusculaires rares. Une description du choix de recrutement ainsi qu'une description clinique de chaque patient sont faites au début de ce chapitre. En tant que bio-informaticien on se limite parfois aux données des patients, mais pour vraiment bien analyser les patients en question il est essentiel de connaitre leur description clinique complète afin de guider l'analyse des données. Il est important de noter que pour certains patients, seulement les données d'ARN ont été extraites dans le cadre de ce projet. Pour ces patients une analyse multi-omique n'était pas possible pour le moment, mais plusieurs analyses du côté de l'ARN ont été effectuées et présentées dans ce mémoire. Aussi, pour ce projet il n'était pas possible d'effectuer de réplique biologique dû au fait que les échantillons de matériel génétique provenaient de biopsie musculaire: il n'était pas possible de faire plusieurs biopsies et les biopsies sont trop petites pour effectuer plusieurs extractions. La seconde partie de ce chapitre sert à expliquer la méthode derrière la mise en place du pipeline ainsi que de présenter les résultats préliminaires du pipeline appliqué sur les données de nos patients. Plusieurs outils et scripts maison pour filtrer et analyser les données multi-omiques des patients seront décrits dans cette deuxième moitié du chapitre afin de faire comprendre le flux de travail utilisé pour ce projet.

L'enchainement des sujets jusqu'à présent était en premier le génome et par la suite les sujets concernant le transcriptome. Il était plus logique d'introduire le génome dans son entièreté et ensuite expliquer le transcriptome pour respecter la suite logique biologique. Afin de bien représenter le déroulement de ce projet, dans les prochains chapitres les sujets reliés aux transcriptomes seront présentés en premier. Les premières données sur lesquelles nous avons effectué des analyses sont celles du RNA-Seq. Les données génomiques sont arrivées en seconde étape pour ce projet.

3.1 Les patients

Les patients recrutés ont subi plusieurs investigations par des cliniciens et des généticiens, mais malgré tout ils n'ont pas reçu de diagnostic. Au niveau moléculaire, on suspecte que leur pathologie s'explique par des biomarqueurs génétiques complexes et rares nécessitant des analyses plus personnalisées à l'aide de la bio-informatique. Le recrutement peut être séparé en trois groupes différents décrivant la situation du patient. Le premier groupe est formé de patients ayant des tests génétiques qui ont identifié plusieurs VUS. Le second groupe concerne les patients où le mode de transmission suspecté est récessif, mais seulement des variants hétérozygotes (dans des gènes différents) ont été identifiés. Puis le troisième groupe concerne les patients où aucun variant ou gène candidat n'a été trouvé.

Dans le cadre de ce projet de maitrise, quatre patients ont été recrutés : Z26; BC01; BC02; HSJNM008. Une description des symptômes; âge; sexe des quatre patients à l'étude se trouvent dans le tableau IV ci-dessous. Pour la patiente Z26, nous avons également accès aux données exomes de son frère (sain). Puisque nous n'avons pas les données transcriptomiques de celui-ci, il sera utilisé pour exclure ou comparer les résultats de Z26 que dans les analyses comportant sur les données ES. Nous avons aussi accès aux données ARN d'un contrôle pédiatrique (un enfant avec biopsie musculaire normal) (HSJNM009) pour le patient HSJNM008. La méthode selon laquelle les données ont été filtrées avec les données familiales ou contrôles est décrite plus loin. Il est important de noter que le contrôle utilisé pour la majorité des analyses transcriptomiques de Z26; BC01; BC02 est l'individu sain B500. Pour ce qui concerne de l'ethnicité du contrôle B500, nous travaillons avec des biopsies musculaires et puis éthiquement c'est très difficile d'avoir accès à des biopsies de muscles sains, donc nous sommes très limités. Avoir une concordance pour l'ethnicité n'est pas toujours possible. Comme dans le cas du patient BC02 provenant de l'Asie du Sud nous n'aurons jamais (très peu probable) accès à du muscle sain de l'Asie du Sud. Dans la mesure du possible, nous allons comparer avec la même ethnicité. La figure 8 décrit l'ensemble des patients et données disponibles pour nos analyses.

Tableau IV. – Description du profil clinique de chaque patient à l'étude

Patient	Début	Sexe	Ethnicité	Description
Z26	Adulte	F	Européenne	Faiblesse musculaire, problèmes d'équilibre, contractions musculaires et crampes, rythme cardiaque anormal, épisodes d'hyperkaliémie, maladie pulmonaire restrictive et convulsions.
BC01	Pédiatrique	H	Européenne	Faiblesse musculaire, hypotonie, fatigue. La biopsie n'était pas informative. Parents pas atteints.
BC02	Pédiatrique	H	Asiatique du Sud	Épisode hémiparésie ressemblant à des accidents vasculaires cérébraux (stroke), rhabdomyolyse, IRM normale.
HSJNM008	Pédiatrique	H	Européenne	Faiblesse musculaire et encéphalopathie myo-neuro-gastro-intestinale (MNGIE).

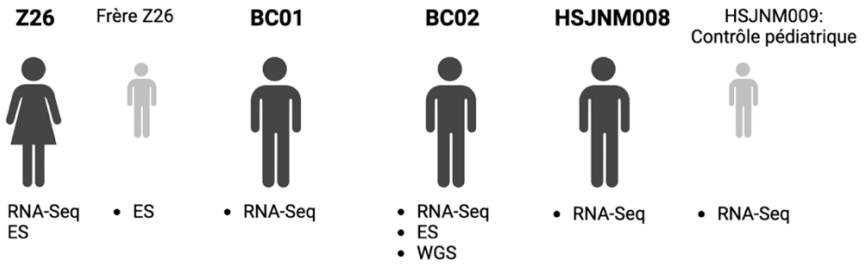

Figure 8. – Le type de données disponibles pour chaque patient.

Les individus en gris foncé représentent les quatre patients à l'étude. Les individus en gris pâle représentent les contrôles respectifs pour le patient, soit le frère de la patiente Z26 pour les données ES et le contrôle pédiatrique HSJNM009 pour le patient HSJNM008.

3.2 Flux de travail de l'approche bio-informatique

Initialement, les échantillons de muscle squelettique de chaque patient sont séquencés par RNA-Seq et, lorsque possible, par ES et WGS en utilisant la technologie d'Illumina. Il est important de noter que les données ES pour ces patients ont été réalisées dans un cadre clinique, une technique couramment utilisée pour le diagnostic de maladies mendéliennes rares (93). Par la suite, les différentes analyses sont effectuées sur les données. Au niveau de l'ARN, on procède à l'analyse de l'épissage alternatif et de l'expression différentielle. Au niveau de l'ADN, on procède

à l'identification des variants codants / non-codants , des expansions de répétitions et les CNVs. Pour intégrer les données ARN ADN: on procède à l'identification des débalancements alléliques et à la détermination de l'impact de variants introniques ou des VUS. Une représentation de l'approche multi-omique utilisée se retrouve sur la figure 9.

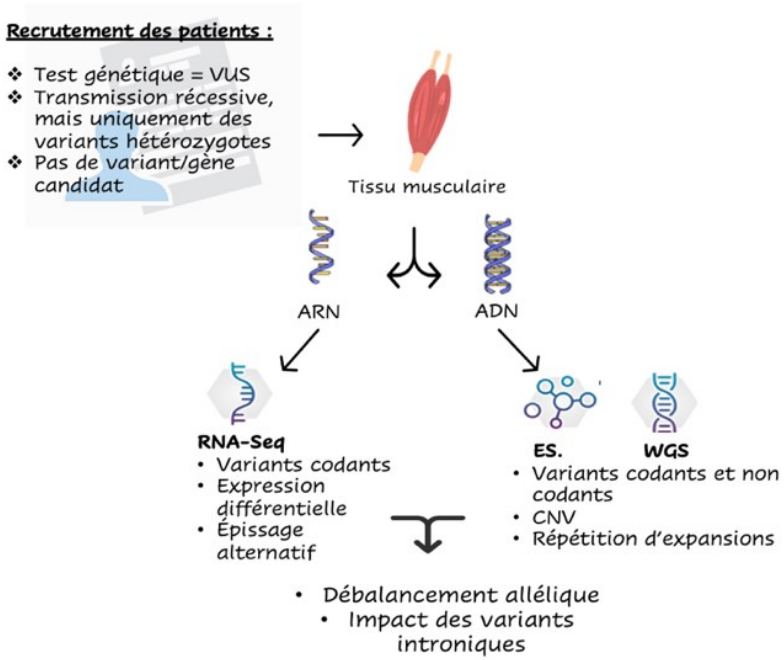

Figure 9. – L'approche multi-omique.

Dans les prochaines sous-sections de ce chapitre, une description du pipeline établit comprenant les différents outils bio-informatiques utilisés ainsi que les scripts maisons personnalisés pour cette approche proposée seront présentés (figure 10). De manière générale, initialement on roule l'alignement, l'annotation et tous les outils présentés dans le chapitre d'introduction (SpliceAI; rMATS; LPEseq; MToolBox; CNVkit; ExpansionHunter). Ceci génère plusieurs fichiers de résultats bruts. Pour chacun des fichiers sortis, des scripts en Python sont utilisés pour filtrer les variants, les jonctions ou les gènes qui n'étaient pas significatifs pour le patient (fréquence allélique; seuil d'épissage; niveaux d'inclusion; etc.). Ces fichiers filtrés sont utilisés pour créer des listes de variants prioritaires pour l'analyse. Comme expliqué précédemment, nous intégrons des variants transcriptomiques et génomiques afin de trouver des marqueurs qui provoquent un débalancement allélique. Nous associons des variants RNA-Seq à des jonctions d'épissage trouvées avec rMATS afin d'associer un variant à un site d'épissage. Nous

comparons ensuite toutes les données décrites avec les DEG. De plus, on vérifie avec CNVkit s'il y a des variants CNV et des expansions de répétition avec ExpansionHunter.

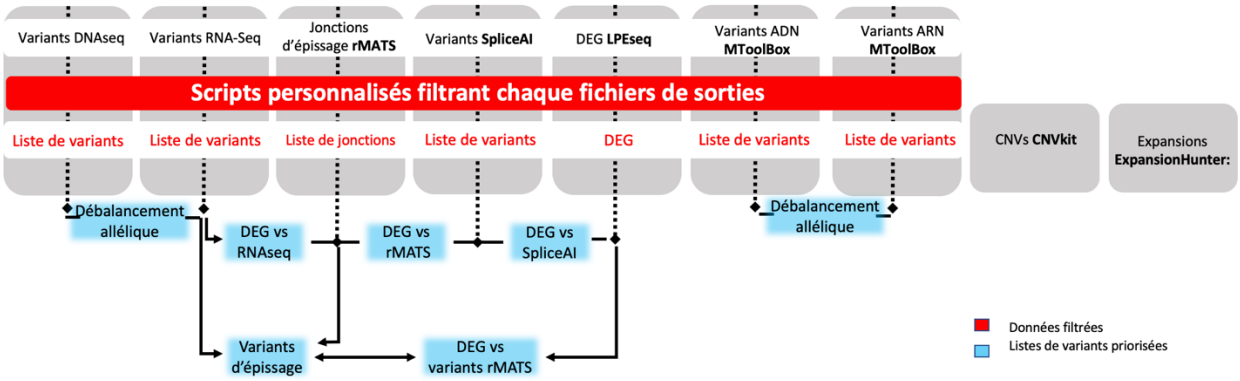

Figure 10. – Le pipeline bio-informatique proposé.

L'application des données de nos patients est également présentée dans ce chapitre. Il est important de noter que le pipeline bio-informatique pour l'alignement et l'annotation des séquences était déjà en place avant le début de ce projet. Les autres composantes du pipeline, soit ceux présentés dans la figure 10, ont été mises en place lors de ce projet.

3.2.1 Alignement et annotation des données de séquençage

Les analyses débutent avec les fichiers bruts FASTQ sortis du séquençage. Un contrôle de qualité est réalisé avant et après l'alignement utilisant l'outil FastQC (94). Cet outil vérifie plusieurs choses, dont la qualité des séquences WGS; ES; RNA-Seq; la présence de contaminants; d'adaptateurs; de lectures dupliquées et du contenu en GC (71, 95). Pour chaque type de séquençage, un outil différent est utilisé pour l'alignement. Pour les données génomiques, nous utilisons un script implémentant l'outil BWA (96). Quant aux données transcriptomiques, un script utilisant l'outil HISAT2 (97). Ces outils permettent l'alignement de l'ADN et de l'ARN sur le génome de référence hg19. L'assemblage et la reconstruction de la séquence génomique permettent d'associer les fragments alignés à un gène ou transcrit selon leur position dans le génome. Il est possible de mesurer l'expression des transcrits en comptant le nombre de fragments alignés qui se chevauchent sur le génome de référence. Un outil puissant utilisé dans ce projet pour quantifier est HTSeq (98). Les variants ont ensuite été appelés utilisant GATK

VarDict. Les fichiers sortis pour cette partie du pipeline sont un BAM (99) contenant les séquences alignées, d'autres fichiers contenant des informations pertinentes sur l'alignement effectué (QC, compte de lecture) et les fichiers VCF contenant les appels de variants.

La prochaine étape consistait à annoter les appels de variants dans le fichier VCF sortie par GATK en utilisant l'outil ANNOVAR (100). Les appels de variants peuvent être annotés avec différents scores de pathogénicité, fréquence allélique et plusieurs informations pertinentes pour l'analyse dans le but de prioriser les variants candidats (101). Parmi les nombreux scores produits par l'annotation, seuls quelques-uns sont très informatifs tels que : AF de gnomAD (102), GERP (103), PhastCons (104), SIFT (105) et CADD (106). Pour cette approche, puisque les patients recrutés présentent des formes plus rares de leurs troubles neuromusculaires, il est approprié de prioriser les variants ayant une AF plus rare, soit un AF inférieur à 0.02 (107). Par définition un variant rare à un AF inférieur à 0.01, nous avons opté pour une valeur moins stricte puisqu'il a déjà été démontré que certains variants pathogéniques dans le contexte des maladies rares avaient une fréquence allélique légèrement supérieure à 0.01 (108). Les variants ayant une fréquence plus élevée, soit un variant plus commun, ne seront pas priorisés pour la suite des analyses. GERP est un outil qui mesure le niveau de conservation du variant. Il adopte une méthode détectant les régions où il y a un manque de substitution permettant de mesurer l'effet de sélection négative par un score qui varie entre -12.3 (le moins conservé) et 6.17 (le plus conservé) (103). Basé sur notre expérience et des variants identifiés préalablement dans le laboratoire nous avons sélectionnés les filtres suivants. Un variant avec un score GERP supérieur ou égal à 5 est gardé. Le score PhastCons est également utilisé pour mesurer le niveau de conservations. L'algorithme utilise un modèle de Markov caché basé sur la phylogénétique prenant en compte le nombre et le type de substitution selon la longueur de la branche phylogénétique (104). Le score de probabilité varie entre 0 et 1, mais pour nos analyses ce score est mis sur une échelle de 1000, donc les variants avec un score plus grand ou égal à 400 sont gardés. Le score de SIFT varie entre 0 et 1, où un score inférieur à 0.05 est interprété comme étant une substitution délétère (109). L'algorithme de SIFT se base sur la séquence et assume qu'un acide aminé important pour la protéine sera conservé. Donc un changement à l'un de ces

acides aminés causant une différence dans les propriétés de celui-ci, sera prédit comme étant délétère, affectant ainsi la fonction de la protéine (105, 110). Il est important de noter que le score SIFT est inversé dans nos résultats puis le filtre choisi n'est pas strict, les variants avec un score SIFT supérieur ou égal à 0.5 sont gardés. Dans un contexte exploratoire, nous trouvons qu'un seuil de 0.95 (dans notre cas 0.05) pour SIFT était trop strict pour nos données, nous voulions être certaines de ne pas exclure des variants candidats. Aussi, lorsque nous regardons l'effet de ce score sur des variants pathogéniques prouvés dans notre laboratoire, on perdait trop d'évènements candidats. Ensuite, le dernier score utilisé pour filtrer et prioriser les variants annotés est CADD. Il prédit ainsi l'effet délétère des SNV et des indels en intégrant plusieurs annotations, un score supérieur à 10 ou 20 indique que le variant est dans les 10% ou 1% des substitutions les plus délétères respectivement (106). Dans nos analyses, un seuil d'au moins 15 pour le score de CADD, est utilisé. CADD combine l'information des outils qui analyse les SNV faux-sens ainsi que les effets de conservation de la séquence. À partir de ses scores, la liste des gènes candidats pour chaque patient peut être filtrée afin d'analyser les variants potentiellement pathogéniques identifiés par les outils bio-informatiques discutés à date.

<u>En résumé:</u>

$$AF \leq 0.02$$

$$GERP \geq 5$$

$$PhastCons \geq 400$$

$$SIFT \geq 0.5$$

$$CADD \geq 15$$

Il est important de noter que les variants candidats identifiés par les algorithmes d'annotation seront utilisés pour analyser de manière plus rigoureuse les résultats obtenus par les explorations décrites dans les prochaines sections. Pour la majorité des outils utilisés, nous utilisons en entrée une liste de variants obtenue par l'alignement. Ensuite, dépendant de l'outil, les variants sortis sont manuellement filtrés pour garder les variants rares, soit une AF de moins de 2%. La raison pour ceci est que la liste de variants candidats générés par l'annotation est de

plus petite taille. Nous optons pour une approche moins stricte pour l'utilisation des outils afin de ne pas manquer un évènement non priorisé par les algorithmes d'ANNOVAR.

3.2.2 Exploration des données du RNA-Seq, WGS et ES

Grâce aux données générées par le séquençage de RNA-Seq, d'ES et WGS ainsi que les étapes d'alignement discutées précédemment, il sera possible d'identifier des variants pathogéniques et des évènements de variation structurale du génome et transcriptome. Avec les données du transcriptome, il est possible d'analyser l'épissage alternatif et les gènes différentiellement exprimés (DEG). Autrement, les CNVs et les expansions de répétitions sont analysés avec les données génomiques. Les variants mitochondriaux sont analysés tant chez le transcriptome que chez le génome.

3.2.2.1 Épissage alternatif et DEG

Pour analyser l'épissage aberrant, l'outil SpliceAI version 1.2.1 est utilisé pour les données ARN et ADN. Cet outil utilise en entrée une liste de variants, soit le fichier VCF extrait de l'alignement. Une fois l'outil lancé pour chaque patient, nous obtenons un VCF de variants annotés par les scores SpliceAI. Les scores sont ensuite filtrés avec un script personnalisé, écrit en Python 3.7. Le filtre garde les variants avec une AF inférieure ou égale à 0.02 et les variants ayant un des quatre scores de SpliceAI supérieure à l'un des scores suivants : 0; 0.2; 0.5; 0.8. Le score de 0.8 est le plus sévère et n'est pas utilisé pour nos analyses.

Par la suite, nous lançons rMATS 4.1.0 qui utilise soit un fichier FASTQ ou BAM en entrée. L'outil génère une série de fichiers .txt pour chaque type d'épissage (A3SS ; A5SS ; MXE ; RI ; SE) contenant les jonctions exon-intron d'épissage. Dans les fichiers de sorties, on obtient de l'information sur la position du début et de la fin de l'exon. Les fichiers de sorties sont filtrés avec un autre script écrit en Python. Le score de différence de niveau d'inclusion (IncLevelDiff) est filtré selon le type de transmission suspectée pour le patient. Pour une transmission dominante, récessive ou inconnue, on filtre pour un score entre 30-60%, 40-100% ou 10-100% respectivement. Le compte de lectures d'intron et saut d'exon d'un des deux échantillons (contrôle ou cas) doit être supérieur ou égal à 10. Avec les fichiers filtrés, il est possible d'effectuer d'autres explorations qui seront discutées plus loin. De plus, la visualisation et la validation de ces

événements d'épissage anormaux sont faites sur le visualiseur de génomique intégrative (IGV) version 2.8.2 en comparant toujours avec un patient contrôle pour différencier les évènements anormaux. Cet outil permet l'exploration et la visualisation des données de séquençage aligné (soit le fichier BAM) de nos patients. Il se sert d'un génome de référence (hg19 pour ce projet) afin de vérifier la présence ou non d'un changement de nucléotide. Il permet également de visualiser la couverture du nombre de lectures et des jonctions d'épissage retrouvées dans le génome.

Pour analyser le niveau d'expression des gènes, cela se fait à l'aide d'un script, écrit en R version 3.6.0, qui utilise l'outil LPEseq. LPEseq est similaire à DESeq, mais convient à l'analyse sans répliques biologiques comme expliqué précédemment. Un changement d'expression significatif pour nos analyses signifie qu'une valeur du L2FC pour le gène est plus grande que la valeur absolue de 2 et une valeur P plus petite que 0.1 (69, 111). Pour le choix du seuil de LPEseq, les auteurs suggèrent de se fier à la valeur P, il existe des possibilités fréquentes d'utiliser de 0.01 à 0.1 (0.1 est plus libéral). La liste des DEG permettra la recherche des variants rares avec un impact fonctionnel inconnu pour associer un variant à un changement d'expression significatif. Cette intégration sera décrite dans la sous-section suivante.

3.2.2.2 Les CNVs et expansion de répétition

L'outil CNVkit prend comme entré le fichier BAM obtenu par l'étape d'alignement décrit précédemment. Il génère plusieurs fichiers, dont un VCF, .cns et .cnr. Pour nos analyses présentées, seuls les résultats contenus dans le VCF seront discutés. Les données sont filtrées selon les seuils de nombre de copies décrit dans le premier chapitre, tableau I. Aussi, il y a deux options de lignes de commandes additionnelles disponibles pour cet outil, dont le "Call" et "Scatter", afin de visualiser à l'aide d'une figure l'allure des chromosomes du patient. La commande "Scatter" prend en entrée une combinaison des .cns et .cnr produit par l'appel des CNV (commande "Call"). Il est possible de spécifier à l'aide des options -g -c pour visualiser un gène ou un chromosome en particulier. Aucune figure ne sera présentée dans ce mémoire pour les résultats de la commande "Scatter".

Afin d'étudier les expansions de répétition pour le patient BC02, l'outil ExpansionHunter a été utilisé. L'outil prend en entrée les lectures du fichier BAM générées par l'alignement, ainsi que les fichiers de référence et variant de catalogue (loci ciblé). L'outil génère un fichier VCF dans lequel les informations sur le nombre de répétitions estimées sont disponibles à l'aide des scores de STR. Pour nos interprétations une analyse de la littérature permet de déterminer si le STR est sain, porteur de prémutation ou pathogénique. Ceci est possible de le faire manuellement en raison du catalogue de petite taille de ExpansionHunter, soit une vingtaine de régions analysées seulement.

3.2.2.3 Exploration des données intégrées

Les divers outils ont permis de récolter un maximum d'informations sur les variations structurelles, les DEGs, l'épissage et SNV. Les fichiers de sorties ont été filtrés avec les seuils décrits dans les sous-sections précédentes. Il est désormais possible d'intégrer toutes ces informations pertinentes afin de prioriser certains événements trouvés dans les fichiers de sortie et réduire considérablement la taille des listes de résultats pour chaque patient. Nous effectuons l'intégration de la liste des variants VCF avec les variants d'épissage donnés par rMATS pour associer un variant à une jonction d'épissage pour les patients. Ensuite, les variants sont intégrés à l'expression génique à l'aide d'un script personnalisé. Nous travaillons avec des données ADN et ARN pour en savoir plus sur les VUS et prioriser ces gènes lors de la recherche des variants pathogéniques. Pour faire ceci, nous utilisons les listes de variants du transcriptome et du génome pour appliquer un filtre écrit en Python qui recherche les débalancements alléliques. Si un variant est homozygote ou possiblement homozygote dans la liste des variants à partir des données ARN et se trouve hétérozygote ou hétérozygote multiple dans la liste des données génomiques et de l'exome (ADN), c'est un candidat potentiel pour un débalancement allélique.

Pour analyser le génome mitochondrial de ces patients, nous allons utiliser MToolBox. Un simple rappel: "les sites dont la valeur de variabilité est proche de 1 présentent un allèle dont l'état est commun à tout un sous-arbre de l'arbre entier représentatif des génomes stockés dans HmtDB ou pourraient être soumis à des mutations récurrentes". De faibles valeurs de variabilité peuvent suggérer un nouveau variant définissant l'haplogroupe ou indiquer un score de mutation

lié à une maladie rare MitVarProt. C'est pourquoi nous avons filtré en utilisant l'estimation de la variabilité du site des nucléotides et des acides aminés via les programmes SiteVar et MitVarProt, soit Variabilité Nt ≤ 0.2 et Aa Variabilité ≤ 0.2.

3.3 Application du pipeline sur les patients

Dans le chapitre d'introduction, nous avons eu une mise en contexte plus clinique de mon projet et une mise en contexte plutôt technique et bio-informatique. Dans le chapitre deux, nous avons exploré les différents avantages et inconvénients de certaines approches d'analyses de données de patient, motivant ainsi ce projet qui intègre les différentes omiques afin d'avoir une approche la plus optimale possible. Au début de ce troisième chapitre, on a décrit plutôt la méthode utilisée pour notre approche, soit les outils bio-informatiques et les scripts filtrant des listes de données de grande taille. Pour la suite de ce chapitre, l'application du pipeline et les résultats sortis sont présentés pour chaque patient. Les résultats pour chaque patient ne sont pas reliés entre eux. Comme expliqué précédemment, ils sont quatre individus non apparentés; pas du même âge ou sexe et présentent tous des formes différentes de maladies neuromusculaires (rappel : le terme neuromusculaire regroupe beaucoup de maladies différentes).

Le coût de certaines expériences ainsi que l'accès aux données peuvent être des facteurs limitants. Idéalement, comme mentionné dans les chapitres précédents, l'accès aux données du génome et transcriptome complet des patients à l'étude dans ce projet auraient permis une analyse en profondeur et complète. Pour les patients Z26 et BC02 les résultats présentés sont plus complets représentant bien le type d'exploration effectuée par notre approche proposée analysant ainsi les données transcriptomiques et génomiques. Pour les deux autres patients, seulement les résultats du RNA-Seq sont présentés.

3.3.1 Les variants candidats identifiés par l'annotation

De nombreux variants existent dans le contenu génétique d'un individu. Dépendant de son emplacement dans le génome et de son impact au niveau fonctionnel, celui-ci peut causer un éventail de problème pour la personne porteuse. Ces variants peuvent être rares, donc difficiles à identifier avec des approches standards. Grâce aux nombreux algorithmes et outils bio-

informatiques, il est possible d'identifier et de prioriser certains variants comme étant potentiellement pathogénique. L'étape d'annotation permet ainsi ce genre de priorisation. À partir des VCF extraits de l'alignement de séquence, il a été possible d'identifier certains variants potentiellement pathogéniques au niveau du transcriptome pour chacun des quatre patients. Ceci a également été possible au niveau des données ES de Z26 et BC02 ainsi que les données WGS de BC02. L'annotation par ANNOVAR génère un rapport des variants identifiés comme étant les meilleurs candidats selon les divers algorithmes utilisés par les outils. Comme expliqué, nous avons filtré ces rapports selon une AF rare (inférieure ou égale à 0.02) et par la suite filtré les variants rares selon les quatre scores de pathogénicité étudiés : GERP; PhastCons; SIFT; CADD. Les variants contiennent ceux provenant de GATK et VarDict pour cette partie de l'analyse.

Dans le cas de maladie plus commune, il est plutôt commun de mettre en place un seuil de couverture génomique minimale lors de l'appel de SNP par RNA-Seq. Dans le contexte de maladie rare, nous avons pris la décision de ne pas utiliser de filtre sur la couverture, car on vérifie manuellement par la suite utilisant soit le séquençage de Sanger ou le LRS. Comme nous sommes en mode exploratoire, nous ne voulons pas éliminer quelque chose qui pourrait être un potentiel candidat. Par exemple, s'il y a une dégradation d'un allèle, on s'attend à ce que la couverture soit plus basse chez notre patient. De plus, il y a certains gènes de syndrome myopathique qui peuvent avoir une expression relativement faible dans le muscle et on ne veut pas éliminer ces gènes d'intérêts. Pour ces raisons, on n'utilise pas de filtre sur l'expression des gènes pour faire de l'appel de SNP.

L'annotation des variants transcriptomiques a été réalisée chez les quatre patients à l'étude, soit Z26, BC01, BC02 et HSJNM008. Pour chaque patient, une liste est générée pour GATK et VarDict. On observe sur la figure 11.A) que pour la majorité, sauf HSJNM008, le nombre de variants bruts de GATK se retrouvant dans cette liste est supérieur à celle de VarDict. On observe également que ces rapports sont déjà préalablement filtrés par le pipeline du laboratoire pour des variants rares, donc pour nos patients très peu de variants ont été exclus à cet égard. Le pipeline mentionné concerne celui mis en place dans le laboratoire préalablement à ce projet qui

gère l'alignement, l'appel de variants et l'annotation des données brutes expliqué précédemment (109). On remarque également que le nombre de variants rares et filtrés est une combinaison des variants uniques à GATK; uniques à VarDict; et en communs. En prenant Z26 comme exemple, on retrouve 60 et 33 variants filtrés de GATK et VarDict respectivement, 26 parmi ceux-ci sont en commun et 42 variants uniquement trouvés chez un seul (35 GATK et 8 VarDict) pour un total de 68 variants combinés. Dans la figure 11.B), nous avons une représentation en diagramme de Venn du nombre de variants potentiellement pathogéniques pour chaque patient. Il est important de noter que pour générer cette figure seulement les variants uniques à VarDict ont été ajoutés à la liste préfiltrée afin de les inclure dans le compte des variants combinés. Aussi, les figures 11 à 13 servent de support visuel pour représenter la distribution des variants en fonctions des différents scores pathogéniques obtenus lors de l'annotation. Ils ont servi de méthode de priorisation et non pour éliminer les variants de nos analyses complètement. Avant de procéder avec le pipeline d'analyse présenté dans ce mémoire, une inspection manuelle de tous les variants bruts sortis a été faite initialement. Par exemple, on observe parmi les gènes ressortis si on n'identifie pas un gène neuromusculaire déjà connu; on observe un filtre à la fois (AF; GERP; CADD; SIFT; phatsCons) quel variant ressort; on observe s'il y a deux variants sur le même gène; on observe les variants homozygotes; etc. Nous prenons en note nos observations, mais il est rare que nous allions identifier un seul variant cible où nous sommes certains qu'il est pathogénique. Comme expliqué, c'est la raison pour laquelle nous devions développer un pipeline diversifié, analysant plusieurs avenues possibles, permettant une analyse plus complète et sophistiquée.

A)

PATIENT	Variants GATK			Variants VarDict			Variants rares filtrés combinés *
	Brutes	Rares	Filtrés	Brutes	Rares	Filtrés	
Z26	588	585	60	527	522	33	**68** (35+8+26)
BC01	190	186	29	146	143	16	**31** (15+2+14)
BC02	350	347	50	266	264	24	**53** (29+3+21)
HSJNM008	284	280	43	451	439	42	**47** (23+4+20)

B) * Nombre total (variants uniques à GATK + variants uniques à VarDict + variants en communs)

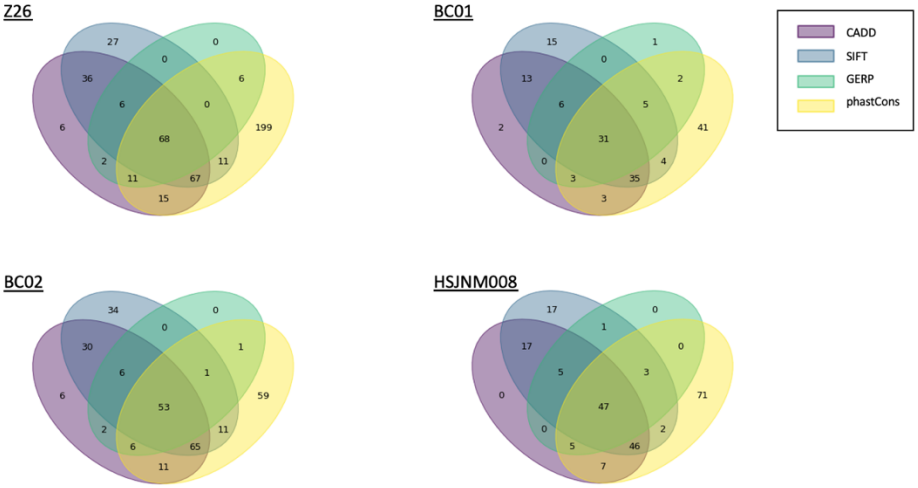

Figure 11. – Diagrammes de Venn des variants RNA-Seq priorisés de l'annotation.

On observe 68; 31; 53; 47 variants rares potentiellement pathogéniques chez Z26; BC01; BC02; HSJNM008 respectivement. Ces variants sont trouvés par les filtres décrits dans le chapitre précédant des scores de GERP; PhastCons; SIFT; CADD.

Une fois que les variants potentiellement pathogéniques sont ciblés du côté du transcriptome, il est intéressant de regarder maintenant au niveau de l'exome si les mêmes variants sont présents, lorsque possible (figure 12). Aussi cela permet d'identifier des variants candidats trouvés au niveau du WGS et ES pour les analyses génomiques. Les mêmes seuils que pour les variants transcriptomique sont appliqués sur les données exome de Z26 et BC02. Comme mentionné précédemment, les variants GATK et VarDict sont combinés pour cette analyse afin de faire une vérification des variants ciblés et s'assurer de ne rien manquer.

On observe sur la figure 12.A) que le nombre de variants bruts pour GATK et VarDict est plus élevé que ceux identifiés au niveau du transcriptome. Pour Z26, on a identifié 53 et 50 variants rares filtrés pour les aligneurs GATK et VarDict, respectivement. Les 50 variants identifiés par VarDict étaient tous inclus dans ceux de GATK, donc aucun variant n'a été ajouté à la liste filtrée de GATK. On a également identifié que 23 des 53 variants sont retrouvés au niveau du ES et RNA-Seq (Annexe A). Avant l'exclusion par les variants communs (soit même variant même génotype) du frère sain de Z26, on retrouvait 77 variants rares filtrés. Donc 24 variants ont été exclus de la liste de variants exomes ciblés de Z26, laissant 53 variants uniques à Z26. Pour le patient BC02, GATK a permis d'identifier 52 variants rares passant tous les seuils décrits dans le chapitre précédent. Autrement, nos filtres ont permis d'identifier 50 variants avec VarDict. Tous les variants trouvés avec VarDict sont inclus dans ceux de GATK, donc le nombre de variants rares et filtrés combinés est de 52. Ceci pourrait s'expliquer par le fait que GATK génère plus de faux positif au niveau de l'ARN. Donc au niveau de l'ADN on s'attend à retrouver un nombre égal (ou très proche) entre GATK et VarDict contrairement à ce qui est observé avec l'ARN. On retrouve 24 de ces variants dans les variants transcriptomiques ciblés pour le patient BC02 (Annexe B). Sur la figure 12.B) une représentation en diagramme de Venn est également produite pour les analyses des données exomes.

A)

PATIENT	Variants GATK			Variants VarDict			Variants rares filtrés combinés *	Nombre de variants communs AND-ARN
	Brutes	Rares	Filtrés	Brutes	Rares	Filtrés		
Z26	428	443	53	449	443	50	**53** (3+0+50)	23
BC02	654	619	52	671	646	50	**52** (2+0+50)	24

*Nombre total (variants uniques à GATK + variants uniques à VarDict + variants en communs)

B)

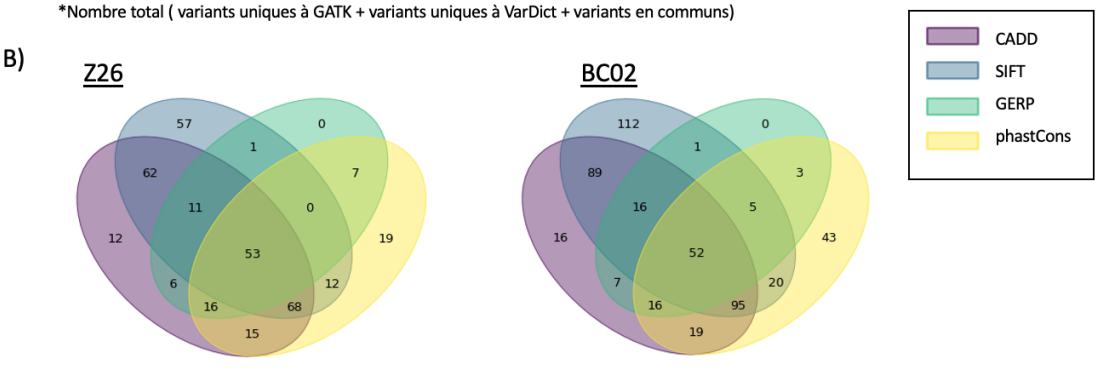

Figure 12. – Diagrammes de Venn des variants ES des patients Z26 et BC02 priorisés de l'annotation

Initialement, pour le patient BC02 nous avions les données transcriptomiques. Il y a plusieurs évènements pathogéniques qui peuvent être révélés en utilisant ce type de données. Comme notre hypothèse stipule, intégrer les données génomiques permet d'investiguer plus en profondeur. Les données exomes de ce patient sont également disponibles comme présenté ultérieurement, mais le fait d'avoir la possibilité d'explorer au niveau du génome complet augmente les chances d'identifier un évènement pathogénique. Par contre, avoir accès à ces données est sans doute plus dispendieux. La méthode pour la figure 13 est la même que celle présentée dans les deux dernières figures. On observe 50 variants filtrés qui sont rares, et ce pour GATK et VarDict. Aucun variant était unique à un seul aligneur. Le nombre de variants rares filtrés communs entre ADN et ARN ne change pas entre les données ES et WGS pour BC02.

A)

PATIENT	Variants GATK			Variants VarDict			Variants rares filtrés combinés *	Nombre de variants communs AND-ARN
	Brutes	Rares	Filtrés	Brutes	Rares	Filtrés		
BC02	654	636	50	714	678	50	**50** (0+0+50)	24

* Nombre total (variants uniques à GATK + variants uniques à VarDict + variants en communs)

B)

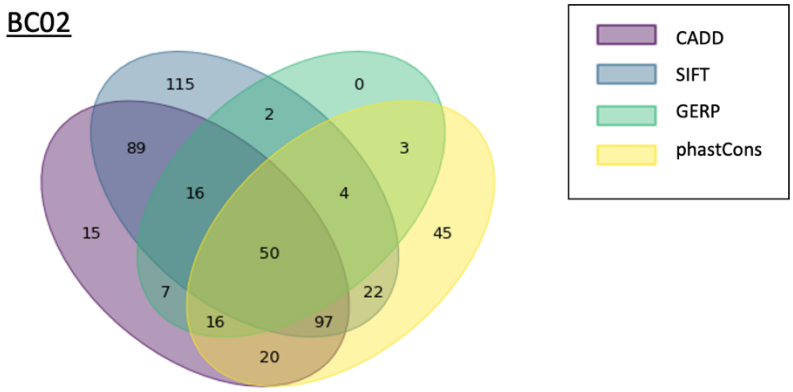

Figure 13. – Diagrammes de Venn des variants WGS du patient BC02 priorisés de l'annotation

3.3.2 Données du pipeline de nos patients

Une fois l'étape d'alignement et d'annotation est complétée on peut maintenant appliquer le pipeline proposé sur les données disponibles pour nos quatre patients. Ceci nous permettra de regarder pour chaque étape du pipeline les évènements les plus intéressants. Le tableau V suivant présente un résumé des résultats obtenus pour chaque outil et script maison. Une description de ce qui est présenté se fera par la suite.

Tableau V. – Résultats sortis du pipeline pour nos quatre patients

		Patients							
		Z26		BC01		BC02		HSJNM008	
		BRUTES	FILTRÉS	BRUTES	FILTRÉS	BRUTES	FILTRÉS	BRUTES	FILTRÉS
SpliceAI	AF ≤ 0.02		818		440		710		596
	DS ≥ 0		751		378		630		535
	DS ≥ 0.2	27454	21	17632	5	24146	11	23131	7
	DS ≥ 0.5		9		1		5		2
Jonctions rMATS	A3SS	4048	20	3616	15	4183	1	3831	1
	A5SS	2487	17	2303	3	2694	4	2485	7
	MXE	2668	19	2729	9	3447	22	3180	5
	RI	3475	29	3137	29	3401	1	3008	16
	SE	26304	181	25646	56	32040	30	29141	10
	=	38982	266	37431	112	45765	58	41645	39
DEG			3447		5251		4582		2909
DEG vs SpliceAI (DS ≥ 0)		82		64		102		70	
DEG vs variants RNA-Seq		131		87		151		95	
DEG vs rMATs		28		10		11		5	
Débalancement allélique		19				5			
MToolBox ARN		71	20	73	20	80	23	60	15
MToolBox ADN		55	12			73	23		
Débalancement allélique mitochondrial		0				0			

L'épissage alternatif est un évènement essentiel au bon fonctionnement de la cellule. Celui-ci est responsable de la diversité des gènes et contrôle lesquels seront transcrits pour former le brin d'ARNm mature. Dans le cas de nos patients, l'accès aux tissus de muscles affectés permet d'observer des anomalies d'épissage cryptique potentiel. Donc, il sera possible d'observer l'épissage qui a eu lieu chez un patient et possiblement identifier des évènements associés à un changement au niveau de l'expression du gène. Le premier outil pour ce genre d'analyse est SpliceAI, celui-ci prédit l'effet des variants identifiés par l'alignement sur l'épissage. Comme mentionné, les résultats obtenus par l'outil ont été filtrés selon leur score DS. Pour chaque patient, le nombre brut correspond aux variants du VCF extrait de l'alignement. On observe sur le tableau V qu'une grande majorité des variants sont exclus de nos analyses, puisqu'ils n'ont pas passé le seuil de AF. Parmi les variants rares de Z26; BC01; BC02; HSJNM008; on observe que 67; 62; 80; 61 variants n'avaient pas au minimum un des quatre scores plus grand ou égal à zéro, respectivement. Il est important de noter que ces variants filtrés pour ce score seront utilisés pour le restant des analyses du pipeline en raison de la taille de la liste de variants sortie. Pour Z26; BC01; BC02; HSJNM008 on retrouve que 2.8%; 1.32%; 1.75%; 1.3% des variants SpliceAI ($DS \geq 0$) ont au moins un des quatre scores DS plus grand ou égal à 0.2, respectivement. Puis pour les variants avec un des quatre scores plus grands que 0.5, on retrouve 1.2% 0.27%; 0.79% 0.37% respectivement.

Pour analyser plutôt les jonctions d'épissage alternatif différentiel, l'outil rMATS a été utilisé. Les fichiers de sorties contiennent des jonctions exon-intron épissées. L'outil produit 38992; 37421; 45765; 41646 jonctions brutes pour Z26; BC01; BC02; HSJNM008. On observe (tableau V) la rigourosité de notre filtre pour une transmission récessive sur les données filtrées de chaque patient. Pour chaque patient, il y a cinq catégories représentées marquant les cinq types d'épissage analysé par l'outil, soit les évènements de A3SS, A5SS, MXE, RI et SE. La catégorie qu'on retrouve en plus grande quantité pour cette analyse est celle des sauts d'exons (SE). La catégorie en minorité est l'épissage A5SS.

Lorsqu'un système est confronté à un changement anormal, il peut affecter une quantité énorme de facteurs. Un décalage dans une séquence codante peut, par exemple, modifier le niveau d'expression d'un gène par un dysfonctionnement d'un processus. Les données transcriptomiques permettent de faire l'analyse de l'expression des gènes. D'ailleurs, nous pouvons examiner ceci directement dans les tissus affectés du patient. Comme expliqué précédemment, aucune réplique n'est disponible donc nous sommes conscients qu'il y a une diminution du pouvoir statistique pour ce type d'analyse dans ce projet. Néanmoins, on utilise LPEseq qui est plus approprié pour explorer les résultats générés. Dans le tableau V, nous avons le résumé du nombre de gènes différentiellement exprimés pour chacun des quatre patients. Ces gènes sont utilisés pour plusieurs autres analyses, donc il ne fallait pas être trop strict, comme nous pouvons l'observer.

Les variants mitochondriaux identifiés au niveau transcriptomique et génomique ont été analysés avec l'outil MToolBox (tableau V). On obtient avec les données ARN 20; 20; 23; 15 variants filtrés pour Z26; BC01; BC02; HSJNM008 respectivement. Avec les données ES, il a été possible d'identifier 12 et 23 variants mitochondriaux pour Z26 et BC02. Les débalancements ont également été vérifiés pour ces données, mais aucun n'a été identifié. Pour la patiente Z26, aucun des variants mitochondriaux filtrés identifiés n'a été conservé lors de l'exclusion des données en commun avec son frère.

La présence d'un débalancement allélique peut parfois causer une maladie. L'accès aux données transcriptomiques et génomiques permet l'identification des ces évènements. Par la suite, il s'agit de déterminer leur impact au niveau fonctionnel. Dans notre cas, un débalancement allélique chez un variant candidat serait intéressant. Il est évident que ce type d'investigation est possible seulement avec les patients dont les deux types de données sont disponibles. Pour la patiente Z26 présentée dans le tableau VI ci-dessous, on observe 19 débalancements alléliques. Pour le patient BC02, un plus petit nombre de débalancements alléliques est reporté, soit cinq (tableau VII). Les évènements ont été vérifiés et validés sur IGV. Ces listes de variants seront utilisées pour prioriser les variants choisis comme potentiellement pathogéniques pour les

patients. Il est important de rappeler encore une fois que nous sommes dans un contexte exploratoire. Parmi les débalancements alléliques présentés, on retrouve parfois une couverture plus faible (exemple couverture de 10X) au niveau de l'ARN. Afin de ne pas exclure des évènements potentiels de débalancement allélique, nous avons décidé de les inclure dans nos analyses (112). Une validation manuelle, soit par séquençage de Sanger ou par LRS devra être effectuée pour chacun de ces évènements ressortis. Dans le cadre de ce projet de maitrise, il n'a pas encore été possible d'accomplir cette étape.

Tableau VI. – Les débalancements alléliques de la patiente Z26

Variants	*REF*	*ALT*	*Gènes*
chr1:112323335	G	A	KCND3
chr1:217793225	C	T	GPATCH2
chr2:48808302	G	A	STON1-GTF2A1L
chr2:175939223	TA	T	ATF2
chr2:219677646	G	A	CYP27A1
chr4:5021161	A	C	CYTL1
chr4:56758797	G	T	EXOC1
chr4:159750363	C	T	FNIP2
chr5:76028758	T	C	F2R
chr6:147655318	G	A	STXBP5
chr6:154771250	A	G	CNKSR3
chr7:73083860	A	G	VPS37D
chr7:75104031	C	T	POM121C
chr16:5134590	C	T	EEF2KMT
chr17:37783480	C	A	PPP1R1B
chr19:7810674	G	A	CD209
chr20:44486301	G	C	ZSWIM3
chr21:46952064	T	C	SLC19A1
chrX:122805523	T	A	THOC2

Tableau VII. – Les débalancements alléliques du patient BC02

Variants	*REF*	*ALT*	*Gènes*
chr1:20067395	C	T	TMCO4
chr5:176930358	G	C	DOK3
chr19:20045052	T	C	ZNF93
chr19:52869706	C	A	ZNF610
chr21:43259753	C	T	PRDM15

L'accès aux données génomiques (ES et WGS) permet d'analyser si le patient à l'étude possède des variants causant des délétions ou insertions d'au moins 1kb dans la séquence nucléotidique. Comme expliqué dans les chapitres précédant, ces évènements peuvent être dévastateurs chez une personne causant ainsi leur pathologie. Dans la figure 14 ci-dessous, on retrouve deux graphiques en beignet décrivant le nombre de variants CNVs identifiés par l'outil CNVkit. Selon les seuils utilisés pour cet outil, il a été possible de classer les variants sortis par l'outil. Il est important de rappeler que cette analyse est mieux réalisée lorsque les données WGS sont utilisées. Donc, deux types de données (ES et WGS) sont présentés pour cette analyse pour le patient BC02. Il n'y avait aucun évènement de CNVkit qui était commun entre Z26 et son frère.

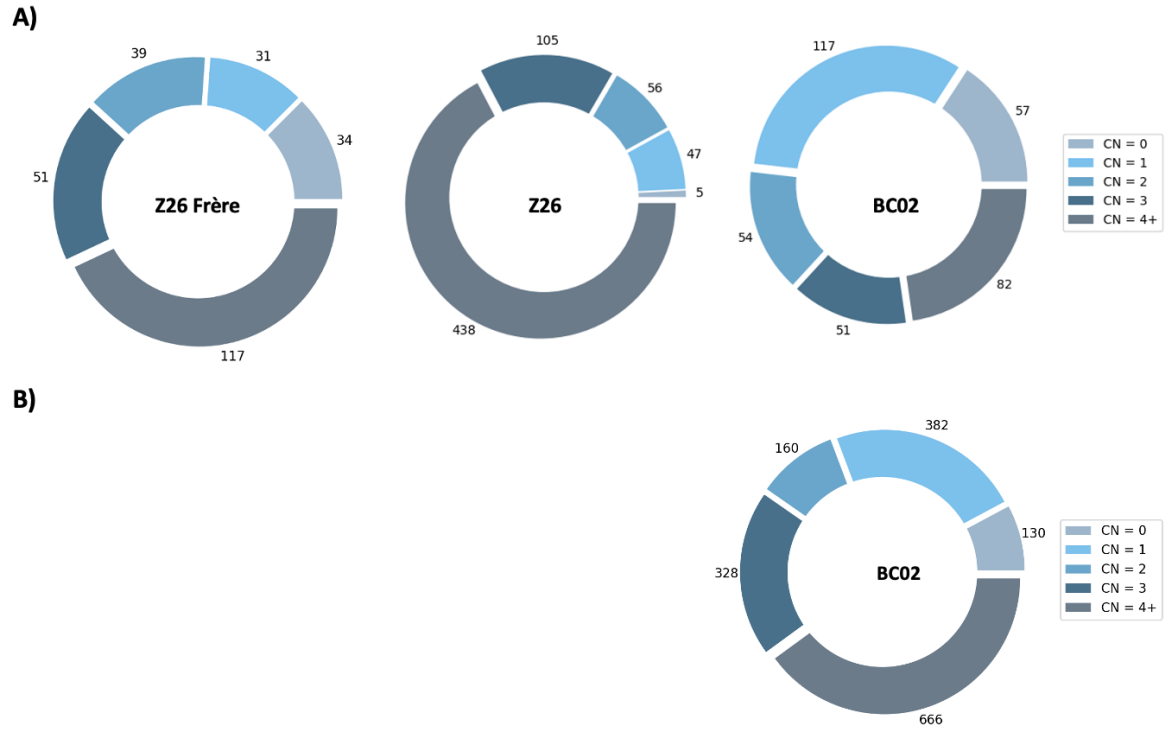

Figure 14. – Les CNVs identifiés chez le frère de Z26, Z26 et BC02 avec l'outil CNVkit.

Dans cette figure on observe en A) les différents variants CNVs ES pour le frère de Z26, la patiente Z26 et le patient BC02 décrit par les différents tons de bleus et en B) en utilisant les données WGS de BC02. Les délétions; les pertes; les gains; AMP sont représentés par CN=0; CN=1; CN=3; CN=4+ respectivement. Le sens associé aux couleurs de la légende et la figure vont dans le sens antihoraire.

Pour la patiente Z26, on observe sur la figure une faible présence de variants causant une délétion (CN = 0). À l'inverse, une très grande proportion de ces variants se trouvent à être des amplifications multicopie (AMP: "multi-copy amplification" en anglais). On observe 47 et 105 variants de perte et gain, respectivement. Le nombre de variants identifié comme étant neutre, soit une copie sur chaque allèle normal, est de 56. On observe une distribution des CNVs similaires avec les données ES du frère de Z26. Pour le patient BC02, au niveau de l'exome on observe une distribution plus égale des variants CNVs classifiés. Les pertes, soit CN=1, sont majoritaires. On retrouve 57;117; 51; 82 variants de délétion; perte; gain; AMP; respectivement. De plus, il y a 56 variants neutres ne causant pas de perte ou gain (CN = 2). Au niveau des CNVs trouvés avec les données WGS de BC02, on retrouve 130; 382; 160; 328; 666 variants de délétion; perte; neutre; gain; AMP; respectivement. Parmi les quatre diagrammes présentés dans la figure, on observe que Z26 possèdent une distribution fortement dominante de CN=4. Ceci ne semble pas être un effet isolé de l'utilisation de données ES puisqu'une distribution similaire se retrouve chez le frère et chez BC02 avec les données WGS. Il est difficile pour le moment de poser une hypothèse concernant ces résultats, il y a intérêt d'explorer plus loin les réarrangements chez la patiente Z26.

Pour chacun des variants candidats ciblés par les scores d'annotation, nous avons regardé si un CNV était relié à ce même gène. Les tableaux VIII et IX décrivent ces évènements retrouvés pour la patiente Z26 et BC02. Pour le tableau IX, nous avons combiné les résultats ES et WGS de l'analyse CNVkit. Alors, les variants candidats priorisés de l'exome sont comparés avec les possibles CNVs identifiés avec les données exomes, puis la même logique est utilisée pour les données WGS du patient BC02. Il est important de noter qu'aucun des débalancements alléliques identifiés précédemment n'est associé avec un CNV obtenu de nos analyses CNVkit, soit pour Z26 ou BC02.

Tableau VIII. – Les variants candidats ES impliqués dans des CNVs pour la patiente Z26

Gènes candidats	*Variants*	*Type*	*Longueur*	*L2FC*
LMOD1	chr1:201857682	DUP	112636	0.8
TTN	chr2:179496891	DUP	10341713	1.1
	chr2:179398813	DUP	7498	1.5
	chr2:179390758	DUP	8055	6.7
CAMK2A	chr5:149509275	DUP	93497	3.2
	chr5:149602772	DEL	-79109	-0.5
MAPK8IP3	chr16:60223	DUP	1747446	0.6
	chr16:1808148	DUP	497794	1.6
KIAA0100	chr17:18166570	DUP	8774955	0.6
	chr17:26941525	DUP	109586	2.7
PLD3	chr19:40871574	DUP	12817	3.2
	chr19:40884997	DEL	-25192	-0.5

Tableau IX. – Les variants candidats ES et WGS impliqués dans des CNVs pour le patient BC02

	Gènes candidats	*Variants*	*Type*	*Longueur*	*L2FC*
ES	CDC45	chr22:19462284	DEL	-957927	-0.569238
	NEB	chr2:152435843	DUP	29619	1.81641
	AMBP	chr9:116840434	DEL	-224212	-0.786107
	SPTB	chr14:65009566	DEL	-226685	-0.747523
	POLE2	chr14:50117778	DEL	-423127	-0.369346
WGS	GCNT2	chr6:10582183	DEL	-831	-1.46112

Les expansions de répétitions ont été analysées avec l'outil ExpansionHunter. Cet outil utilise une liste prédéterminée de régions d'expansion d'intérêt. En utilisant la référence, il a été possible de vérifier si pour chaque allèle du gène étudié il est annoté comme sain, porteur de la prémutation ou pathogénique. Nos analyses pour les deux patients étudiés ont seulement trouvé des expansions porteuses de prémutation, comme on peut observer dans le tableau X. Un outil n'utilisant pas une liste de gènes définis serait possiblement plus intéressant, surtout étant donné que plusieurs des gènes parmi ceux-ci ne sont pas liés aux myopathies ou dystrophies

musculaires. Pour la patiente Z26, les évènements communs avec son frère ont été exclus ne laissant aucune expansion à explorer. Pour le patient BC02, on peut observer sur le tableau X quatre évènements d'expansion interprétés comme des porteurs de prémutation. Les gènes concernés sont majoritairement liés à des ataxies ou épilepsies. D'autres investigations sont requises pour ces évènements.

Tableau X. – Les expansions potentiellement pathogéniques.

Patient	***Chromosome***	***Position***	***Ref***	***Alt***	***Gènes***
BC02	chr4	39350044	A	<STR10>,<STR43>	RFC1 (113)
	chr4	41747989	A	<STR27>	PHOX2B (114)
	chr9	71652177	T	<STR27>,<STR41>	FXN_A (115)
	chr21	45196324	G	<STR16>	CSTB (116)

3.3.3 Les variants candidats intégrés avec les résultats du pipeline

Afin de prioriser les variants potentiellement pathogéniques sortis de nos analyses, nous avons intégré les listes dans le but de garder les variants se trouvant plusieurs fois dans les fichiers de sorties filtrés. Dans le tableau XI ci-dessous, nous pouvons retrouver une intégration des variants candidats présentés au début avec les outils d'analyses de notre approche proposée. Pour chaque patient, les variants candidats rares et filtrés sont comparés aux listes des outils d'analyses, soit : SpliceAI; rMATS; LPEseq. Nous avons également fait ceci pour les listes produites par nos scripts maison, soit: les variants SpliceAI DEG; les jonctions rMATS DEG; les variants associés aux jonctions rMATS. Lorsque le filtre utilisé pour l'outil est strict, par exemple rMATS, on ne retrouve aucune correspondance entre les variants cibles. Ceci est également le cas pour les variants rMATS, on retrouve qu'un seul variant candidat dans ce groupe pour le patient Z26 et aucun chez les autres patients. Les variants concernés du tableau XI sont explicitement décrits dans les tableaux présentés dans l'Annexe C-F.

Tableau XI. – Le nombre de variants prioritaires retrouvés au travers des résultats du pipeline

Patient	*Variants candidats **	*Candidats vs SpliceAI*	*Candidats vs rMATS***	*Candidats vs DEG*	*Candidats vs DEG SpliceAI*	*Candidats vs DEG rMATS*	*Candidats vs Variants rMATS****
Z26	68	25	3	6	3	0	1
BC01	31	12	0	3	0	0	0
BC02	53	21	1	6	2	1	0
HSJNM008	47	25	0	6	2	0	0

*Les variants candidats présentés ici concernent ceux proposés par les outils d'annotation pour GATK et VarDict (voir figure 11)

** Les variants candidats ciblés comparés avec les jonctions d'épissage sorties de l'outil rMATS.

*** Les variants candidats ciblés comparés avec les variants RNAseq associés à une jonction d'épissage sortie de l'outil rMATS.

Chapitre 4 – Variants candidats pour les patients

Effectuer ce type d'analyse sur des cas uniques est difficile, surtout pour la priorisation des variants. Avoir une cohorte de patients présentant le même type de phénotype, il serait par exemple plus convenable de chercher des variants communs chez ce groupe de personnes. Dans le contexte des maladies très rares et dans le cadre de ce projet, c'est vraiment du cas par cas. Le pipeline proposé ici permet de regrouper des variants identifiés par plusieurs outils, comme étant soi-disant intéressants, expliquant potentiellement la pathologie du patient. Comme nous pouvons l'observer jusqu'à présent, les résultats présentés démontrent l'importance d'une approche personnalisée afin d'explorer plusieurs hypothèses de diagnostic moléculaire possible. Le fait de bien filtrer les données analysées permet de générer une liste de taille réduite des variants candidats potentiels. Dans ce chapitre, une seconde itération d'exclusion de variants sera présentée. Celle-ci est effectuée manuellement en se basant sur les résultats décrits dans le chapitre précédent et la littérature. Pour chaque patient, une description des meilleurs candidats pathogéniques ainsi que le raisonnement à l'appui de la décision est présentée.

Dans le chapitre précédent, plusieurs variants ont été ciblés par nos analyses que ce soit par notre pipeline ou par les outils d'annotation. Le tableau XII ci-dessous regroupe les candidats les plus prometteurs pour chaque individu. À ce stade dans nos analyses, il n'est pas possible de poser un diagnostic puisque des analyses expérimentales de validation sont requises. Néanmoins, on est confiant de proposer de tels évènements comme étant potentiellement pathogénique pour chaque patient concerné.

Tableau XII. – Les variants candidats finaux pour Z26, BC01, BC02 et HSJNM008.

Patients	*Variants*	*Ref*	*Alt*	*HMZ*	*Protéines*	*Gènes*	*Scores: AF; GERP; PhastCons; SIFT; CADD*	*SpliceAI*	*rMATS**	*L2FC DEG*	*Deb. All.***
Z26	chr1:112323335	G	A	hom	p.L450F	KCND3	6.5E-5; 5.26; 542; 0.2; 23.8	DS = 0	Non	1.45	Oui
	chr1:112524545	C	G	het	p.M268I	KCND3	0; 5.51; 656; 0.93; 17.76	NA	Non	1.45	Non
	chr6:5369210	C	A	het	p.P136H	FARS2	0.0003; 5.38; 537; 1 ;31	DS = 0	Non	0.27	Non
BC01	chr21:17250163	G	A	het	p.E950K	USP25	0; 5.58; 711;0.67; 23.1	DS > 0	Oui	-1.54	NA
BC02	chr17:12897799	A	G	het	p.I644T	ELAC2	0; 5.2; 441; 1; 31	NA	Non	0.28	Non
	chr14:36008779	C	T	hom	p.R2075H	RALGAPA1	0.0001; 5.53; 699; 0.99; 34	DS = 0	Non	-1.89	Non
HSJNM008	chr11:46563850	T	C	het	p.N573D	AMBRA1	0; 5.73; 644; 0.98; 25.2	DS > 0	Oui	0.09	NA
	chr11:46564927	G	C	het	p.P214A	AMBRA1	0.001; 5.93; 619; 0.97; 17.47	DS > 0	Oui	0.09	NA

* Nous avons vérifié si un évènement relié à ce variant a été reporté par rMATS, l'utilisation du "non" signifie aucun évènement et "oui" signifie qu'un évènement pour le même gène a été identifié (pas nécessairement à la même position génomique).
** Décrit s'il y a un débalancement allélique pour ce variant.

Pour chaque variant nous avons vérifié avec l'outil I-Mutant (117) si le changement de la protéine, causé par le variant, impacte la stabilité de la protéine générée en calculant la valeur DDG (où DDG est la variation de l'énergie libre de Gibbs). Un score inférieur à -0.5 est interprété comme une forte diminution de la stabilité des protéines. Polyphen-2 est un score qui permet de prédire l'impact d'un variant en se basant sur la séquence et la structure. Un score de Polyphen-2 supérieure à 0.5 est délétère alors qu'un score inférieur est toléré (118). Ces résultats se retrouvent dans le tableau XIII ci-dessous. Il faut noter que les deux scores ne signifient pas la même chose : PolyPhen-2 évaluent leurs fonctions biologiques alors que la stabilité structurelle des protéines est évaluée par I-mutant (119).

Tableau XIII. – Stabilité structurelle des protéines pour les différents variants candidats.

Patients	***Variants***	***Protéines***	***Gènes***	***PolyPhen-2****		***I-Mutant****	
Z26	chr1:112323335	p.L450F	KCND3	0.05	Toléré	-1.04	--
	chr1:112524545	p.M268I	KCND3	0.01	Toléré	-0.47	-
	chr6:5369210	p.P136H	FARS2	0.96	Délétère	-1.67	--
BC01	chr21:17250163	p.E950K	USP25	0.11	Toléré	-0.79	--
BC02	chr17:12897799	p.I644T	ELAC2	1	Délétère	-0.25	-
	chr14:36008779	p.R2075H	RALGAPA1	0.99	Délétère	-1.18	--
HSJNM008	chr11:46563850	p.N573D	AMBRA1	0.99	Délétère	-0.54	--
	chr11:46564927	p.P214A	AMBRA1	0.27	Toléré	-0.91	--

* Les scores PolyPhen-2 sont décrits à gauche et leur interprétation à droite.
** Les scores I-Mutant DDG sont décrits à gauche et leur interprétation à droite. Le symbole - signifie une faible diminution de stabilité (sous le seuil) et le symbole -- signifie une forte diminution de la stabilité de la protéine.

Les prochaines sections de ce chapitre décrivent pour chaque patient les variants les plus prometteurs et le raisonnement menant à la décision. Les variants proposés sont présentés dans le tableau XII. Nous avons également inclus les résultats de stabilité de la protéine présentés du tableau XIII. Un changement au niveau de l'acide aminé causé par un variant n'est pas nécessairement néfaste. C'est pourquoi nous avons analysé ceci pour chacun des variants proposés. Une validation par IGV est présentée pour chaque variant ainsi qu'une prédiction de l'effet du variant en question sur la structure secondaire de l'ARN par l'outil RNAfold (120). L'outil RNAfold prend en entrée un format FASTA, soit les séquences de nucléotides. Manuellement nous avons changé la séquence de référence trouvée sur UCSC (121) afin d'introduire le changement de nucléotide du variant. Nous avons seulement pris les 100 nucléotides autour du variant afin de mieux visualiser la structure. Les figures présentées pour les analyses de RNAfold ont un système de couleurs décrivant la valeur minimale de l'énergie libre de la structure. Le bleu signifie une valeur de 0, le vert une valeur de 0.5 et le rouge une valeur de 1 (voir ci-dessous). Il a été montré que l'impact d'un changement de la structure secondaire peut être associé à des maladies (122, 123). Lorsque la structure reste inchangée, il est possible d'observer un changement au niveau de l'énergie libre indiquant un potentiel de risque de dysfonctionnement de l'ARN. Aussi, l'impact de ces changements de structure a été lié avec des maladies neurologiques telles que le Parkinson et l'Alzheimer affectant les éléments de reconnaissance des interactions ARN-protéine et ARN-ARN (124).

4.1 Patiente Z26

La présentation clinique de cette patiente est très complexe. On suspecte un génotype multigénique ou possiblement un dysfonctionnement au niveau mitochondrial. Lors de nos recherches de variants candidats, il était important de connaitre son profil clinique afin de mieux prioriser certains évènements. La patiente Z26 a subi déjà plusieurs investigations et un variant proposé était *KCND3*:NM_004980:exon4:c.1348C>T:p.L450F (voir description de la patiente ici : https://undiagnosed.hms.harvard.edu/participants/participant-013/). Nos outils ont également

identifié ce variant avec le RNA-Seq et ES ainsi que d'autres marqueurs potentiellement intéressants.

4.1.1 *KCND3*

Comme mentionné, le variant KCND3:NM_004980:exon4:c.1348C>T:p.L450F est également présent dans nos résultats d'analyses. Cependant, ce variant n'a pas passé tous nos filtres en raison de son score SIFT trop faible. Néanmoins, pour la majorité des scores ce variant semble potentiellement pathogénique. Jusqu'à présent, nous n'avons que validé ce qui était déjà connu pour cette patiente. Voici les nouvelles observations qui ont été découvertes durant ce projet concernant ce variant. Grâce à notre approche, l'intégration des données transcriptomiques et génomiques ont permis d'identifier un débalancement allélique à la position de ce variant (figure 15).

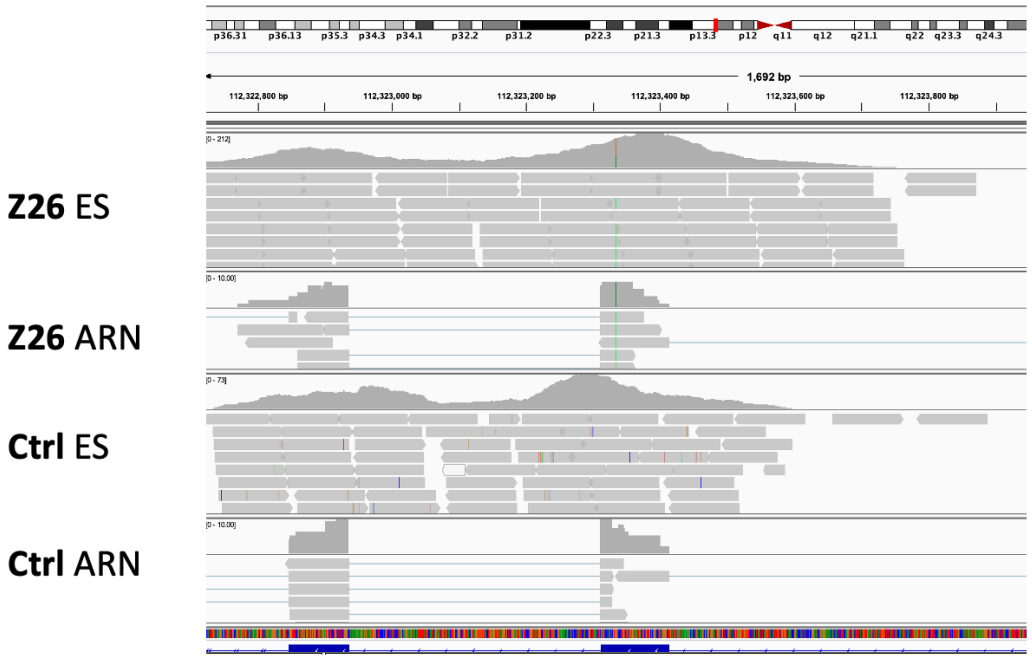

Figure 15. – Visualisation sur IGV du débalancement allélique du variant chr1:112323335 *KCND3* de Z26.

Le variant est homozygote au niveau transcriptomique et hétérozygote au niveau génomique (changement de couleur visible sur la figure 15). Le changement du ratio allélique suggère un débalancement de l'expression allélique du variant. C'est un réarrangement cellulaire

causé par l'inactivation épigénétique d'un allèle ou des variants dans des régions régulatrices atténuant un allèle, résultant à l'expression exclusive de l'autre allèle (125). Il a été démontré que cette différence d'expression allélique peut être responsable d'une prédisposition génétique chez les individus atteints d'une maladie mendélienne (126). Le phénomène de débalancement allélique peut être tissu spécifique, donc il est encore plus intéressant de retrouver dans le contenu musculaire de la patiente un tel évènement (127).

Cet évènement a été identifié par notre script maison. La fréquence allélique de ce variant n'est pas nulle, mais très faible suggérant que c'est quand même un variant rare. Tous sauf le score de SIFT passent les seuils de pathogénicités utilisés. Donc, pour la majorité des outils de prédiction d'ANNOVAR, ce variant est un bon candidat. De plus, la structure secondaire et l'énergie minimum libre sont modifiées par ce variant (figure 16). Le score PolyPhen-2 suggère que ce variant est toléré, mais le score I-Mutant pour ce variant décrit une diminution de la stabilité de la protéine. Ceci peut être causé par le changement de la structure secondaire. D'autres analyses sont nécessaires pour mieux comprendre l'impact fonctionnel du variant. Malgré le fait que ce variant en particulier n'était pas priorisé par les outils d'annotation de pathogénicité, étant donné le débalancement allélique identifié, il reste un bon candidat pour la patiente Z26.

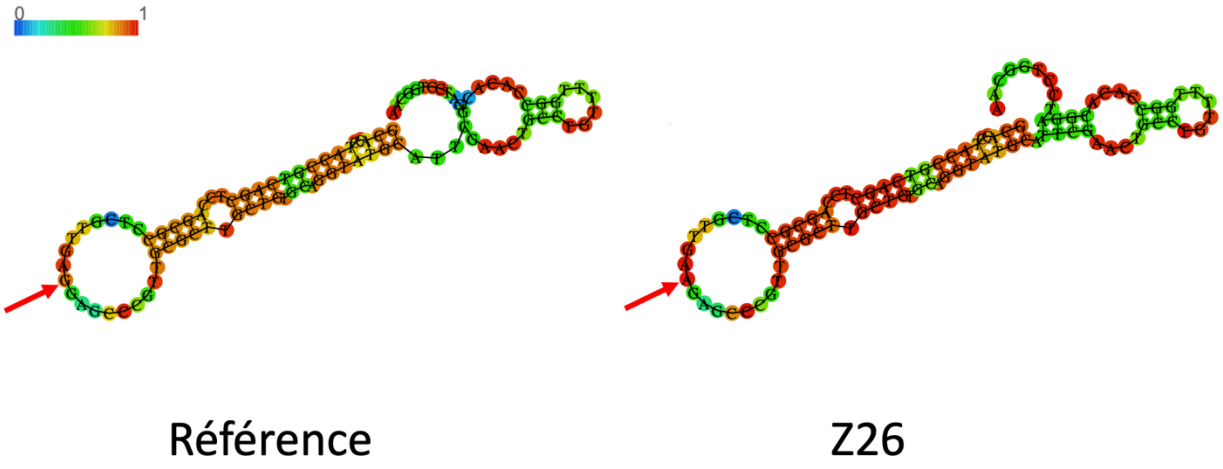

Figure 16. – Structure secondaire RNAfold du variant du variant chr1:112323335 *KCND3* de Z26.

Autrement, un second variant pour ce gène est priorisé par nos analyses de variants candidats, soit le variant KCND3:NM_004980:exon2:c.804G>C:p.M268I (figure 17). Il est important de noter que ce variant n'est pas identifié par ES, donc il est possible que ce variant soit manqué par ES due à une faible couverture ou que ce soit un artefact (128). Une autre possibilité est que la couverture au niveau de l'exome n'est pas assez forte causée par l'âge et la détérioration du tissu musculaire affecté de la patiente (129, 130).

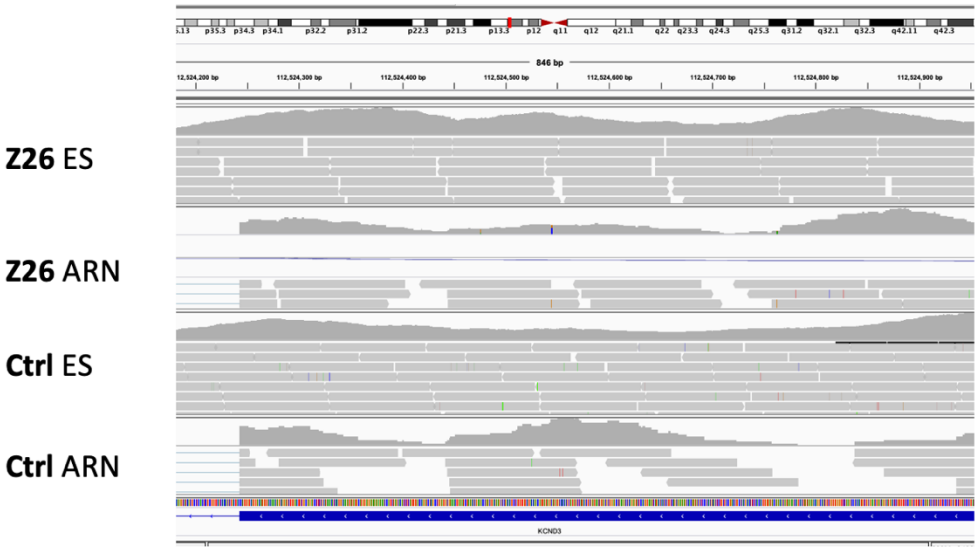

Figure 17. – Visualisation sur IGV du variant chr1:112524545 *KCND3* de Z26.

Pour ce variant tous les scores d'annotation passent les seuils, sa fréquence allélique est de zéro. La structure secondaire présentée sur la figure 18 est modifiée comparativement à celle prédite pour la référence ainsi que l'énergie libre. Ceci peut avoir un effet néfaste sur l'ARN. Le score de PolyPhen-2 prédit que ce variant est toléré et que la stabilité est faiblement diminuée selon I-Mutant. Ces résultats nécessitent plus d'analyse. Une hypothèse intéressante serait que ce variant rare soit *de novo* et donc acquis au courant de la vie de la patiente (131).

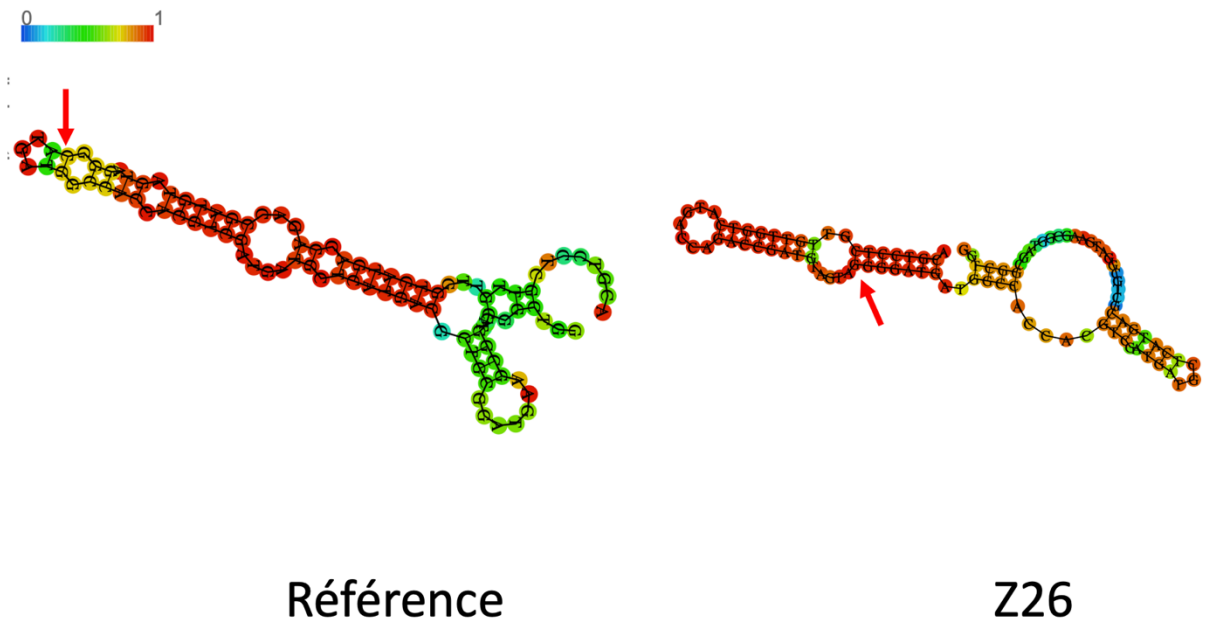

Figure 18. – Structure secondaire RNAfold du variant du variant chr1:112524545 *KCND3* de Z26.

À ce stade, nous avons suggéré deux variants intéressants dans le même gène pour cette patiente. Premièrement, nous avons un déséquilibre allélique présent en amont du gène et un variant hétérozygote potentiellement pathogénique en aval de ce variant. Nous n'avons pas identifié d'évènements d'épissage, ni avec SpliceAI et ni avec rMATS. Le score de L2FC est faible, mais montre une faible augmentation de l'expression du gène. Ce gène appartenant à la famille des canaux potassiques dépendant du voltage est impliqué dans l'échange d'action potentielle dans notre corps (132). Ces canaux sont responsables de la contraction musculaire (133). Il n'y a pas encore d'étude analysant l'effet de ces canaux potassiques sur les muscles squelettiques. À ce jour, dans la littérature, on retrouve seulement que le gène est impliqué dans des maladies telles que les ataxies congénitales et Brugada syndrome 9 (132). La patiente présente des similarités phénotypiques avec Brugada syndrome 9, ce chevauchement clinique confirme que ce gène est un bon candidat. Cependant, on ne peut pas affirmer que ce gène explique la totalité des symptômes de la patiente, spécialement l'atteinte des muscles squelettiques.

4.1.2 *FARS2*

Puisque le phénotype de la patiente est très multisystémique, on suspecte que plusieurs gènes peuvent causer les phénotypes de la patiente. Un autre candidat intéressant serait FARS2:NM_001318872:exon2:c.407C>A:p.P136H (figure 19).

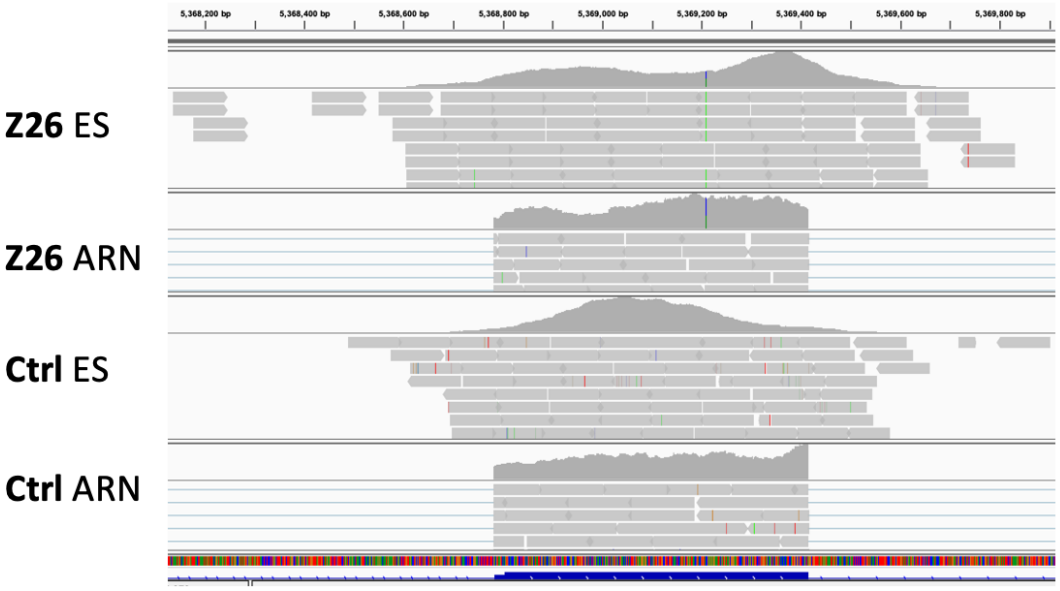

Figure 19. – Visualisation sur IGV du variant chr6:5369210 *FARS2* de Z26.

Ce variant faux-sens a été priorisé par les outils d'annotation, il a passé tous les seuils définit comme étant potentiellement pathogénique. Les scores de GERP; PhastCons; SIFT; et CADD sont tous élevés signifiant une forte conservation génomique et une substitution d'acide aminé délétère pour la protéine. C'est un variant rare avec une fréquence de 0.03% et une interprétation de pathogénicité conflictuelle selon GnomAD. Selon notre pipeline, aucun évènement d'épissage n'est relié à ce variant, donc pour le moment, nous pouvons exclure cette possibilité en ce qui concerne l'interprétation fonctionnelle du variant. L'expression du gène ne semble pas affectée non plus. Le variant est également identifié dans les données ES et priorisé par nos filtres.

La structure secondaire de cette région de la séquence et l'énergie minimum libre ne semblent pas être modifiées selon l'outil de prédiction RNAfold (figure 20). Cela étant dit, nous avons exclu la possibilité d'un changement au niveau de la structure de l'ARN, mais il est possible

que ce variant cause un effet dévastateur sur la protéine qui encode. Selon le score PolyPhen-2 le variant à un impact délétère sur la protéine et déstabilise grandement la protéine selon I-Mutant.

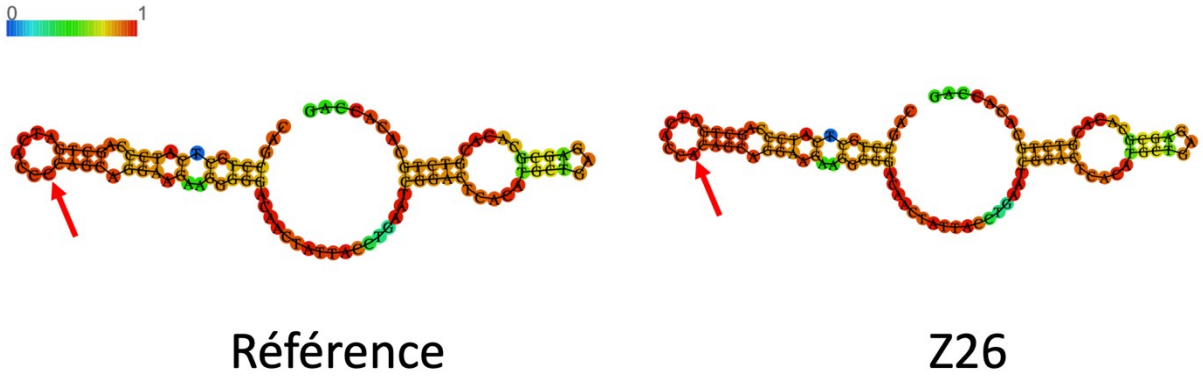

Figure 20. – Structure secondaire RNAfold du variant du variant chr6:5369210 *FARS2* de Z26.

L'implication du gène au niveau de la mitochondrie fait en sorte que ce variant est un excellent candidat pour cette patiente. Comme la présentation du phénotype affecte plusieurs organes, une hypothèse serait que le biomarqueur affecte le bon fonctionnement mitochondrial causant ainsi un large éventail de symptômes chez la patiente. C'est un gène nucléaire encodant pour une protéine de la mitochondrie, il est estimé que dans 99% du temps c'est le cas (134). La transcription et traduction du génome mitochondrial se fait à l'intérieur des mitochondries, mais la majorité des protéines nécessaires à ces processus sont codées dans le génome nucléaire (135). Le gène *FARS2* encode pour le phénylalanyl-ARNt synthétase mitochondriale (mtPheRS) (136). Cette protéine générée par ce gène joue un rôle important dans la traduction de protéine mitochondriale (137). Plusieurs rapports de cas concernant ce gène et divers autres variants ont été rapportés. On associe, à la grande majorité des variants identifiés à ce jour, pour ce gène avec les maladies d'encéphalopathie épileptique, de paraplégie spastique tardive et de convulsions focales (136). Une étude a aussi démontré deux variants composés hétérozygotes (c.925G>A p.Gly309Ser et c.943G>C p.Gly315Arg) pour ce gène chez un frère et une sœur qui ont été associés avec une maladie mitochondriale rare COXPD14 (déficit combiné en phosphorylation oxydative 14) autosomique récessive (138). Plusieurs types de diagnostics COXPD14 existent, ils sont dépendants de l'âge de début et des phénotypes reliés qui varient beaucoup. Dans certains

cas, les phénotypes reliés sont la faiblesse musculaire et des convulsions (138). Ce profil phénotypique est en partie similaire à celui de la patiente. Dans le but de vérifier si ce gène contient un second variant hétérozygote négligé par nos outils de priorisation, une recherche d'un second variant a été effectuée dans les données de la patiente. Un variant faux-sens hétérozygote a été identifié dans le même gène, soit FARS2:NM_001318872:exon4:c.839A>G:p.N280S. Les scores d'annotation suggèrent que ce variant n'est pas pathogénique avec un score CADD de 12.8 et une fréquences alléliques de 0.16. Le variant possède un score de GERP de 3.13 de phastCons 638 et de SIFT 0.5. Ce variant n'est pas un bon candidat comme variant composé pour ce gène. Il est possible qu'avec les données WGS, on puisse identifier un second variant pathogénique pour ce gène.

Nous avons proposé des évènements potentiellement pathogéniques pour cette patiente. D'autres investigations sont nécessaires afin de valider ce qui est déjà trouvé et d'évaluer l'effet de ces biomarqueurs d'un point de vue plus fonctionnel afin de mieux comprendre ce qui se passe au niveau métabolique et protéomique. Pour le gène *KCND3* préalablement ciblé par des analyses effectuées ailleurs, nous avons identifié un évènement de débalancement pour ce même variant candidat proposé. Ceci était possible grâce à notre pipeline proposé. Nous avons pu exclure plusieurs possibilités en ayant accès aux données transcriptomiques de cette patiente. Une investigation de l'épissage alternatif et de l'expression des gènes était possible. Nous avons également analysé les évènements de CNVs et les expansions de répétition pour cette patiente, mais n'avons rien trouvé de significatif. Il est possible qu'avec les données ES, ce ne fût pas suffisant pour identifier quelque chose à ce niveau-là. Cependant les données RNA-Seq et ES ont quand même permis d'identifier des variants potentiellement pathogéniques, tels que deux variants dans le gène *KCND3* ainsi que dans le gène *FARS2*. Il y a des études qui démontrent l'effet pathogénique de variants hétérozygote composé ("compound heterozygous variants" en anglais) chez des maladies neuromusculaires (139, 140). D'autres investigations, par exemple le LRS, sont requises pour confirmer ceci pour les deux variants *KCND3* de cette patiente.

4.2 Patient BC01

Le patient BC01 est un jeune garçon âgé de moins de 5 ans présentant un phénotype progressant depuis sa naissance. Pour faire un rappel de ce qui était présenté dans le chapitre 3,

ce qu'on connait sur ce patient est qu'il présente des faiblesses musculaires, de l'hypotonie, de la fatigue. La biopsie n'était pas informative puis les parents ne sont pas atteints. On suspecte donc dans ce cas, une transmission autosomique récessive ou un variant *de novo*. Pour ce patient, nous avons accès aux données transcriptomiques.

4.2.1 *USP25*

Un excellent candidat identifié lors de nos analyses est le variant faux-sens USP25:NM_013396:exon23:c.2848G>A:p.E950K (figure 21). Nous avons identifié ce gène par les outils d'annotation.

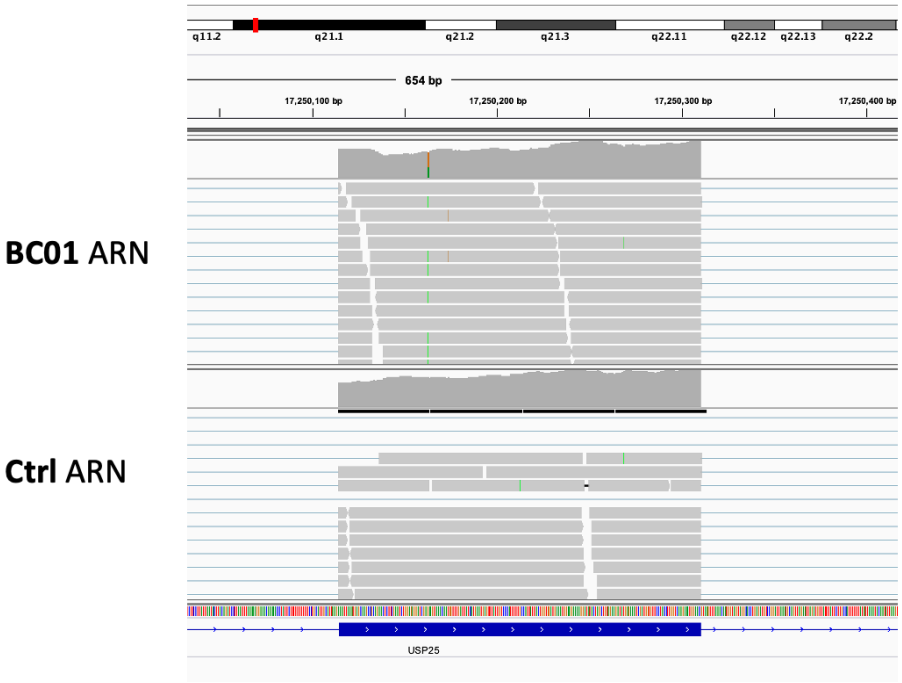

Figure 21. – Visualisation sur IGV du variant chr21:17250163 *USP25* de BC01.

Nous avons également eu l'opportunité d'investiguer au niveau de l'épissage s'il y avait un évènement. Pour SpliceAI, il était ciblé par notre filtre puisqu'au moins un des quatre scores DS est plus grand que 0, soit 0.03; 0.01; 0.04; 0.01 pour DS_AG; DS_AL; DS_DG; DS_DL respectivement. Du côté de rMATS, deux évènements de SE et MXE sur le même gène sont identifiés (tableau XIV). Ces évènements n'ont pas été présentés dans nos résultats du chapitre

trois, puisque leur score de IncLevelDiff est plus bas que le seuil défini pour une transmission récessive.

Tableau XIV. – Les deux évènements d'épissage rMATS pour le variant *USP25*

	Type	*Début*	*Fin*	*IJC_1*	*SJC_1*	*IJC_2*	*SJC_2*	*Incleveldiff*
USP25	MXE	17219981	17220095	391	283	83	102	0.132
USP25	SE	17219981	17220095	391	119	83	42	0.126

Nous avons vérifié avec IGV pour voir s'il y a un épissage au niveau de la position indiquée par rMATS et SpliceAI. Dans les deux cas, après avoir mis le minimum de jonction à 5, les jonctions du patient semblaient identiques à ceux du contrôle. Pour ce variant, la fréquence allélique est de zéro, les scores d'annotation suggèrent que ce variant est dans une région fortement conservée et que l'effet sur la protéine est délétère. Le score GERP; phastCons; SIFT ; CADD est de 5.58; 711; 0.67; 23.1 respectivement. L'expression du gène est diminuée, mais une valeur de L2FC de -1.54 ne passe pas notre seuil défini dans le pipeline. Cependant, notre approche nous permet de quand même avoir les valeurs de DEG des gènes pas sorties dans le fichier filtré final, pour nous permettre de vérifier manuellement lorsque nécessaire un gène. De plus, nous avons été sévères avec le seuil de L2FC, plusieurs études utilisent un seuil de 1 ou 1.5 pour déterminer si un gène est DEG (141-143). Donc, on peut conclure que le gène *USP25* diminue au moins de moitié son expression comparée au contrôle B500, ce qui est significatif et intéressant (141). Enfin, notre pipeline nous permet d'explorer et de tirer des conclusions sur divers sujets (épissage; expression) concernant ce variant. En outre, la structure secondaire prédite suggère un changement structurel causé par ce variant (figure 22). On remarque également que plusieurs changements au niveau de l'énergie libre sont à proximité du variant. Il y a un effet clair sur la structure de l'ARN causé par ce variant. Le score de PolyPhen-2 suggère que ce variant est toléré, mais selon I-Mutant, il déstabilise la protéine.

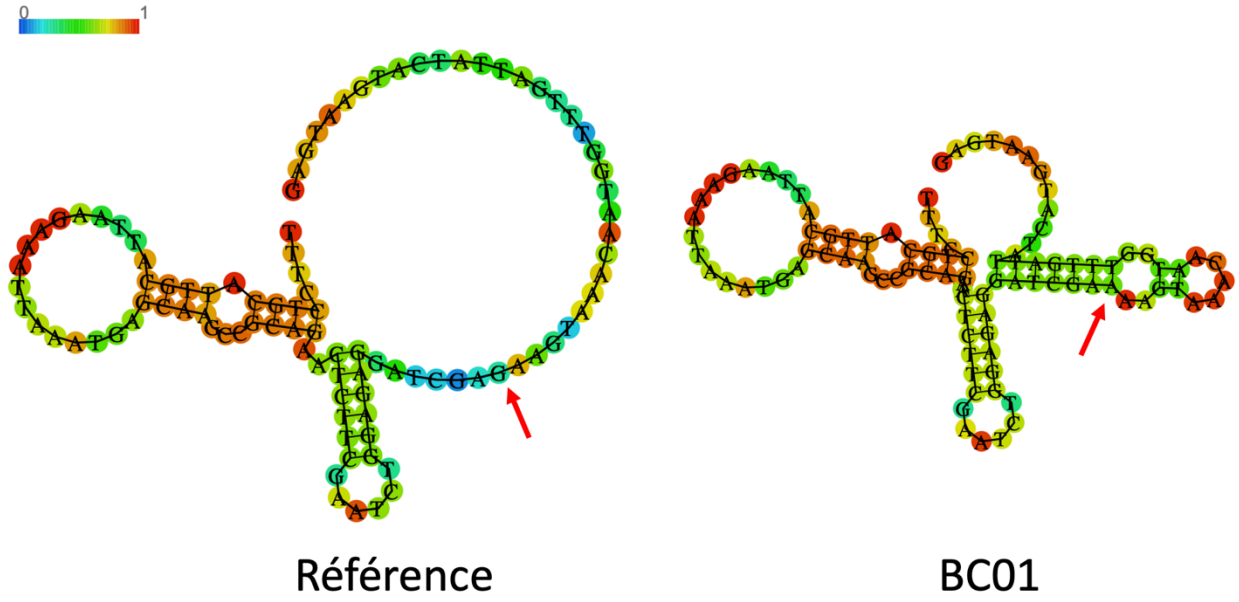

Figure 22. – Structure secondaire RNAfold du variant chr21:17250163 *USP25* de BC01.

Nous avons également effectué une expérience de validation par PCR afin de vérifier si le gène est *de novo*. La salive des deux parents non atteints a été utilisée pour faire ceci. Comme on peut l'observer sur la figure 23, le variant n'est pas présent chez les parents du patient BC01, donc la mutation est présente que chez le patient. Ceci permet de conclure que ce variant hétérozygote peut très bien être la cause derrière le phénotype du patient (144-146).

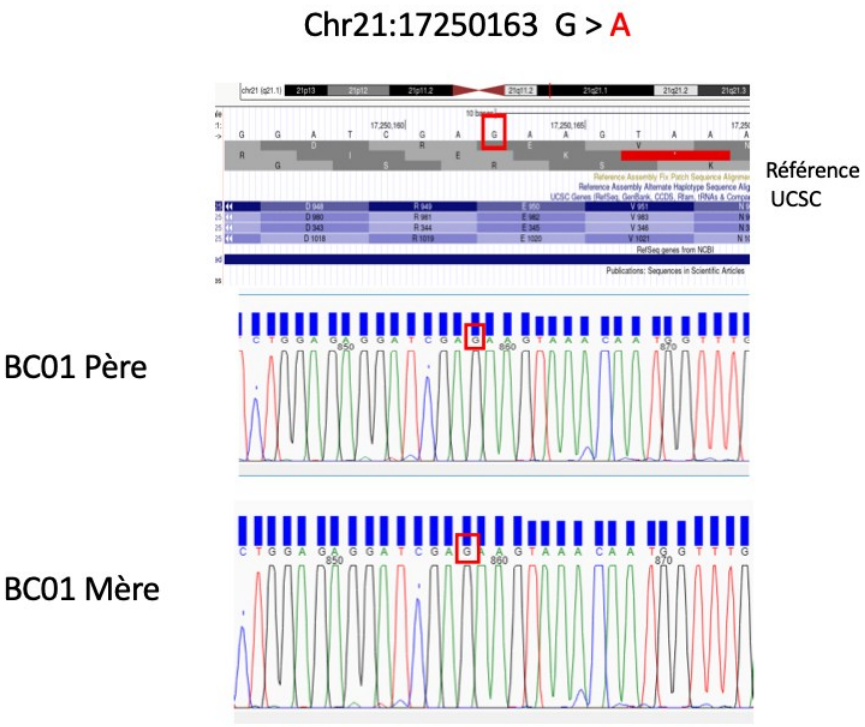

Figure 23. – Validation que le variant chr21:17250163 *USP25* de BC01 est *de novo* par PCR.

Le gène encode pour une enzyme ayant un rôle essentiel dans la dégradation des protéines via le protéasome 26S régulant ainsi de nombreuses voies cellulaires, assurant un équilibre entre substrats ubiquitinés et désubiquitinés (147). Ce gène est impliqué pour plusieurs processus pathologiques comme le cancer, les réponses immunitaires aux infections virales et la neurodégénérescence (148). *USP25* interagit avec trois protéines sarcomériques précédemment impliquées dans la pathogenèse de myopathie sévère ayant tous un rôle essentiel dans le maintien et différentiation musculaire, soit *ACTA1*, la *FLNC* et *MyBPC1* (149). Le gène interagit avec *ACTA1* dans le domaine C-terminal donc le variant proposé ici pourrait affecter cette interaction. *ACTA1* est lié avec plusieurs myopathies, dont l'une, avec disproportion des fibres (150).

Plusieurs facteurs présentés précédemment , nous indiquent que ce variant semble être un très bon candidat pour ce patient. La fonction du gène est très intéressante et pourrait possiblement expliquer le phénotype de notre patient. Les scores d'annotation prédisent tous un effet pathogénique. La structure d'ARN modifiée peut possiblement avoir un impact, mais nous ne pouvons pas confirmer ceci pour le moment. De plus, l'absence du variant chez les parents

non atteints ainsi que la diminution d'expression suggèrent un rôle pathogénique. D'autres analyses sont requises pour évaluer l'effet de ce variant au niveau fonctionnel. Un modèle animal, soit le poisson-zèbre, pourrait être une bonne option afin d'évaluer l'effet de ce variant dans un complexe d'interaction.

4.3 Patient BC02

Le patient BC02 est un jeune garçon âgé de moins de cinq ans, présentant depuis sa naissance des épisodes d'hémiparésies ressemblant à des accidents vasculaires cérébraux (stroke), une rhabdomyolyse et des niveaux de CK très élevé. Les imageries par résonance magnétique sont normales. Le phénotype de ce patient est sévère et progresse rapidement. Pour avoir le maximum d'information pour ce patient, nous avons extrait les données pour séquencer avec le RNA-Seq, ES, et WGS. Nous avons vérifié plusieurs évènements, dont les CNVs, les expansions de répétition, l'épissage alternatif, les gènes différentiellement exprimés et les variants candidats ciblés par les outils d'annotation. Comme nous avons observé dans le dernier chapitre, quelques évènements sont sortis de nos analyses concernant les expansions de répétition et les CNVs. Les gènes ciblés avec ExpansionHunter ne sont pas reliés vraiment avec des problèmes présentés par notre patient tels que des gènes ataxiques et épileptiques. Il y a quelques variants CNVs priorisés, tels que celui pour le gène *NEB*, mais à ce point-ci dans nos analyses, nous n'avons pas eu la possibilité d'approfondir nos recherches concernant les expansions de répétitions à l'aide de la technique du LRS.

4.3.1 *ELAC2*

À partir de nos listes de variants prioritaires, un gène qui est ressorti de nos analyses pour ce patient est *ELAC2*. Un variant candidat identifié pour ce patient est ELAC2:NM_001165962:exon21:c.1931T>C:p.I644T (figure 24). C'est un variant hétérozygote retrouvé dans les données RNA-Seq, ES et WGS du patient.

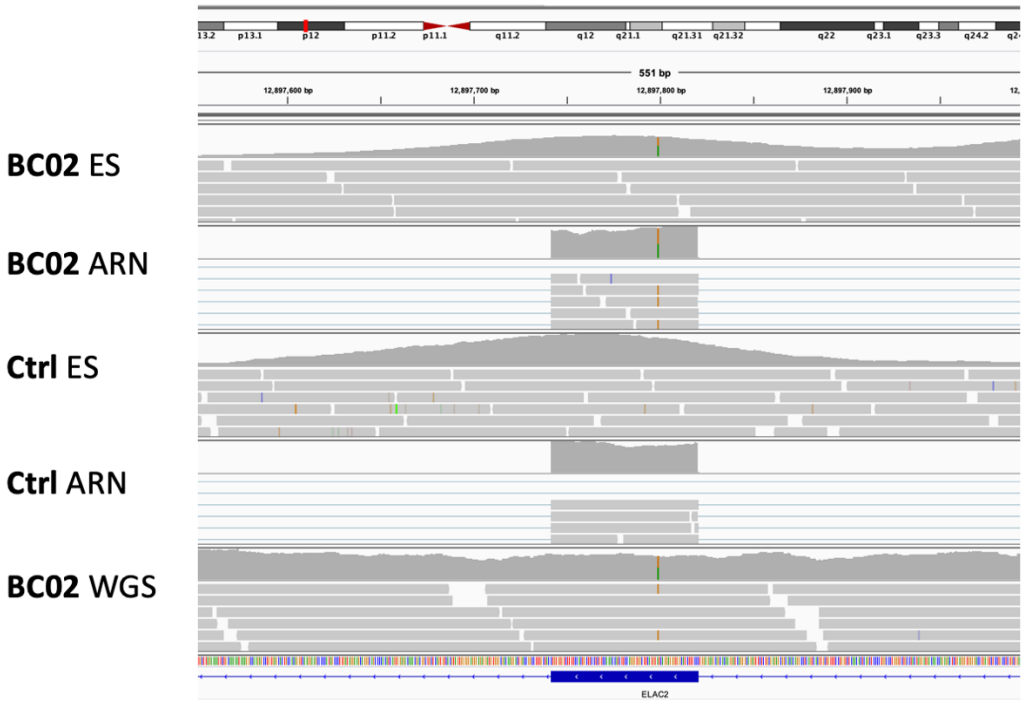

Figure 24. – Visualisation sur IGV du variant chr17:12897799 *ELAC2* de BC02.

Le variant possède une fréquence allélique de zéro, donc il n'a pas été précédemment identifié dans d'autres bases de données. Les scores de conservation GERP et phastCons suggèrent que c'est dans une région génomique conservée, ayant un score de 5.2 et 441. Le changement d'acide nucléotide semble ne pas être toléré, affectant la fonction de la protéine avec un score SIFT de 1. Le score CADD suggère également un impact délétère sur la protéine avec un score de 31. On peut exclure la possibilité d'un épissage concernant ce gène, ainsi que l'expression du gène est très peu modifiée avec un score en bas de 1 pour L2FC. La structure secondaire est modifiée avec ce variant (figure 25), mais peu de différence est observable pour l'énergie libre minimum comparativement à la référence. Selon PolyPhen-2, le variant est délétère, puis une diminution forte de la stabilité de la protéine selon I-Mutant. Les outils de prédictions indiquent que ce variant semble affecter la fonction de la protéine. Ceci est intéressant et solidifie notre affirmation que ce variant semble potentiellement pathogénique pour ce patient.

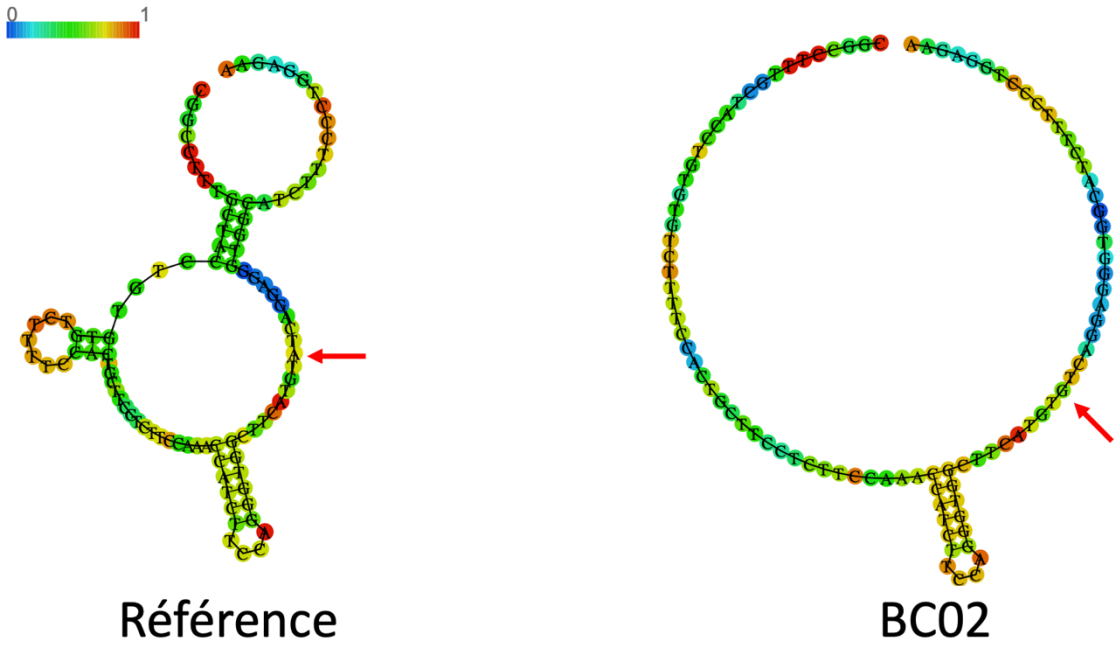

Figure 25. – Structure secondaire RNAfold du variant chr17:12897799 *ELAC2* de BC02.

Ce gène code pour une protéine reliée à une activité d'endonucléase mitochondriale de traitement de l'ARNt 3' qui est essentiel pour la maturation de l'ARNt mitochondrial (151). Le gène était auparavant impliqué dans des cas de cancer de la prostate et de cardiomyopathie (152-154). Bien que les phénotypes reportés à ce jour ne sont pas nécessairement ceux du patient étudié, les symptômes au niveau musculaires concordent. Ils décrivent dans une étude que chez trois patients atteints de cardiomyopathie sévère présentaient une hypotonie, une acidose lactique, une croissance médiocre et un développement psychomoteur retardé(153) . De plus, cette protéine est un élément clé à la synthèse des protéines mitochondriales et à la fonction OXPHOS (153).

Ce variant hétérozygote était le seul à être identifié comme pathogénique. Nous avons vérifié dans les données WGS si un variant hétérozygote dans *ELAC2* également potentiellement pathogénique était présent. Plusieurs autres variants ont été identifiés, mais aucun avec des scores de pathogénicité ou une AF intéressante. Si on considère ce variant, il faudrait vérifier s'il est *de novo*. De plus, il est très possible que le phénotype de patient s'explique par des variants digénique. Pour que le phénotype se présente, deux variants dans deux gènes différents

interagissant ensemble peuvent causer la maladie (155). D'autres investigations sont requises pour ce variant.

4.3.2 *RALGAPA1*

Dans le cas où le variant pathogénique pour ce patient soit hérité, une transmission autosomique récessive est suspectée. Alors, un variant homozygote, où les deux allèles expriment la variation nucléotidique, pourrait expliquer le phénotype de ce patient. Un variant faux-sens homozygote est proposé pour ce patient, soit le RALGAPA1:NM_001346246:exon40:c.6224G>A:p.R2075H (figure 26). Ce variant est présent dans les données RNA-Seq, ES et WGS du patient BC02.

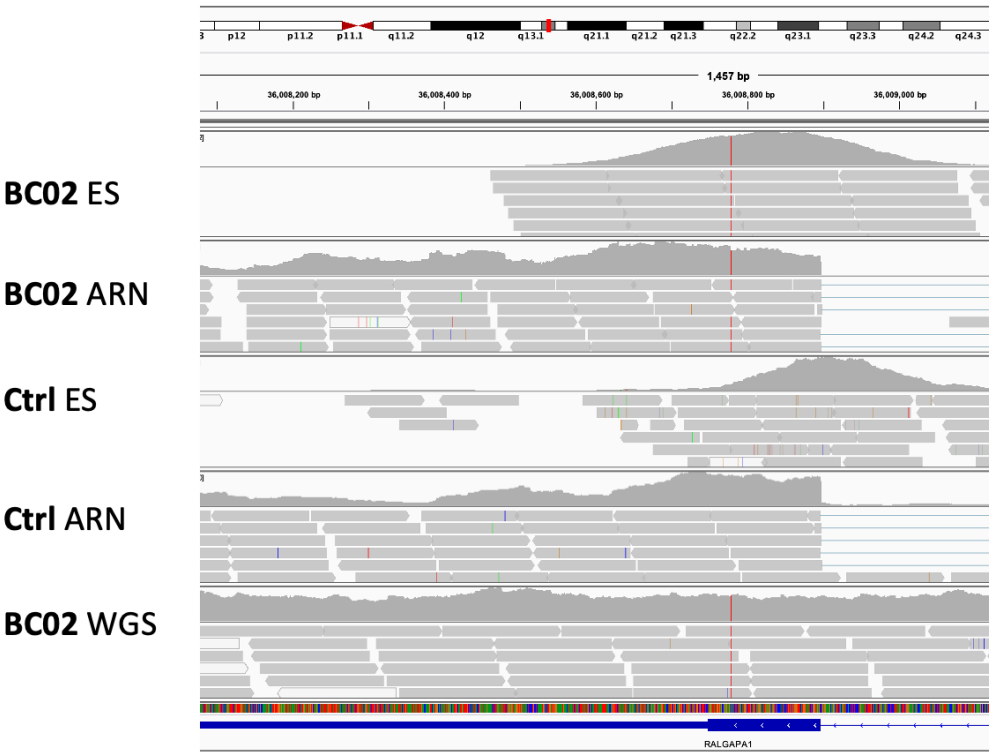

Figure 26. – Visualisation sur IGV du variant chr14:36008779 *RALGAPA1* de BC02.

La fréquence allélique de 0.0001 est faible donc c'est bien un variant rare. Les scores suggèrent que c'est un variant ayant un effet délétère sur la protéine avec un score SIFT et CADD de 0.99 et 34. Il est également prédit d'être fortement conservé avec un score de GERP et phastCons de 5.53

et 699. Aucun épissage n'a été observé avec notre pipeline. L'expression du gène semble diminuée d'au moins de moitié comparativement à un contrôle avec un score de L2FC de -1.89. Ce variant est homozygote donc, on peut facilement lier ce variant à une transmission autosomale récessive, car les deux allèles code pour le variant. De manière intéressante, la structure secondaire présentée sur la figure 27 ne semble pas modifiée par le variant. On peut également remarquer de très faibles changements au niveau de l'énergie libre. Cependant, pour les deux scores de stabilité structurelle, soit PolyPhen-2 et I-Mutant, le variant à un effet délétère sur la protéine générée. Donc bien que la structure prédite reste inchangée, il est tout de même possible que le variant cause un dysfonctionnement de la protéine affectant la fonction et le complexe d'interaction de celle-ci.

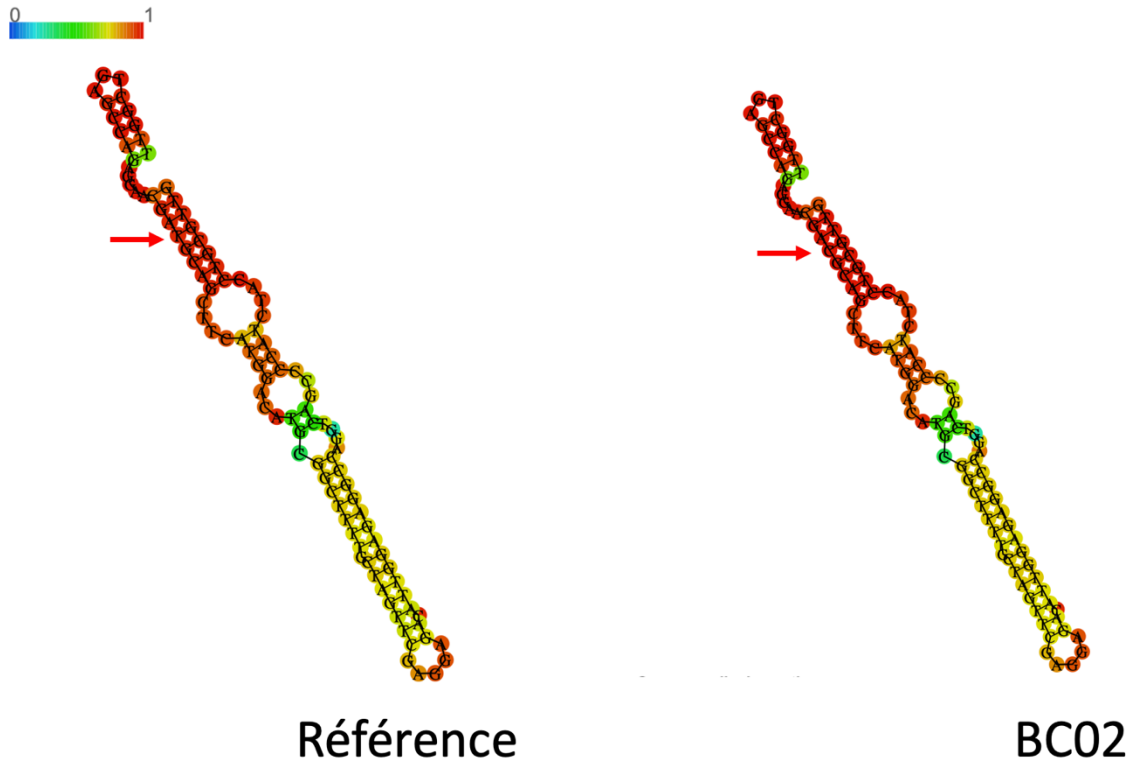

Figure 27. – Structure secondaire RNAfold du variant chr14:36008779 *RALGAPA1* de BC02.

On peut examiner au niveau fonctionnel, pourquoi ce gène pourrait être intéressant. Une étude démontre que la perte de *RALGAPA1* perturbe le complexe RalGAP, conduit à l'activation constitutive de RalA et provoque une maladie neurodéveloppementale sévère avec hypotonie

musculaire profonde, spasmes infantiles et difficultés d'alimentation (156). Il y a évidence dans la littérature que le processus de rhabdomyolyse peut se présenter sous forme d'hypotonie, une étude de cas le démontre (157). Lorsque deux facteurs affectent le muscle, il est très plausible qu'ils soient reliés d'une manière ou d'une autre, ce qui pourrait supporter notre proposition de gène candidat pour notre patient. Des expériences de validations fonctionnelles sont nécessaires.

Dans le cas d'une transmission digénique, il est intéressant d'analyser si les deux gènes proposés, soit *ELAC2* et *RALGAPA1*, interagissent ensemble supportant notre proposition (158). En utilisant l'outil Gene Mania (159) en ligne, il est possible d'analyser si les deux gènes interagissent ensemble. La figure 28 ci-dessous démontre le complexe d'interaction entre les deux gènes.

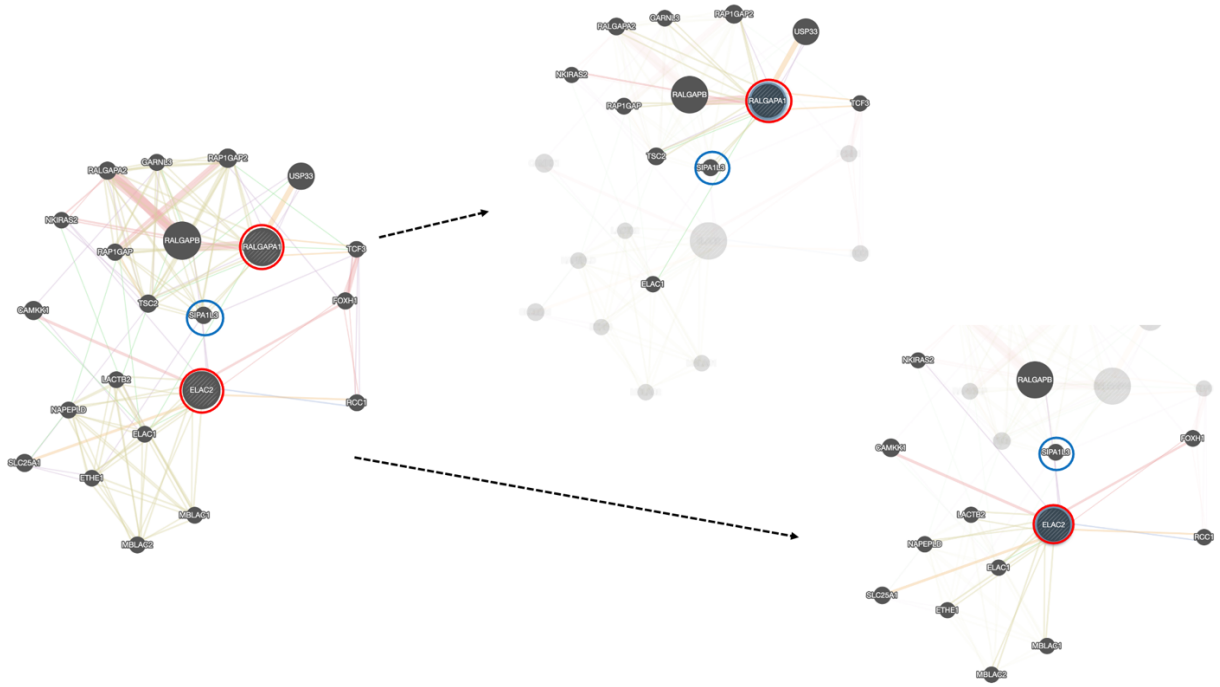

Figure 28. – Complexe d'interaction entre le gène *ELAC2* et *RALGAPA1* selon Gene Mania.

Cette figure représente les réseaux d'interactions entre les gènes. Les lignes représentent les interactions et les nœuds représentent les gènes. Les nœuds encerclés en rouge représentent les gènes *ELAC2* et *RALGAPA1*. Le gène encerclé en bleu réfère au gène *SIPA1L3* qui semblerait interagir avec nos gènes d'intérêt. Les flèches pointent vers des réseaux d'interaction propres à RALGAPA1 en haut et ELAC2 en bas. Dans ces deux réseaux, il y a SIPAIL3 en bleu.

On observe que l'interaction entre les deux gènes ne se fait pas de manière directe, mais plutôt par un gène intermédiaire nommé *SIPA1L3*. Une co-expression est obtenue comme résultat. Dans la littérature, ce gène est impliqué dans des anormalités de polarité cellulaire, la morphogénèse de cellules épithéliales et de l'organisation cytoskelettique (160). Il n'y a aucune évidence pour l'instant de sa fonction au niveau des cellules musculaires. Aussi il n'est pas clair comment ces gènes interagissent ensemble, mais définitivement intéressant qu'une interaction soit possible.

Nous avons proposé deux variants qui semblent les plus prometteurs pour le patient BC02. Nos analyses ont permis d'investiguer l'expression génique, l'épissage alternatif cryptique, les CNVs et les expansions de répétition pour ce patient. Nous avons également analysé l'effet des variants sur la structure de l'ARN et de la stabilité des protéines générées. L'information recueillie à jour nous permet d'avancer dans la recherche du génotype pour ce patient.

4.4 Patient HSJNM008

Le quatrième patient à l'étude pour ce projet est également un jeune garçon présentant des faiblesses musculaires et une encéphalopathie myo-neuro-gastro-intestinale (MNGIE). Son phénotype n'est pas spécifique aux muscles, plus d'un organe est affecté dans son cas. Pour ce patient, nous avons que les données transcriptomiques, plusieurs analyses ont été possible avec ces données. Également, nous avions accès à un contrôle pédiatrique HSJNM009 afin d'exclure pour chaque analyse effectuée sur ce patient, les évènements partagés avec un patient sain. Avoir la possibilité d'un contrôle d'échantillon provenant de muscle squelettique pédiatrique est plus difficile et rare. Ceci nous a permis d'évaluer avec plus de rigueur les données du patient HSJNM008.

4.4.1 *AMBRA1*

Deux excellents candidats ont été identifiés et priorisés par notre pipeline dans le même gène. Ce sont deux variants hétérozygotes ayant des scores de pathogénicité élevés. Le premier variant est AMBRA1:NM_001300731:exon7:c.1717A>G:p.N573D (figure 29) et le second AMBRA1:NM_001267782:exon7:c.640C>G:p.P214A (figure 30).

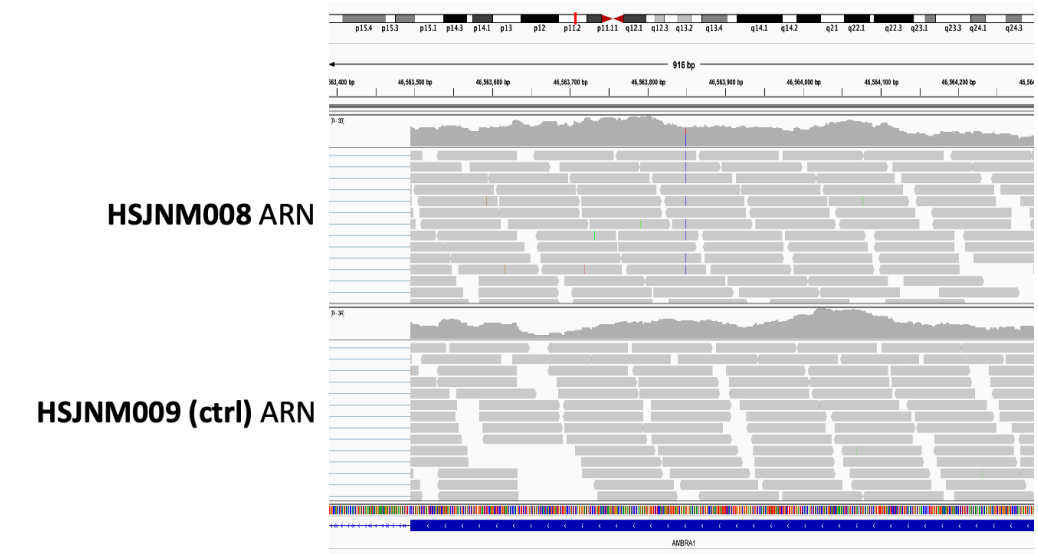

Figure 29. – Visualisation sur IGV du variant chr11:46563850 *AMBRA1* de HSJNM008.

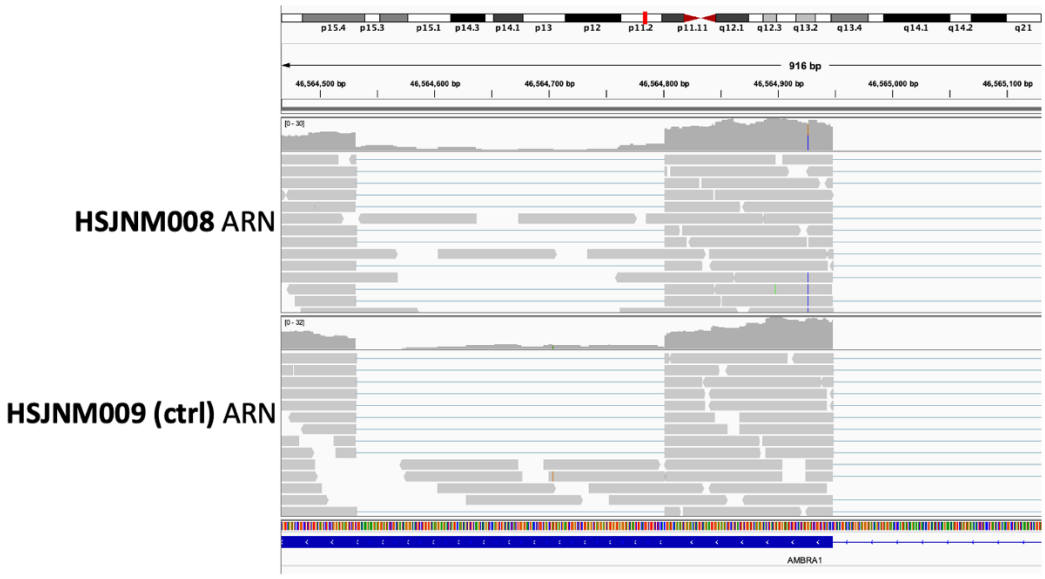

Figure 30. – Visualisation sur IGV du variant chr11:46564927 *AMBRA1* de HSJNM008.

La fréquence allélique du premier variant est égale à zéro et le second est de 0.001. Les scores de conservation GERP et phastCons suggèrent que ces variants sont dans une région génomique conservée, ayant un score de 5.73; 5.93 et 644; 619 pour les deux variants respectivement. Le changement d'acide nucléotide des variants semble ne pas être toléré, affectant la fonction de la protéine avec un score SIFT de 0.98 et 0.97. Les scores CADD suggèrent

également un impact délétère sur la protéine avec un score de 25.2 et 17.47. L'expression du gène est très peu modifiée avec un score en bas de 1 pour L2FC.

La structure secondaire de l'ARN pour le premier variant n'a pas de modification apparente (figure 31), mais on peut observer un changement de l'énergie minimum. Le variant ne semble pas modifier la structure de l'ARN. Cependant, pour le second variant on observe un changement tant au niveau de la structure secondaire de l'ARN que de l'énergie libre (figure 32). De plus, selon les scores de stabilité, le premier variant a un effet prédit délétère sur la protéine selon PolyPhen-2 et un effet déstabilisant selon I-Mutant. Pour le second variant, le score de PolyPhen-2 indique que le variant est toléré, mais avec I-Mutant on remarque un effet prédit fortement déstabilisant.

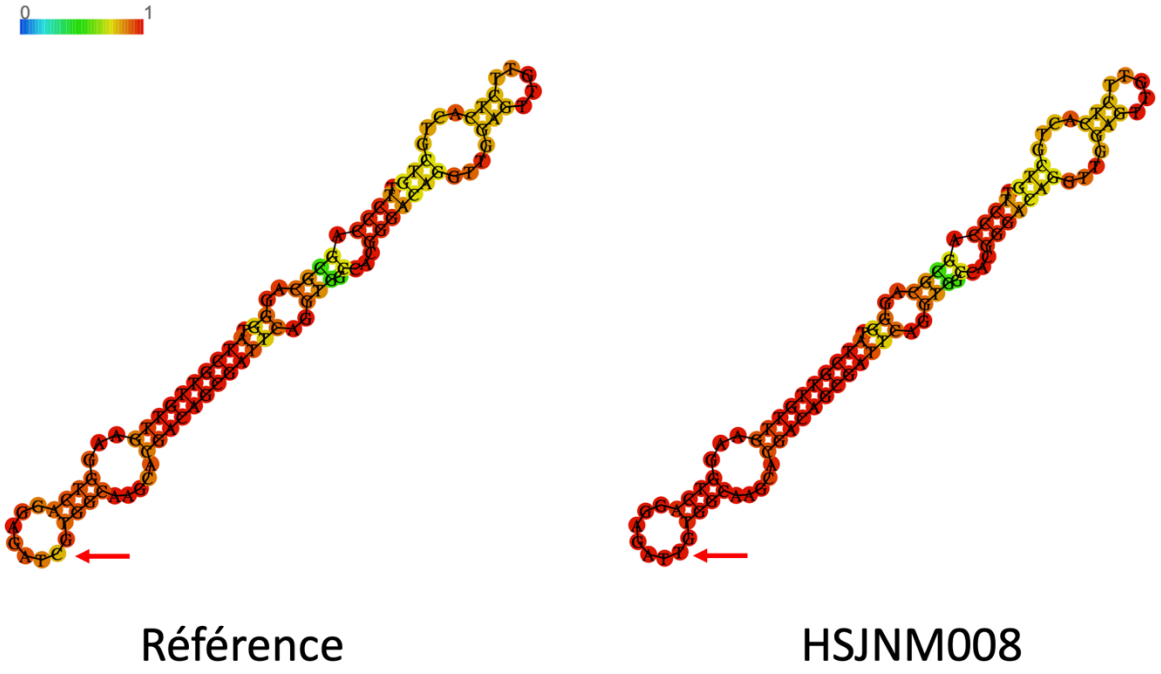

Figure 31. – Structure secondaire RNAfold du variant chr11:46563850 *AMBRA1* de HSJNM008.

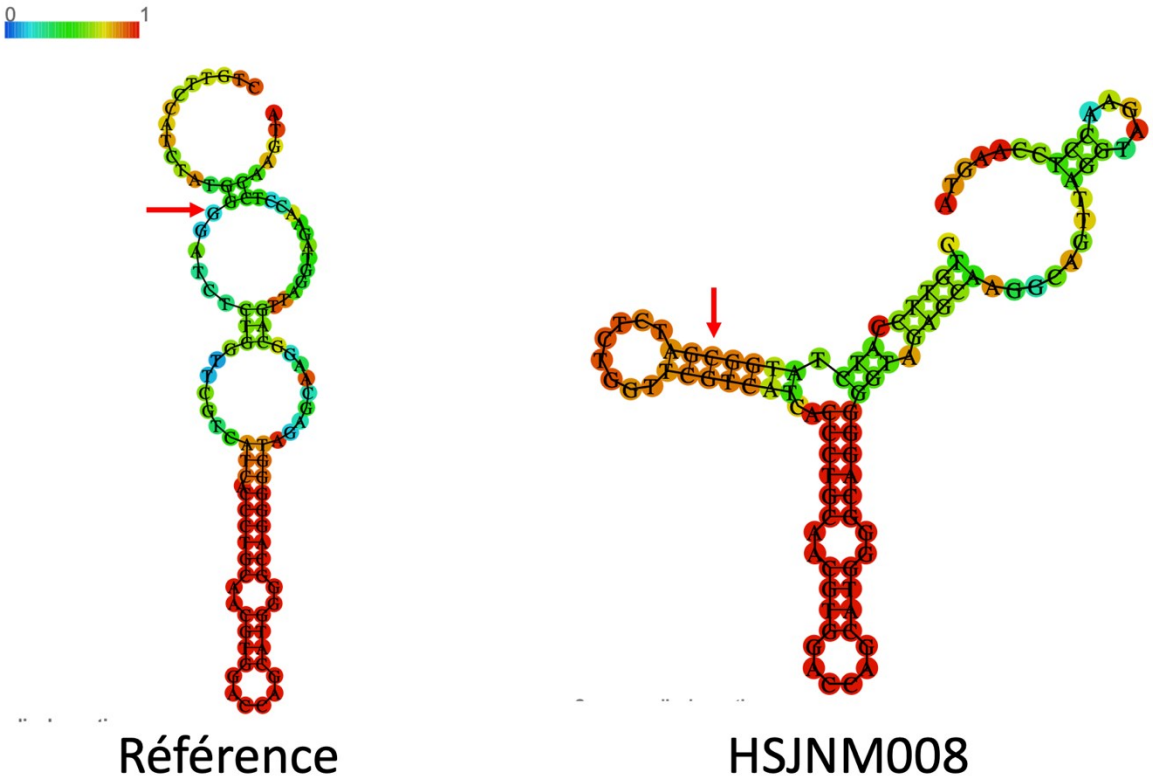

Figure 32. – Structure secondaire RNAfold du variant chr11:46564927 *AMBRA1* de HSJNM008.

Dans l'optique où le diagnostic de ce patient serait par deux variants hétérozygotes composés, ces deux variants seraient d'excellents candidats pour expliquer la pathologie du patient. Avec notre pipeline, ce gène était priorisé par nos filtres de SpliceAI, puisqu'au moins un des quatre scores DS est plus grand que 0, soit 0.03; 0.01; 0.02; 0.00 pour DS_AG; DS_AL; DS_DG; DS_DL respectivement pour le second variant. Il y avait également pour le premier variant *AMBRA1* un score DS_AG égal à 0.01. Avec IGV, une validation de ces évènements était possible, mais le nombre de jonctions était trop faible pour conclure un évènement d'épissage alternatif. Du côté de rMATS, un évènement de MXE sur le même gène est identifié, mais avec un score de IncLevelDiff plus bas que le seuil définit pour récessif (tableau XV). Malgré le fait qu'un évènement est identifié, la position génomique n'est pas proche d'aucun des deux variants proposés. Il est dans ce cas plus difficile d'associer à ce point un lien entre les variants et un épissage cryptique. Possiblement avec du LRS, on pourrait vérifier si l'épissage est réel et vérifier si les deux variants hétérozygotes sont en *trans* ou en *cis*.

Tableau XV. – Évènements d'épissage rMATS pour *AMBRA1*

	Type	*Début*	*Fin*	*IJC_1*	*SJC_1*	*IJC_2*	*SJC_2*	*Incleveldiff*
AMBRA1	MXE	46529740	46529920	9	40	4	56	0.19

Comme expliqué précédemment, deux variants hétérozygotes composés dans les mêmes gènes peuvent causer la pathologie du patient. Maintenant, pourquoi avons-nous considéré ses deux variants à la base? En fait, le gène est impliqué dans le processus de mitophagie, le métabolisme, la mort cellulaire et la division cellulaire (161). Le rôle de la mitophagie est de retirer les mitochondries défectueuses pour prévenir une accumulation dangereuse, sa dérégulation s'est avérée être impliquée dans plusieurs troubles affectant les muscles, y compris les dystrophies musculaires (162). Ce processus est essentiel pour la fonction musculaire et un dysfonctionnement de celle-ci pourrait très bien expliquer le phénotype de notre patient. Il a été démontré par un modèle de poisson-zèbre que l'expression de ce gène est impliquée dans la formation de la myogénèse, au bon développement et à la morphogenèse du muscle squelettique (163). Il y a beaucoup d'évidence dans la littérature que ce gène est un facteur important pour les muscles et qu'une diminution de sa fonction serait pathogénique.

Chapitre 5 – Discussion et conclusion

Les maladies neuromusculaires sont complexes et rares. Des analyses de précisions personnalisées sont requises afin de repérer un ou plusieurs biomarqueurs pathogéniques. Pour contribuer aux recherches, nous avons mis en place une approche bio-informatique intégrant plusieurs types de données analysées provenant de différents omiques. Nous avons également eu l'occasion d'appliquer notre approche sur les données disponibles de quatre patients présentant tous des formes différentes de maladies neuromusculaires.

Pour faire un rappel, voici une description rapide du flux de travail général ainsi que certains des outils utilisés au cours du projet. Initialement, les échantillons de muscle squelettique de chaque patient ont été séquencés par RNA-Seq et, si possible, par ES ou WGS. Un pipeline bio-informatique pour l'alignement et l'annotation a été décrit dans le chapitre 3. Cette étape a permis de prioriser des variants candidats utilisés pour nos analyses. Pour le séquençage d'ARN, nous avons analysé l'épissage alternatif ainsi que l'expression différentielle. Pour l'objectif d'épissage, SpliceAI est l'un des outils utilisés pour les données ARN utilisant en entrée une liste de variants. Une fois l'outil lancé pour les patients, nous obtenons un VCF de variants annotés par les scores SpliceAI qui sont ensuite filtrés avec un script personnalisé. Par la suite, rMATS utilise la liste des jonctions d'épissage en entrée générant une série de fichiers .txt annotés pour chaque A3SS ; A5SS ; MXE ; RI ; SE. Les fichiers de sortie sont filtrés avec d'autres scripts écrits en Python. Avec les fichiers filtrés, la confirmation de ces événements d'épissage anormaux sur IGV est désormais possible. Pour analyser le niveau d'expression des gènes, cela se fait à l'aide d'un script écrit en R utilisant LPEseq. Par la suite, des scripts écrits en Python permettent d'intégrer les informations recueillies des diverses listes d'évènements dans le but de réduire la liste de variants ciblés. On recherche des variants rares avec un impact fonctionnel inconnu pour associer le variant à un changement d'expression significatif. Donc, nous avons effectué l'intégration de la liste des variants VCF avec les variants d'épissage donnés par rMATS pour associer un variant à une jonction d'épissage pour les patients. Ensuite, les variants ont été intégrés à l'expression

génique à l'aide d'un script personnalisé. On travaille avec des données transcriptomiques et génomiques pour en savoir plus sur les VUSs et prioriser ces gènes lors de la recherche des variants pathogéniques. Alors, nous avons utilisé les listes de variants du transcriptome et du génome pour appliquer un script écrit en Python qui recherche les débalancements alléliques. De plus, pour plusieurs patients, une myopathie mitochondriale est suspectée, nous avons donc mis en place un pipeline d'analyse pour examiner le génome mitochondrial. L'analyse de gène nucléaire relié au génome mitochondrial était également possible par notre approche. Pour la recherche de CNVs et d'expansion de répétitions, cela a été fait avec les outils CNVkit et ExpansionHunter.

La raison pour ce rappel de la méthode est vraiment pour mettre en valeur les divers outils et scripts instaurés lors de ce projet. Il faut comprendre aussi que dans ce domaine, plusieurs outils d'analyse sont présentés à la communauté scientifique. Chacun stipulant être le plus innovateur, le plus rapide et le plus précis. Il fallait durant ce projet prendre une décision sur lequel choisir pour nos analyses. Un travail d'analyse comparative ("benchmarking" en anglais) a été fait lors de mon stage sur les divers outils d'épissage alternatif qui était disponible à ce moment-là. Nous avons comparé la performance entre SpliceAI ; Alamut® Batch version 1.11; rMATS; LeafCutter (164). Pour nos analyses effectuées dans notre domaine de recherche, nous avons décidé de procéder par une combinaison d'outils, soit SpliceAI et rMATS. Plusieurs études d'analyse comparative sont disponibles sur ces outils d'épissage et le consensus est que SpliceAI est un excellent outil de prédiction d'épissage (60, 61, 165). Nous avons décidé d'utiliser rMATS puisqu'il permet de regarder au niveau des jonctions d'épissage directement lesquels semble causer un épissage cryptique. De plus, il est convenable d'avoir les différents types d'épissage analysés, soit A3SS; A5SS; MXE; RI; SE.

Pour faire un bref retour sur les résultats de l'application du pipeline, nous allons discuter de certains points importants à soulever. Initialement, nous avons filtré les données sorties de l'analyse d'annotation par ANNOVAR et ainsi priorisé des variants d'intérêt. Comme démontré, plusieurs seuils ont été définis sur différents scores de pathogénicité. Cette approche nous a

permis de prioriser un bon nombre de variants candidats par patients. Ceci a été réalisé sur les données RNA-Seq, et lorsque possible, ES et WGS. Ces gènes ciblés ont servi au courant de nos analyses pour vérifier et prioriser les évènements intéressants pour chacun des patients. Pour cette étape, nous avons décidé de combiner les variants identifiés par GATK et VarDict. Nous voulions vraiment cibler les meilleurs candidats selon ANNOVAR. De plus, il est montré que VarDict surpasse GATK surtout lorsqu'il s'agit de donnée RNA-seq (32). Par contre, pour la majorité de nos analyses les fichiers sortis avec GATK sont utilisés en raison de compatibilité de format accepté par les autres outils d'analyses. Ceci est en fait une limitation dans notre approche. Nous avons besoin de tester la performance des nouvelles versions de GATK afin de mieux décider quel appel de variants est optimal. Bref, pour la majorité des variants ciblés et discutés, ils ont été appelés par les deux GATK et VarDict. On remarque également que le nombre de variants communs, entre les deux outils d'appel de variants, est plus élevé au niveau génomique, probablement le résultat de la moins bonne performance de GATK avec les données transcriptomiques. Autrement, les seuils utilisés lors de nos analyses ne concordent pas toujours avec ceux suggérés dans la littérature. Ce choix de seuils différents a été pris selon les besoins de ce projet. Nous ne voulions pas être trop stricts dès le début avec l'appel de variants candidats qui seront utilisés au courant de nos analyses pour prioriser les évènements potentiellement pathogéniques.

Il est pertinent de discuter d'ethnicité dans un contexte de projet de recherche clinique. La présence de certains variants et leur fréquence allélique sont influencées par l'ethnicité d'une personne. Bien que l'ethnicité soit très importante en recherche, il est parfois difficile d'obtenir cette information de la part des cliniciens. Une raison étant l'identification des patients: plus que nous avons d'informations (sexe; âge; ethnicité) plus qu'il y a des chances d'identifier le patient. Donc certains sont réticents à donner cette information sur des bases éthiques. Le fait de ne pas avoir l'information sur l'ethnicité peut faire en sorte qu'un variant nous semble rare, mais en réalité cela est dû au fait qu'il n'y a pas assez d'information sur cette population ou bien nous regardons la fréquence totale et non spécifique aux patients. Nous pouvons tout de même exclure les variants communs, puisque comme nous travaillons avec des maladies rares nous cherchons

des variants rares peu importe la population. L'impact de l'ethnicité sur la découverte de nouveaux variants n'est donc pas aussi important dans ce contexte en comparaison avec des études de génétique de population ou études d'association. De plus, puisque nous travaillons avec le muscle, il n'est pas toujours possible d'avoir des contrôles correspondant avec l'ethnicité du patient en question. Lors du recrutement de patient, malgré le fait qu'on n'exclut personne, la réalité est que nous sommes un laboratoire au Canada collaborant avec des cliniques canadiennes où la grande majorité des patients sont d'ascendance Européenne.

Notre pipeline a permis de générer plusieurs listes de variants et évènements prioritaires comme nous avons démontré dans les résultats. Ceci a vraiment permis de ressortir ce qui était pertinent dans les listes de milliers de variants candidats étudiés par patient. Il est important de mentionner qu'il est possible d'introduire de faux positifs avec notre pipeline. Nous avons souvent été moins stringent au niveau des seuils suggérés pour plusieurs scores. Puisque nous sommes moins stringent, on augmente le risque d'introduire des évènements priorisés qui ne sont pas réellement pathogéniques. Par exemple, on remarque dans ce mémoire que la statistique de SIFT semble souvent contredire la majorité des autres scores étudiés pour un variant. Il aurait peut-être été mieux d'exclure ce score de nos analyses de priorisation afin de réduire le risque d'erreur de faux négatifs. Cela étant dit, pour chaque patient une inspection manuelle des variants obtenus dans nos analyses est effectuée initialement afin d'explorer le plus possible les données obtenues pour nos patients. Aussi, un faible risque de faux négatif est présent dès que nous utilisons des seuils pour prioriser certains évènements. Il est donc possible que certains évènements ne soient pas priorisés face à cet égard. Nous avons décidé d'être moins sévères avec le seuil de SIFT afin de limiter les faux négatifs, mais ceci augmente le risque de faux positifs. Il est certain que pour optimiser ce pipeline, un ajustement de certains seuils est possible afin d'éviter le plus possible ces erreurs.

Le but de ce pipeline est de permettre d'effectuer plusieurs analyses au niveau du transcriptome et du génome. Comme nous l'avons observé dans ce projet, analyser l'épissage; les DEG; les variants mitochondriaux; les CNVs; les expansions et les débalancements alléliques

génèrent beaucoup de données. Il est difficile de tirer des conclusions pertinentes. Avec notre approche, il est possible de gérer toutes ces données de manières efficaces. Nos filtres réduisent davantage les listes de variants bruts de chacun des outils. Nos scripts personnalisés permettent d'analyser les évènements potentiellement pathogéniques possiblement impliqués à plusieurs facteurs soit au niveau de l'épissage, de l'expression différentielle, des variants candidats annotés et des répétitions. Faire cela manuellement, sans l'aide de scripts, serait une tâche lourde avec possiblement beaucoup d'erreurs. En perspective, une optimisation du pipeline présenté consisterait tout d'abord de réévaluer la pertinence de certains outils présentant des résultats contradictoires. Pour le moment, on présente la première itération de ce pipeline dans ce mémoire mis en pratique sur des données de patients.

Nous avons présenté les variants les plus prometteurs pour chacun de nos patients dans le chapitre quatre. Il est possible grâce à notre pipeline, d'analyser des variants ciblés par plusieurs avenues différentes, soit l'épissage; l'expression; les CNVs; et les expansions de répétitions. Notre approche nous a permis, pour chacun des variants cibles, de faire un rapport complet sur ce qui est présent ou pas chez ces candidats. Donc, il a été possible d'exclure pour le moment certaines hypothèses concernant l'implication d'un épissage cryptique chez certains patients et de même établir de nouvelles hypothèses liées à nos découvertes. Cela a permis de regrouper les candidats pathogéniques les plus pertinents provenant d'une analyse complexe et profonde.

Nous sommes conscients que certains variants RNA-Seq ont une couverture plutôt faible, mais comme expliqué nous ne voulons pas filtrer sur la couverture due à un potentiel dégradation d'allèle et de certains gènes connus pour causer des syndromes musculaires qui sont faiblement exprimés (myopathie métabolique ou mitochondriale par exemple). Nous compensons pour ceci en validant par séquençage de Sanger ou LRS. Dans le cas précis du gène *KCND3* de Z26, une couverture faible est retrouvée. Nous savons que ce gène explique une partie des symptômes (voir fiche de la patiente) donc nous avons décidé de ne pas l'exclure. Comme expliqué, nous préférons avoir de faux positifs lors de nos analyses que d'éliminer des candidats potentiellement pathogéniques. Il faut valider par d'autres méthodes avant d'exclure les variants présentés dans

ce mémoire. De plus, dans le cas des patients où l'on suspecte une myopathie mitochondriale, on s'attend à ce qu'ils aient une plus faible couverture : le muscle est l'un des tissus où il y a le plus de mitochondrie. Encore une fois, ceci justifie pourquoi nous ne voulons pas filtrer pour le nombre de lectures/la couverture. Une couverture plus faible, soit 10X, ne devrait pas être ignorée lors de nos analyses de recherche de variants potentiellement pathogénique (112). Dernièrement, nos analyses proviennent de biopsie de muscle pathologique qui peut avoir parfois des infiltrations graisseuses. Donc, il se pourrait que dans l'échantillon musculaire recueilli on retrouve moins de muscle que dans une biopsie musculaire saine (pas atteint de maladie neuromusculaire). Ceci pourrait influencer en partie le nombre de lectures retrouvées dans nos analyses. En somme, plusieurs facteurs justifient pourquoi nous préférons ne pas filtrer sur la couverture / nombre de lectures.

Il y a constamment de nouveaux outils bio-informatiques qui sont développés. Ce domaine est soumis à des changements constants et de nouvelles approches deviennent disponibles. Comme il a été mentionné, nous apprenons et testons constamment de nouveaux outils. Proposer un pipeline, comme nous l'avons fait dans ce projet, en utilisant de nombreux outils peut rapidement devenir obsolète. Au moment de l'établissement de notre flux de travail, les outils présentés étaient parmi les meilleurs outils disponibles. Aujourd'hui, on peut déjà proposer deux nouveaux outils performants. Par exemple, Pangolin (166) et Introme (https://github.com/KCCG/introme). Une étude d'analyse comparative utilisant Pangolin démontre qu'il est plus performant que SpliceAI (166). Il serait intéressant de valider ceci pour nos données. Du côté de l'analyse de DEG, il est difficile d’avoir accès à des contrôles pédiatriques de biopsies musculaires. C'est pourquoi il n'était pas possible d'avoir accès à des répliques dans nos expériences. Avoir des répliques fournit un plus grand pouvoir statistique lors des analyses de DEG, nous avons proposé un outil LPEseq qui est conçu pour ce genre de cas. Une possibilité que nous avons considérée est d'utiliser les données publiques GTEx (167) pour les analyses de DEG. Le problème avec cette proposition est qu'elle aurait introduit des effets de lots. Une solution élaborée dans une étude démontre l'utilisation d'un modèle de régression linéaire multiple pour corriger les effets de lots (168). L'utilisation des données GTEx est une approche

qui pourrait potentiellement améliorer les résultats d'analyses DEG pour des cas non répliqués. Malgré tout, encore à ce jour, LPEseq semble la meilleure approche pour traiter des données sans réplique. De plus, il a été démontré que cet outil performait mieux que DESeq dans le cas de sans réplique (72). Il serait pertinent de le comparer avec un autre outil également conçu pour des données non répliquées, mais il n'y en a pas présentement. Aussi, nous avons proposé l'outil ExpansionHunter pour l'analyse des expansions de répétitions. L'outil est trop spécifique dans ses recherches, utilisant une liste prédéterminée de région d'expansion ciblée. Ceci est intéressant comme analyse, mais dans notre cas, il serait plus pertinent d'utiliser une approche recherchant des expansions dans d'autres gènes que ceux ciblés par l'outil. Nos résultats présentés pour cet outil étaient plutôt des gènes qui n'étaient pas reliés au phénotype de nos patients, par exemple des gènes reliés aux ataxies ou épilepsies. ExpansionHunter DeNovo (46) serait plus pertinent pour nos analyses. Il n'utilise pas de catalogue de gènes prédéfinis, puis il recherche pour l'entièreté du génome des évènements d'expansions de répétitions. Ceci permettrait une analyse et une recherche plus complète d'évènements non reportés précédemment. Dans le cadre de ce projet, il n'a pas été possible de se servir du LRS ciblé. C'est une excellente manière de valider des évènements intéressants identifiés tels que l'épissage alternatif, les expansions de répétitions, les CNVs et même de vérifier si des variants sont transmis en *trans* ou en *cis*. Les séquences de longues lectures permettent de bien visualiser ces évènements. Nous ainsi que d'autres chercheurs de la littérature avons démontré l'efficacité de cette approche en identifiant des variants et des gènes candidats pour plusieurs patients (82, 169, 170). Bref, le pipeline que nous avons proposé s'est montré puissant et efficace. Il serait cependant intéressant de tester notre approche avec les nouveaux outils.

Notre approche multi-omiques était plutôt centrée sur les données transcriptomiques et génomiques. Comme mentionné dans le premier chapitre, il y a plusieurs omiques que nous pouvons explorer tels que la métabolomique et la protéomique. La métabolomique pourrait permettre d'identifier des marqueurs affectant le processus métabolite normal d'un organe ou d'un tissu afin de bien comprendre la progression de la maladie (171). Ces métabolites peuvent également être utilisés comme test préliminaire pour obtenir un diagnostic et servir comme

approche thérapeutique. Si nous arrivons à diagnostiquer le patient plus rapidement, on pourrait réduire la progression de la maladie et améliorer ses symptômes plus tôt. L'idée est similaire pour l'utilisation d'une approche protéomique. La combinaison de protéomique et génomique est déjà utilisée dans le cas de maladie neuromusculaire rare (172, 173). Cela pourrait grandement enrichir et mieux compléter nos analyses afin de mieux étudier l'impact de nos variants sur la stabilité des protéines encodés. Une autre possibilité d'analyse supplémentaire, pour nos patients sans diagnostic, serait une approche épigénétique. En fait, cela permettrait l'analyse de changement de l'activité génique causé par des facteurs autres que des variants, soit la méthylation de l'ADN et les modifications d'histones (174). La combinaison de cette approche pourrait grandement avancer et améliorer la recherche de variant pathogénique hérité ou sporadique dans le cas de maladie neuromusculaire rare (175).

En somme, WGS, WES et RNA-Seq ont montré d'excellents résultats en termes de découverte de variants. L'approche multi-omique proposée a permis d'identifier des marqueurs génomiques prometteurs chez des patients atteints de maladies orphelines, soit les myopathies et les dystrophies musculaires. La composante multisystémique des patients recrutés rend le diagnostic moléculaire très complexe. Ce projet aura un bénéfice direct pour les patients et leurs familles en mettant un terme à leur odyssée diagnostique. Les patients bénéficieront d'un conseil génétique et d'une prise en charge clinique plus adaptée et personnalisée. De plus, l'identification de nouveaux gènes candidats contribuera à l'avancement des connaissances sur les mécanismes physiopathologiques à l'origine des myopathies. Définir l'étiologie génétique est la première étape vers l'identification de cibles thérapeutiques potentielles et le développement de traitements curatifs. L'objectif principal de ce projet d'identifier des variants pathogéniques pour nos patients avec notre pipeline est atteint. Pour le futur, il serait intéressant de recruter des patients supplémentaires sans diagnostic moléculaire et ainsi d'optimiser le pipeline. Cette approche pourrait être utilisée par d'autres chercheurs travaillant sur les maladies rares.

Références bibliographiques

1. Zatz M, Passos-Bueno MR, Vainzof M. Neuromuscular disorders: genes, genetic counseling and therapeutic trials. Genet Mol Biol. 2016;39(3):339-48.
2. Mary P, Servais L, Vialle R. Neuromuscular diseases: Diagnosis and management. Orthopaedics & Traumatology: Surgery & Research: Supplement. 2018;104(1 Supplement):S89-S95.
3. Laing NG. Genetics of neuromuscular disorders. Crit Rev Clin Lab Sci. 2012;49(2):33-48.
4. McDonald CM. Clinical Approach to the Diagnostic Evaluation of Hereditary and Acquired Neuromuscular Diseases. Phys Med Rehabil Clin N Am. 2012;23(3):495-563.
5. Hassan Nagy KDV. StatPearls: Myopathies2021 Jan.
6. Cardamone M, Darras, B. T., & Ryan, M. M. Inherited myopathies and muscular dystrophies. Semin Neurol. 2008.
7. Lovering RM, Porter NC, Bloch RJ. The muscular dystrophies: from genes to therapies. Phys Ther. 2005;85(12):1372-88.
8. Brinkmeyer-Langford C, Kornegay J. Comparative Genomics of X-linked Muscular Dystrophies: The Golden Retriever Model. Current Genomics. 2013;14(5):330-42.
9. Mendell JR, Lloyd-Puryear M. Report of MDA muscle disease symposium on newborn screening for Duchenne muscular dystrophy. Muscle Nerve. 2013;48(1):21-6.
10. Gao QQ, McNally EM. The Dystrophin Complex: Structure, Function, and Implications for Therapy. Comprehensive Physiology. 2015:1223-39.
11. Milone M, Wong LJ. Diagnosis of mitochondrial myopathies. Mol Genet Metab. 2013;110(1-2):35-41.
12. Ahmed ST, Craven L, Russell OM, Turnbull DM, Vincent AE. Diagnosis and Treatment of Mitochondrial Myopathies. Neurotherapeutics. 2018;15(4):943-53.
13. Ahuja AS. Understanding mitochondrial myopathies: a review. PeerJ. 2018;6:e4790.
14. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1).
15. Simic G. Rare diseases and omics-driven personalized medicine.(EDITORIAL). Croat Med J. 2019;60(6):485(3).
16. Ahmed Z, Zeeshan S, Mendhe D, Dong X. Human gene and disease associations for clinical-genomics and precision medicine research. Clinical and Translational Medicine. 2020;10(1):297-318.
17. Goldman AD, Landweber LF. What Is a Genome? PLoS Genet. 2016;12(7):e1006181.
18. Scherrer K, Jost J. Gene and genon concept: coding versus regulation. Theory Biosci. 2007;126(2):65-113.
19. Brown T. Genomes. 2nd edition. Oxford: Wiley-Lies. 2002:Chapter 1, The Human Genome.
20. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. Briefings in Functional Genomics. 2016;15(5 Special Issue: Marine genomics: insights and challenges):374-84.
21. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications. 2014;7(9):1026-42.

22. Ross JP, Dion PA, Rouleau GA. Exome sequencing in genetic disease: recent advances and considerations. F1000Research. 2020;9:336.
23. Kim M-J, Yum M-S, Seo GH, Lee Y, Jang HN, Ko T-S, et al. Clinical Application of Whole Exome Sequencing to Identify Rare but Remediable Neurologic Disorders. Journal of Clinical Medicine. 2020;9(11):3724.
24. Naomi Laflamme* VT, Laurence Martineau, Dènahin Hinnoutondji Toffa, Annie Laplante, Patrick Cossette, Éric Samarut, Martine Tétreault, Dang Khoa Nguyen. X-linked bilateral polymicrogyria with epilepsy and intellectual disability associated with a novel KIF4A variant. In revision. 2022.
25. Souche E, Beltran S, Brosens E, Belmont JW, Fossum M, Riess O, et al. Recommendations for whole genome sequencing in diagnostics for rare diseases. Eur J Hum Genet. 2022.
26. Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. The Lancet Neurology. 2022;21(3):234-45.
27. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583(7814):96-102.
28. Ellingford JM, Barton S, Bhaskar S, Williams SG, Sergouniotis PI, O'Sullivan J, et al. Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. Ophthalmology. 2016;123(5):1143-50.
29. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med. 2018;20(4):435-43.
30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
31. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108-e.
32. Sandmann S, De Graaf AO, Karimi M, Van Der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep. 2017;7(1):43169.
33. Zhang F, Gu W, Hurles ME, Lupski JR. Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics. 2009;10(1):451-81.
34. Moreno-Cabrera JM, Del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur J Hum Genet. 2020;28(12):1645-55.
35. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences. 2015;112(17):5473-8.
36. Välipakka S, Savarese M, Johari M, Sagath L, Arumilli M, Kiiski K, et al. Copy number variation analysis increases the diagnostic yield in muscle diseases. Neurology Genetics. 2017;3(6):e204.

37. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12(4):e1004873.
38. Chen Y-C, Seifuddin F, Nguyen C, Yang Z, Chen W, Yan C, et al. Comprehensive Assessment of Somatic Copy Number Variation Calling Using Next-Generation Sequencing Data. 2021.
39. Kuśmirek W. Different Strategies for Counting the Depth of Coverage in Copy Number Variation Calling Tools. Bioinform Biol Insights. 2022;16:117793222211155.
40. Cava C, Bertoli G, Castiglioni I. Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. BMC Syst Biol. 2015;9(1).
41. Paulson H. Repeat expansion diseases. Elsevier; 2018. p. 105-23.
42. Zhao X-N, Usdin K. The Repeat Expansion Diseases: The dark side of DNA repair. DNA Repair. 2015;32:96-105.
43. Chen Z, Yan Yau W, Jaunmuktane Z, Tucci A, Sivakumar P, Gagliano Taliun SA, et al. Neuronal intranuclear inclusion disease is genetically heterogeneous. Annals of Clinical and Translational Neurology. 2020;7(9):1716-25.
44. Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome Med. 2022;14(1).
45. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35(22):4754-6.
46. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, Van Vugt JJFA, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020;21(1).
47. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57-63.
48. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harbor Protocols. 2015;2015(11):pdb.top084970.
49. Gonorazky H, Liang M, Cummings B, Lek M, Micallef J, Hawkins C, et al. RNAseq analysis for the diagnosis of muscular dystrophy. Annals of clinical and translational neurology. 2016;3(1):55-60.
50. Peymani F, Farzeen A, Prokisch H. RNA sequencing role and application in clinical diagnostic. Pediatric Investigation. 2022;6(1):29-35.
51. Uapinyoying P, Goecks J, Knoblach SM, Panchapakesan K, Bonnemann CG, Partridge TA, et al. A long-read RNA-seq approach to identify novel transcripts of very large genes. Genome Res. 2020;30(6):885-97.
52. Wang Y, Liu J, Huang B, Xu Y-M, Li J, Huang L-F, et al. Mechanism of alternative splicing and its regulation. Biomedical Reports. 2015;3(2):152-8.
53. Humphrey J, Emmett W, Fratta P, Isaacs AM, Plagnol V. Quantitative analysis of cryptic splicing associated with TDP-43 depletion. BMC Med Genomics. 2017;10(1).
54. Douglas AGL, Wood MJA. Splicing therapy for neuromuscular disease. Molecular and Cellular Neuroscience. 2013;56:169-85.
55. Aartsma-Rus A, Arechavala-Gomeza V, Khoo B. Splicing modulation therapy in the treatment of genetic diseases. The Application of Clinical Genetics. 2014:245.

56. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-48.e24.
57. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences. 2014;111(51):E5593-E601.
58. Ha C, Kim J-W, Jang J-H. Performance Evaluation of SpliceAI for the Prediction of Splicing of NF1 Variants. Genes. 2021;12(9):1308.
59. Jang W, Park J, Chae H, Kim M. Comparison of In Silico Tools for Splice-Altering Variant Prediction Using Established Spliceogenic Variants: An End-User's Point of View. International Journal of Genomics. 2022;2022:1-6.
60. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genetics in Medicine : Official journal of the American College of Medical Genetics and Genomics. 2020;22(6):1005-14.
61. Riepe TV, Khan M, Roosing S, Cremers FPM, t Hoen PAC. Benchmarking deep learning splice prediction tools using functional splice assays. Hum Mutat. 2021;42(7):799-810.
62. Ding L, Rath E, Bai Y. Comparison of Alternative Splicing Junction Detection Tools Using RNA-Seq Data. Current Genomics. 2017;18(3):268-77.
63. Mehmood A, Laiho A, Venäläinen MS, McGlinchey AJ, Wang N, Elo LL. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform. 2019;21(6):2052-65.
64. Guillermo M-P, Sergio L, Javier B, Juan Carlos T. RNA-Seq Perspectives to Improve Clinical Diagnosis. Frontiers in Genetics [Internet]. 10.
65. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136(6):665-77.
66. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386):eaal5209.
67. Vainzof M, Zatz M. Protein defects in neuromuscular diseases. Braz J Med Biol Res. 2003;36(5):543-55.
68. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11(12):220.
69. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12).
70. Ren X, Kuan P-F. Negative binomial additive model for RNA-Seq data analysis. BMC Bioinformatics. 2020;21(1).
71. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1).
72. Gim J, Won S, Park T. LPEseq: Local-Pooled-Error Test for RNA Sequencing Experiments with a Small Number of Replicates. PLoS One. 2016;11(8):e0159182.
73. Heyer EE, Deveson IW, Wooi D, Selinger CI, Lyons RJ, Hayes VM, et al. Diagnosis of fusion genes using targeted RNA sequencing. Nature Communications. 2019;10(1).

74. Latysheva NS, Babu MM. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res. 2016;44(10):4487-503.
75. Yan, Duanmu, Zeng, Liu, Song. Mitochondrial DNA: Distribution, Mutations, and Elimination. Cells. 2019;8(4):379.
76. Ortiz GG, Mireles-Ramírez MA, González-Usigli H, Macías-Islas MA, Bitzer-Quintero OK, Torres-Sánchez ED, et al. Mitochondrial Aging and Metabolism: The Importance of a Good Relationship in the Central Nervous System. InTech; 2018.
77. Calabrese C, Simone D, Diroma MA, Santorsola M, Guttà C, Gasparre G, et al. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics. 2014;30(21):3115-7.
78. Rubino F, Piredda R, Calabrese FM, Simone D, Lang M, Calabrese C, et al. HmtDB, a genomic resource for mitochondrion-based human variability studies. Nucleic Acids Res. 2012;40(Database issue):D1150-D9.
79. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1).
80. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol. 2015;33(7):736-42.
81. Liu Q, Tong Y, Wang K. Genome-wide detection of short tandem repeat expansions by long-read sequencing. BMC Bioinformatics. 2020;21(S21).
82. Mezreani J, Audet S, Martin F, Charbonneau J, Triassi V, Bareke E, et al. Novel homozygous nonsense mutation of MLIP and compensatory alternative splicing. npj Genomic Medicine. 2022;7(1).
83. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. Frontiers in Genetics. 2019;10.
84. Quaio CRDAC, Obando MJR, Perazzio SF, Dutra AP, Chung CH, Moreira CM, et al. Exome sequencing and targeted gene panels: a simulated comparison of diagnostic yield using data from 158 patients with rare diseases. Genet Mol Biol. 2021;44(4).
85. Guo Y, Long J, He J, Li C-I, Cai Q, Shu X-O, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012;13(1):194.
86. Efthymiou S, Manole A, Houlden H. Next-generation sequencing in neuromuscular diseases. Curr Opin Neurol. 2016;29(5):527-36.
87. Waddell LB, Bryen SJ, Cummings BB, Bournazos A, Evesson FJ, Joshi H, et al. WGS and RNA Studies Diagnose Noncoding DMD Variants in Males With High Creatine Kinase. Neurology Genetics. 2021;7(1):e554.
88. Yépez VA, Gusic M, Kopajtich R, Mertes C, Smith NH, Alston CL, et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. 2021.
89. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018;19(5):299-310.
90. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. The American Journal of Human Genetics. 2019;104(3):466-83.
91. French JD, Edwards SL. Allelic imbalance in human breast cancer. Oncotarget. 2017;8(7):10763-4.

92. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017;136(9):1093-111.
93. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. JAMA. 2014;312(18):1880.
94. Andrew S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom. 2010.
95. Kanzi AM, San JE, Chimukangara B, Wilkinson E, Fish M, Ramsuran V, et al. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Frontiers in Genetics. 2020.
96. Heng L, Richard D. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics [Internet]. 2009; 25(14):[1754-60 pp.].
97. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-15.
98. Simon A, Paul Theodor P, Wolfgang H. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics [Internet]. 2015; 31(2):[166-9 pp.].
99. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
100. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
101. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10(10):1556-66.
102. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
103. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-13.
104. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50.
105. Pauline CN, Steven H. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res [Internet]. 2003; 31(13):[3812-4 pp.].
106. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1 Database issue):D886-D94.
107. Raychaudhuri S. Mapping Rare and Common Causal Alleles for Complex Human Diseases. Cell. 2011;147(1):57-69.
108. Choquet K, Tétreault M, Yang S, La Piana R, Dicaire M-J, Vanstone MR, et al. SPG7 mutations explain a significant proportion of French Canadian spastic ataxia cases. Eur J Hum Genet. 2016;24(7):1016-21.
109. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(Web Server issue):W452-W7.
110. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. Genome Res. 2001;11(5):863-74.

111. Zhou D, Sun Y, Jia Y, Liu D, Wang J, Chen X, et al. Bioinformatics and functional analyses of key genes in smoking-associated lung adenocarcinoma. Oncol Lett. 2019.
112. Quaglieri A, Flensburg C, Speed TP, Majewski IJ. Finding a suitable library size to call variants in RNA-Seq. BMC Bioinformatics. 2020;21(1).
113. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51(4):649-58.
114. Weese-Mayer DE RC, Khaytin I, et al. Congenital Central Hypoventilation Syndrome. GeneReviews. Seattle (WA): University of Washington. 2004 Jan 28 [Updated 2021 Jan 28].
115. Bidichandani SI DM. Friedreich Ataxia. GeneReviews. Seattle (WA): University of Washington. 1998 [Updated 2017 Jun 1].
116. Lehesjoki AE KR. Progressive Myoclonic Epilepsy Type 1. GeneReviews. Seattle (WA): University of Washington. 2004 Jun 24 [Updated 2020 Jul 2].
117. Emidio C, Piero F, Rita C. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res [Internet]. 2005; 33(2):[W306-W10 pp.].
118. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Current Protocols in Human Genetics. 2013;76(1):7.20.1-7..41.
119. Ghosh N, Nandi S, Saha I. A review on evolution of emerging SARS-CoV-2 variants based on spike glycoprotein. Int Immunopharmacol. 2022;105.
120. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Res. 2008;36(Web Server issue):W70-W4.
121. Lee BT, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2022 update. Nucleic Acids Res. 2021;50(D1):D1115-D22.
122. Martens L, Rühle F, Witten A, Meder B, Katus HA, Arbustini E, et al. A genetic variant alters the secondary structure of the lncRNA H19 and is associated with dilated cardiomyopathy. RNA Biol. 2021;18(sup1):409-15.
123. Joseane Biso de C, Guilherme Loss de M, Thays Cristine dos Santos V, Natana Chaves R, Juan Clinton L, Sayonara Maria de Carvalho G, et al. miRNA Genetic Variants Alter Their Secondary Structure and Expression in Patients With RASopathies Syndromes. Frontiers in Genetics [Internet]. 10.
124. Bernat V, Matthew. RNA Structures as Mediators of Neurological Diseases and as Drug Targets. Neuron. 2015;87(1):28-46.
125. Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M. Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. PLoS Comput Biol. 2010;6(7):e1000849.
126. De La Chapelle A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. Oncogene. 2009;28(38):3345-8.
127. Pinter SF, Colognori D, Beliveau BJ, Sadreyev RI, Payer B, Yildirim E, et al. Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome. Genetics. 2015;200(2):537-49.
128. Curry PDK, Broda KL, Carroll CJ. The Role of RNA-Sequencing as a New Genetic Diagnosis Tool. Current Genetic Medicine Reports. 2021;9(2):13-21.
129. Thuriot F, Gravel E, Buote C, Doyon M, Lapointe E, Marcoux L, et al. Molecular diagnosis of muscular diseases in outpatient clinics. Neurology Genetics. 2020;6(2):e408.

130. Scicchitano BM, Rizzuto E, Musarò A. Counteracting muscle wasting in aging and neuromuscular diseases: the critical role of IGF-1. Aging. 2009;1(5):451-7.
131. Mani A. Pathogenicity of De Novo Rare Variants. Circ Cardiovasc Genet. 2017;10(6).
132. Pollini L, Galosi S, Tolve M, Caputi C, Carducci C, Angeloni A, et al. KCND3-Related Neurological Disorders: From Old to Emerging Clinical Phenotypes. Int J Mol Sci. 2020;21(16):5802.
133. Grande M, Suàrez E, Vicente R, Cantó C, Coma M, Tamkun MM, et al. Voltage-dependent K⁺channel β subunits in muscle: Differential regulation during postnatal development and myogenesis. J Cell Physiol. 2003;195(2):187-93.
134. El-Hattab AW, Craigen WJ, Scaglia F. Mitochondrial DNA maintenance defects. BBA - Molecular Basis of Disease. 2017;1863(6):1539-55.
135. Ali AT, Boehme L, Carbajosa G, Seitan VC, Small KS, Hodgkinson A. Nuclear genetic regulation of the human mitochondrial transcriptome. eLife. 2019;8.
136. Hotait M, Nasreddine W, El-Khoury R, Dirani M, Nawfal O, Beydoun A. FARS2 Mutations: More Than Two Phenotypes? A Case Report. Frontiers in Genetics. 2020;11.
137. Banerjee R, Chakraborty S. Phenylalanyl-tRNA synthetase. Research and Reports in Biochemistry. 2016:25.
138. Li L, Ma J, Wang J, Dong L, Liu S. Two Chinese siblings of combined oxidative phosphorylation deficiency 14 caused by compound heterozygous variants in FARS2. Eur J Med Res. 2022;27(1).
139. Jiaming L, Yinghua C, Xin W, Qinglei W, Hongjun W, Bo Y. The Novel Compound Heterozygous Mutations of GAA Gene in Mainland Chinese Patient with Classic Infantile-Onset Pompe Disease A Case Report and Literature Review. Int Heart J [Internet]. 2020; 61(1):[178-82 pp.].
140. Schaaf CP, Blazo M, Lewis RA, Tonini RE, Takei H, Wang J, et al. Early-onset severe neuromuscular phenotype associated with compound heterozygosity for OPA1 mutations. Mol Genet Metab. 2011;103(4):383-7.
141. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: Review and integration. Brief Bioinform. 2019;20(6):2044-54.
142. Peart MJ, Smyth GK, Van Laar RK, Bowtell DD, Richon VM, Marks PA, et al. Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. Proceedings of the National Academy of Sciences. 2005;102(10):3697-702.
143. Raouf A, Zhao Y, To K, Stingl J, Delaney A, Barbara M, et al. Transcriptome Analysis of the Normal Human Mammary Cell Commitment and Differentiation Process. Cell Stem Cell. 2008;3(1):109-18.
144. Waldrop M, Amornvit J, Sahenk Z, Pierson CR, Boue DR. A novel de novo heterozygous SCN4A mutation causing congenital myopathy, myotonia and multiple congenital anomalies. Journal of Neuromuscular Diseases. 2019;6(4):467-73.
145. Haijes HA, Koster MJE, Rehmann H, Li D, Hakonarson H, Cappuccio G, et al. De Novo Heterozygous POLR2A Variants Cause a Neurodevelopmental Syndrome with Profound Infantile-Onset Hypotonia. The American Journal of Human Genetics. 2019;105(2):283-301.
146. Hernandez-Lain A, Husson I, Monnier N, Farnoux C, Brochier G, Lacène E, et al. de novo RYR1 heterozygous mutation (I4898T) causing lethal core-rod myopathy in twins. Eur J Med Genet. 2011;54(1):29-33.

147. Valero R, Marfany G, González-Angulo O, González-González G, Puelles L, Gonzàlez-Duarte R. USP25, a Novel Gene Encoding a Deubiquitinating Enzyme, Is Located in the Gene-Poor Region 21q11.2. Genomics. 1999;62(3):395-405.
148. Zhu W, Zheng D, Wang D, Yang L, Zhao C, Huang X. Emerging Roles of Ubiquitin-Specific Protease 25 in Diseases. Frontiers in Cell and Developmental Biology. 2021;9.
149. Bosch-Comas A, Lindsten K, Gonzàlez-Duarte R, Masucci MG, Marfany G. The ubiquitin-specific protease USP25 interacts with three sarcomeric proteins. Cellular and Molecular Life Sciences CMLS. 2006;63(6):723-34.
150. Laing NG, Dye DE, Wallgren-Pettersson C, Richard G, Monnier N, Lillis S, et al. Mutations and polymorphisms of the skeletal muscle Î±-actin gene (ACTA1). Hum Mutat. 2009;30(9):1267-77.
151. Brzezniak LK, Bijata M, Szczesny RJ, Stepien PP. Involvement of human ELAC2 gene product in 3' end processing of mitochondrial tRNAs. RNA Biol. 2011;8(4):616-26.
152. Saoura M, Powell CA, Kopajtich R, Alahmad A, Al-Balool HH, Albash B, et al. Mutations in ELAC2 associated with hypertrophic cardiomyopathy impair mitochondrial tRNA 3'-end processing. Hum Mutat. 2019;40(10):1731-48.
153. Tobias, Kopajtich R, Freisinger P, Wieland T, Rorbach J, Thomas, et al. ELAC2 Mutations Cause a Mitochondrial RNA Processing Defect Associated with Hypertrophic Cardiomyopathy. The American Journal of Human Genetics. 2013;93(2):211-23.
154. Schroeder C, Navid-Hill E, Meiners J, Hube-Magg C, Kluth M, Makrypidi-Fraune G, et al. Nuclear ELAC2 overexpression is associated with increased hazard for relapse after radical prostatectomy. Oncotarget. 2019;10(48):4973-86.
155. Gazzo A, Raimondi D, Daneels D, Moreau Y, Smits G, Van Dooren S, et al. Understanding mutational effects in digenic diseases. Nucleic Acids Res. 2017;45(15):e140.
156. Wagner M, Skorobogatko Y, Pode-Shakked B, Powell CM, Alhaddad B, Seibt A, et al. Bi-allelic Variants in RALGAPA1 Cause Profound Neurodevelopmental Disability, Muscular Hypotonia, Infantile Spasms, and Feeding Abnormalities. The American Journal of Human Genetics. 2020;106(2):246-55.
157. Nance JR, Mammen AL. Diagnostic evaluation of rhabdomyolysis. Muscle & Nerve. 2015;51(6):793-810.
158. Schäffer AA. Digenic inheritance in medical genetics. J Med Genet. 2013;50(10):641-52.
159. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214-W20.
160. Greenlees R, Mihelec M, Yousoof S, Speidel D, Wu SK, Rinkwitz S, et al. Mutations in SIPA1L3 cause eye defects through disruption of cell polarity and cytoskeleton organization. Hum Mol Genet. 2015;24(20):5789-804.
161. Cianfanelli V, De Zio D, Di Bartolomeo S, Nazio F, Strappazzon F, Cecconi F. Ambra1 at a glance. J Cell Sci. 2015;128(11):2003-8.
162. Gambarotto L, Metti S, Chrisam M, Cerqua C, Sabatelli P, Armani A, et al. Ambra1 deficiency impairs mitophagy in skeletal muscle. Journal of Cachexia, Sarcopenia and Muscle. 2022.

163. Skobo T, Benato F, Grumati P, Meneghetti G, Cianfanelli V, Castagnaro S, et al. Zebrafish ambra1a and ambra1b Knockdown Impairs Skeletal Muscle Development. PLoS One. 2014;9(6):e99210.
164. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet. 2018;50(1):151-8.
165. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Correction: Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet Med. 2020;22(6):1129.
166. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. 2022;23(1).
167. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5.
168. Somekh J, Shen-Orr SS, Kohane IS. Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset. BMC Bioinformatics. 2019;20(1).
169. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1).
170. Sanford Kobayashi E, Batalov S, Wenger AM, Lambert C, Dhillon H, Hall RJ, et al. Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. Sci Rep. 2022;12(1).
171. Zhang A, Sun H, Yan G, Wang P, Wang X. Metabolomics for Biomarker Discovery: Moving to the Clinic. BioMed Research International. 2015;2015:1-6.
172. Roos A, Thompson R, Horvath R, Lochmüller H, Sickmann A. Intersection of Proteomics and Genomics to “Solve the Unsolved” in Rare Disorders such as Neurodegenerative and Neuromuscular Diseases. PROTEOMICS - Clinical Applications. 2018;12(2):1700073.
173. Dowling P, Murphy S, Ohlendieck K. Proteomic profiling of muscle fibre type shifting in neuromuscular diseases. Expert Review of Proteomics. 2016;13(8):783-99.
174. Zhang L, Lu Q, Chang C. Epigenetics in Health and Disease. Springer Singapore; 2020. p. 3-55.
175. Coppedè F. Epigenetics of neuromuscular disorders. Epigenomics. 2020;12(23):2125-39.

Annexe A : Les variants candidats communs au niveau ADN ARN de la patiente Z26

Variants	*Ref*	*Alt*	*HMZ*	*Protéine*	*Gènes*
chr1:169349823	G	A	0/1	p.R258H	BLZF1
chr1:201915330	C	A	0/1	p.V47L	LMOD1
chr1:217793225	C	T	0/1	p.D225N	GPATCH2
chr2:176812292	G	C	0/1	p.P85A	LNPK
chr4:6862791	G	C	0/1	p.E228Q	KIAA0232
chr6:5369210	C	A	0/1	p.P136H	FARS2
chr6:127469804	G	A	0/1	p.V37I	RSPO3
chr6:145103190	A	G	0/1	p.H2922R	UTRN
chr7:115889084	C	T	0/1	p.R42C	TES
chr8:17794873	C	G	0/1	p.F109L	PCM1
chr8:107782313	A	G	0/1	p.W36R	ABRA
chr9:106900424	G	A	0/1	p.R1132H	SMC2
chr10:50534135	G	A	0/1	p.R1182Q	C10orf71
chr10:72289028	A	G	0/1	p.K77E	PALD1
chr11:57464259	C	T	0/1	p.P346S	ZDHHC5
chr11:76253389	C	G	0/1	p.A897G	EMSY
chr13:75861080	G	A	0/1	p.R1186W	TBC1D4
chr14:21859651	C	T	0/1	p.E2346K	CHD8
chr16:1812884	G	A	0/1	p.R585H	MAPK8IP3
chr16:11941607	T	C	0/1	p.D101G	RSL1D1
chr17:26948488	C	T	0/1	p.R1662H	KIAA0100
chr19:40880407	G	A	0/1	p.C300Y	PLD3
chr21:35186357	C	T	0/1	p.P898L	ITSN1

Annexe B : Les variants candidats communs au niveau ADN ARN du patient BC02

Variants	*Ref*	*Alt*	*HMZ*	*Protéine*	*Gènes*
chr1:147131132	G	A	0/1	p.R168W	ACP6
chr2:85596901	C	T	0/1	p.S103F	ELMOD3
chr2:152582014	T	C	0/1	p.T119A	NEB
chr2:166769088	G	A	0/1	p.P753L	TTC21B
chr3:52233295	G	A	0/1	p.R13Q	ALAS1
chr3:184102934	G	A	0/1	p.G576S	CHRD
chr3:195013022	G	A	0/1	p.A642V	ACAP2
chr4:123192841	G	T	0/1	p.G2721V	KIAA1109
chr4:156643302	G	A	0/1	p.C610Y	GUCY1A3
chr5:42699994	G	C	0/1	p.D148H	GHR
chr6:99353380	G	C	0/1	p.S342C	FBXL4
chr6:127765360	C	T	0/1	p.R660H	KIAA0408
chr6:152464847	C	T	0/1	p.D522N	SYNE1
chr9:135526484	T	A	0/1	p.D323V	DDX31
chr12:107360943	G	A	0/1	p.D17N	TMEM263
chr12:109604770	G	A	0/1	p.R253H	ACACB
chr14:36008779	C	T	1/1	p.R2075H	RALGAPA1
chr14:65239614	C	T	0/1	p.R1746Q	SPTB
chr15:48738953	T	C	0/1	p.N1913S	FBN1
chr17:4448967	G	A	0/1	p.R671W	MYBBP1A
chr17:12897799	A	G	0/1	p.I644T	ELAC2
chr17:28011678	G	C	0/1	p.Q108E	SSH2
chr19:11323895	G	A	0/1	p.A1483V	DOCK6
chrX:129185946	A	G	1/1	p.K1603R	BCORL1

Annexe C : Les variants candidats transcriptomiques sorties de l'analyse pour la patiente Z26

	Variants	*Ref*	*Alt*	*HMZ*	*Protéine*	*Gènes*	*L2FC*	*pvalue*	*qvalue*
Candidats vs SpliceAI	chr1:169349823	G	A	0/1	p.R258H	BLZF1			
	chr1:201915330	C	A	0/1	p.V47L	LMOD1			
	chr1:217793225	C	T	0/1	p.D225N	GPATCH2			
	chr2:176812292	G	C	0/1	p.P85A	LNPK			
	chr4:6862791	G	C	0/1	p.E228Q	KIAA0232			
	chr5:176522675	C	T	0/1	p.S551F	FGFR4			
	chr6:5369210	C	A	0/1	p.P136H	FARS2			
	chr6:145103190	A	G	0/1	p.H2922R	UTRN			
	chr8:107782313	A	G	0/1	p.W36R	**ABRA**	-2.57	0.01	0.10
	chr9:106900424	G	A	0/1	p.R1132H	SMC2			
	chr10:50534135	G	A	0/1	p.R1182Q	C10orf71			
	chr10:72289028	A	G	0/1	p.K77E	PALD1			
	chr11:7509386	C	T	0/1	p.T53M	**OLFML1**	2.43	0.02	0.13
	chr11:57464259	C	T	0/1	p.P346S	ZDHHC5			
	chr12:70965843	C	T	0/1	p.S648N	PTPRB			
	chr13:75861080	G	A	0/1	p.R1186W	TBC1D4			
	chr14:21859651	C	T	0/1	p.E2346K	CHD8			
	chr14:103932417	G	C	0/1	p.G217R	MARK3			
	chr16:1812884	G	A	0/1	p.R585H	**MAPK8IP3**	2.06	0.04	0.24
	chr16:11941607	T	C	0/1	p.D101G	RSL1D1			
	chr17:26948488	C	T	0/1	p.R1662H	KIAA0100			
	chr18:19153835	C	T	0/1	p.E324K	ESCO1			
	chr19:40880407	G	A	0/1	p.C300Y	PLD3			
	chr21:35186357	C	T	0/1	p.P898L	ITSN1			
	chrX:47085721	G	A	0/1	p.R353Q	CDK16			
Candidats vs DEG	chr8:107782313	A	G	0/1	p.W36R	**ABRA**	3.07	0.00	0.03
	chr9:125616333	C	G	1/1*	p.E1005D	RC3H2	2.43	0.02	0.13
	chr11:7509386	C	T	0/1	p.T53M	**OLFML1**	2.06	0.04	0.24
	chr16:1812884	G	A	0/1	p.R585H	**MAPK8IP3**	-2.36	0.02	0.14
	chr2:27716842	T	A	0/1	p.T137S	FNDC4	-2.40	0.02	0.14
	chr8:17794873	C	G	0/1	p.F109L	PCM1	-2.57	0.01	0.10
Candidats vs rMATS	chr2:61484452	A	T	0/1	p.F1960I	USP34			
	chr16:4414612	C	T	0/1	p.G290R	CORO7			
	chr14:72055953	G	C	0/1	p.C455S	SIPA1L1			
Autres	chr1:161176376	A	C	0/1	p.T128P	NDUFS2			
	chr1:161180487	G	C	0/1	p.D325H	NDUFS2			
	chr1:112524545	C	G	0/1	p.M268I	KCND3			
	chr13:110864935	T	C	1/1	p.N104D	COL4A1			

Annexe D: Les variants candidats transcriptomiques sorties de l'analyse pour le patient BC01

	Variants	*Ref*	*Alt*	*HMZ*	*Protéine*	*Gènes*	*L2FC*	*pvalue*	*qvalue*
Candidats vs SpliceAI	chr1:146758135	C	A	0/1	p.P523T	CHD1L			
	chr1:154941278	G	A	0/1	p.P200L	SHC1			
	chr5:31526792	G	A	1/1	p.P83L	DROSHA			
	chr12:93221765	G	C	0/1	p.L443V	EEA1			
	chr14:69521637	C	T	0/1	p.R588H	DCAF5			
	chr21:17250163	G	A	0/1	p.E950K	USP25			
	chr6:129636760	C	A	0/1	p.P1232H	LAMA2			
	chr22:35806779	C	G	0/1	p.I265M	MCM5			
	chr16:30012323	C	T	0/1	p.L120F	INO80E			
	chr7:128494922	G	A	0/1	p.R2331H	FLNC			
	chr17:76046849	C	T	0/1	p.P569L	TNRC6C			
	chr11:113675654	T	C	0/1	p.K506E	USP28			
Candidats vs DEG	chr1:21083673	A	G	0/1	p.S323P	HP1BP3	-2.39	0.02	0.09
	chr13:110864935	T	C	1/1	p.N104D	COL4A1	-2.06	0.04	0.15
	chr9:130868499	C	T	0/1	p.A324V	SLC25A25	-3.91	0.00	0.00
Autres	chr1:167971799	A	C	1/1	p.E297A	DCAF6			
	chr3:52485532	C	G	0/1	p.G110A	TNNC1			

Annexe E : Les variants candidats transcriptomiques sorties de l'analyse pour le patient BC02

	Variants	*Ref*	*Alt*	*HMZ*	*Protéine*	*Gènes*	*L2FC*	*pvalue*	*qvalue*
Candidats vs SpliceAI	chr1:147131132	G	A	0/1	p.R168W	ACP6			
	chr2:152582014	T	C	0/1	p.T119A	**NEB**	-2.14	0.03	0.16
	chr2:166769088	G	A	0/1	p.P753L	TTC21B			
	chr3:52233295	G	A	0/1	p.R13Q	ALAS1			
	chr3:184102934	G	A	0/1	p.G576S	CHRD			
	chr3:195013022	G	A	0/1	p.A642V	ACAP2			
	chr4:123192841	G	T	0/1	p.G2721V	KIAA1109			
	chr4:156643302	G	A	0/1	p.C610Y	GUCY1A3			
	chr5:42699994	G	C	0/1	p.D148H	GHR			
	chr6:99353380	G	C	0/1	p.S342C	FBXL4			
	chr6:127765360	C	T	0/1	p.R660H	**KIAA0408**	3.04	0.00	0.02
	chr12:49226215	A	G	0/1	p.S649P	DDX23			
	chr12:109604770	G	A	0/1	p.R253H	ACACB			
	chr14:36008779	C	T	1/1	p.R2075H	RALGAPA1			
	chr14:65239614	C	T	0/1	p.R1746Q	SPTB			
	chr15:48738953	T	C	0/1	p.N1913S	FBN1			
	chr17:4448967	G	A	0/1	p.R671W	MYBBP1A			
	chr17:49302426	G	A	0/1	p.P33S	MBTD1			
	chr19:11323895	G	A	0/1	p.A1483V	DOCK6			
	chrX:129185946	A	G	1/1	p.K1603R	BCORL1			
	chr2:85596901	C	T	0/1	p.S103F	ELMOD3			
Candidats vs DEG	chr2:152582014	T	C	0/1	p.T119A	**NEB**	-2.14	0.03	0.16
	chr6:127765360	C	T	0/1	p.R660H	**KIAA0408**	3.04	0.00	0.02
	chr9:103204491	G	C	0/1	p.E91Q	MSANTD3-TMEFF1	2.25	0.02	0.13
	chr9:127915822	A	G	0/1	p.V198A	PPP6C	-2.31	0.02	0.11
	chr2:218745633	C	T	0/1	p.D348N	TNS1	-3.85	0.00	0.00
	chr10:98744266	G	A	0/1	p.C1350Y	LCOR	-3.15	0.00	0.02
Candidats vs rMATS	chr2:152582014	T	C	0/1	p.T119A	NEB	-2.14	0.03	0.16
Autres	chr17:12897799	A	G	0/1	p.I644T	ELAC2			

Annexe F : Les variants candidats transcriptomiques sorties de l'analyse pour le patient HSJNM008

	Variants	*Ref*	*Alt*	*HMZ*	*Protéine*	*Gènes*	*L2FC*	*pvalue*	*qvalue*
Candidats vs SpliceAI	chr1:82456107	G	C	0/1	p.A1164P	ADGRL2			
	chr2:101622522	C	A	0/1	p.T112N	RPL31			
	chr2:113333172	C	T	0/1	p.R881C	POLR1B			
	chr2:170371458	A	C	0/1	p.K455Q	KLHL41			
	chr2:190924861	A	G	0/1	p.I225T	**MSTN**	-2.21	0.03	0.22
	chr3:132401637	A	G	0/1	p.L1241S	NPHP3			
	chr4:13605304	G	A	0/1	p.R1074W	BOD1L1			
	chr4:170671841	C	G	0/1	p.G82R	HPF1			
	chr6:17781508	T	G	0/1	p.Q1177P	KIF13A			
	chr6:74331606	G	A	0/1	p.S300F	SLC17A5			
	chr7:50513410	C	G	0/1	p.V526L	FIGNL1			
	chr9:71851924	A	G	0/1	p.D688G	TJP2			
	chr10:64953193	T	C	0/1	p.E1706G	JMJD1C			
	chr11:46563850	T	C	0/1	p.N573D	AMBRA1			
	chr11:46564927	G	C	0/1	p.P214A	AMBRA1			
	chr12:56718381	C	T	0/1	p.R571H	PAN2			
	chr12:72028035	A	G	0/1	p.S804P	ZFC3H1			
	chr12:110290475	G	A	0/1	p.T172M	GLTP			
	chr14:71502824	G	T	0/1	p.G1162C	PCNX1			
	chr17:3917677	C	T	0/1	p.G2760R	ZZEF1			
	chr17:80194070	G	C	0/1	p.W62C	**SLC16A3**	-8.72	0.00	0.00
	chr19:11323881	G	A	0/1	p.L1488F	DOCK6			
	chr22:37904575	C	T	0/1	p.V342M	CARD10			
	chr12:100478425	T	C	0/1	p.N373D	UHRF1BP1L			
	chr19:36215616	G	A	0/1	p.R1138Q	KMT2B			
Candidats vs DEG	chr2:190924861	A	G	0/1	p.I225T	**MSTN**	-2.21	0.03	0.22
	chr17:80194070	G	C	0/1	p.W62C	**SLC16A3**	-8.72	0.00	0.00
	chr14:23755149	A	C	0/1	p.W5G	HOMEZ	-2.22	0.03	0.21
	chr5:88024434	T	A	0/1	p.S270C	MEF2C	2.39	0.02	0.16
	chr12:56975195	C	T	0/1	p.P212L	RBMS2	2.21	0.03	0.22
	chr15:86697673	G	A	0/1	p.G92E	AGBL1	2.09	0.04	0.25