Université de Montréal

Caractérisation de variants génétiques pour estimer la prévalence de Niemann-Pick type C au Québec

Par

Marjorie Labrecque

Département de biochimie et médecine moléculaire, Faculté de médecine

Mémoire présenté en vue de l'obtention du grade de Maîtrise (M.Sc.)

en bio-informatique

Juillet 2022

Université de Montréal

Département de biochimie et médecine moléculaire, Faculté de médecine

---

Ce mémoire intitulé

Caractérisation de variants génétiques pour estimer la prévalence de Niemann-Pick type C au Québec

Présenté par

Marjorie Labrecque

A été évalué par un jury composé des personnes suivantes

Sébastien Jacquemont
Président-rapporteur

Martine Tétreault
Directeur de recherche

Sarah Gagliano-Taliun
Membre du jury

Résumé

La maladie de Niemann-Pick type C (NP-C) est une maladie autosomal récessive rare neurodégénérative, pan-ethnique et avec variabilité phénotypique. La forme classique se trouve chez les patients juvéniles, mais des patients de tous les âges existent. Les symptômes incluent des signes viscéraux, moteurs et neurologiques. La maladie est causée par une mutation dans le gène *NPC1* ou *NPC2*. La prévalence mondiale se trouve à environ un cas par 100 000 naissances, mais varie beaucoup selon les populations. Pour cette raison, nous avons voulu identifier et classifier des variants qui se trouve dans la population québécoise pour faire une estimation de la prévalence de NP-C au Québec. Nous croyons que cette maladie neurodégénérative est sous-diagnostiquée.

Pour identifier le pool génétique de la population québécoise, nous avons utilisé une approche bio-informatique. À l'aide des données de séquençage des 1109 participants sains de la cohorte CARTaGENE, nous avons identifié des variants rares, ayant des fréquences alléliques inférieures à 1%, dans les gènes *NPC1* et *NPC2*. Les données de séquençage de l'ARN et d'exome ont été alignées, les variants ont été détectés et annotés avec différents scores de pathogénicité. Les variants ont ensuite été classifiés à l'aide des lignes directrices de l'ACMG.

À l'aide de notre pipeline bio-informatique, nous avons identifié 37 variants rares. Parmi ces variants, un, p.I1061T, a été classifié comme pathogénique comme il l'est dans d'autres bases de données et un, p.P543L, initialement classifié comme potentiellement pathogénique a été classifié comme pathogénique dans notre population. Le variant p.P543L est d'ailleurs possiblement une mutation fondatrice chez les Canadiens-Français. La prévalence mesurée à l'aide des fréquences alléliques de ces deux variants est de 0,61 cas par 100 000 naissances.

Cette étude a permis d'identifier deux variants pathogéniques dans une population saine, c'est-à dire sans maladie neurodégénérative connue. Nous avons ensuite pu estimer pour la première fois la prévalence minimale de NP-C au Québec. Les résultats suggèrent que NP-C est sous-diagnostiquée dans notre population. Avec ces informations, les méthodes de diagnostic pourront être ajustées pour accélérer la détection de NP-C au Québec et ainsi aider les patients en donnant accès au traitement disponible pour réduire les symptômes neurologiques.

Mots-clés : Niemann-Pick type C, bio-informatique, classification, prévalence, effet fondateur, Québec

Abstract

Niemann-Pick type C disease (NP-C) is a rare autosomal recessive neurodegenerative, pan-ethnic disease with heterogenous symptoms. The classical form mainly affects juvenile patients, but patients of varying ages exist. The main symptoms are visceral, motor and neurological. The disease is caused by mutations in the *NPC1* or *NPC2* gene. The worldwide prevalence is approximately one case per 100 000 births but varies between populations. Therefore, we wanted to identify and classify rare variants found in Quebec's population to estimate the prevalence of NP-C in this population. We hypothesized that NP-C is under-diagnosed in Quebec.

To determine the genetic pool of NP-C in Quebec's population, we used a bioinformatics pipeline. With the sequencing data of 1109 healthy individuals of the CARTaGENE cohort, we identified rare variants, with a minor allele frequency inferior to 1%, in the *NPC1* and *NPC2* genes. The sequencing data from RNA and exome sequencing was aligned and the variants were found and annotated with different pathogenicity scores. The variants were then classified using the ACMG guidelines.

Using our bioinformatics pipeline, we identified a total of 37 rare variants. In those variants, one, p.I1061T, was directly classified as pathogenic since it was classified as that in all databases. The other one, p.P543L, was initially classified as likely pathogenic, but we were able to reclassify it as pathogenic in our population. The p.P543L variant is possibly a founder mutation in the French-Canadian population. Next, we estimated the prevalence based on the allelic frequencies of those two variants in our cohort. We found a prevalence of 0,61 case per 100 000 births.

This study allowed us to identify two pathogenic variants in a healthy population, without known neurodegenerative disease. We were also able to estimate the first ever minimal prevalence for NP-C in Quebec. Our results suggests that NP-C is underdiagnosed in our population. With the information collected here, we would be able to adjust the diagnostic methods of NP-C in Quebec to then be able to help the patients by giving them access to the available treatment to reduce neurological symptoms.

Keywords : Niemann-Pick type C, bioinformatics, classification, prevalence, founder effect, Quebec

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

ACMG : Collège Américain de Génétique et Génomique Médical

AF : Fréquence allélique

CaG : CARTaGENE

CNV : Variant du nombre de copies

Exome-seq : Séquençage des exomes

FPKM : Fragments Par Million de Kilobase

Indels : Insertions et délétions

NFE : Européens Non-Finlandais

NP-C : Niemann-Pick type C

NPC1 : Maladie Niemann-Pick de type C1

NPC2 : Maladie Niemann-Pick de type C2

OR : Rapport de côte

RNA-seq : Séquençage de l'ARN

RPKM : Reads Par Million de Kilobase

SNV : Variant d'un seul nucléotide

VUS : Variant de signification incertaine

Remerciements

Tout d'abord j'aimerais remercier ma merveilleuse directrice de recherche, Dre Martine Tétreault. Merci de m'avoir fait confiance au tout début comme stagiaire durant ma deuxième année de baccalauréat en bio-informatique. Merci de m'avoir fait participer à autant de projets de collaboration qui m'ont permis d'évoluer et de nourrir ma passion pour la recherche. Et finalement merci de m'avoir épaulé et encouragé durant cette maîtrise faite à distance pendant la pandémie. Merci également à Julie Hussin pour ton parrainage et tes conseils pour l'écriture de ce mémoire.

Mes prochains remerciements vont à mes excellents co-auteurs, Lahoud Touma, Dre Claude Bhérer et Dr. Antoine Duquette. Vos idées et conseils, nos conversations stimulantes et notre entraide ont mené à deux excellents papiers qui me rendent extrêmement fière.

Merci aussi à toute l'équipe du labo Tétreault au CRCHUM. Malgré la distance, je savais que je pouvais tout de même compter sur vous. Un merci particulier à Éric Bareke qui a été mon premier mentor au laboratoire et qui m'a montré toute la base pour les analyses bio-informatiques. Merci à Annie Laplante, avec qui j'ai eu les conversations les plus inusitées et divertissantes. Un merci aussi à mon amie Valérie Triassi qui me supporte et m'encourage depuis le début du baccalauréat et avec qui j'ai eu la chance de faire ma maîtrise.

Je remercie également des êtres très chers dans ma vie, mes parents, Francine et Michel, mon frère Maxime, ma mamie Liane et mes amis, qui sans toujours comprendre ce que je faisais, ont toujours été mes plus grands supporteurs. Merci également à mon copain François, qui m'a beaucoup fait grandir comme personne durant ces deux années.

Merci au Département de Biochimie, programme de Bio-informatique, et la Faculté des Études Supérieures et Postdoctorales pour le soutien financier.

Finalement, merci aux membres du jury qui ont accepté de donner de leur temps pour évaluer ce travail.

Chapitre 1 – Introduction

Les troubles de stockage lysosomal sont un groupe d'environ 40 maladies qui sont causés par un manque d'enzymes dans les lysosomes ce qui engendre l'accumulation de molécules non dégradées dans les cellules (1). La première maladie à avoir été découverte dans ce groupe est la maladie de Tay-Sachs en 1881 (2). De nos jours, la maladie de stockage lysosomal la plus commune est celle de Gaucher, avec une prévalence de 1 sur 40 000 personnes (3). Dans la fin des années 1920, les scientifiques Albert Niemann et Ludwig Pick ont découvert un groupe de troubles autosomaux récessifs faisant aussi partie des troubles de stockage lysosomal. Ils ont donc nommé ces maladies Niemann-Pick (4). Plus tard, dans les années 1950, deux autres scientifiques, Crocker et Farber, ont découvert qu'il y avait une grande variabilité des symptômes (5). Cela a mené à la création de quatre sous-groupes de Niemann-Pick, soit les types A, B, C et D (6). Les types A et B partagent le même mécanisme, impliquant le gène *SMPD1*, encodant la sphingomyéline phosphodiestérase. Le type A atteint uniquement les jeunes enfants. Les principaux symptômes sont de l'hépatosplénomégalie et une grande dégénération du système nerveux central (7). L'hépatosplénomégalie est une enflement du foie (hépatomégalie) et/ou de la rate (splénomégalie). De l'hypotonie qui affecte les muscles, la motricité et les nerfs moteurs peut aussi se détecter dans les premiers mois et progresser rapidement par la suite. La majorité des patients avec le type A meurent avant l'âge de 3 ans (8). Dans le type B, le symptôme principal est l'hépatosplénomégalie. Comparé au type A, il n'y a pas d'altération au niveau du système nerveux central, mais il peut il y avoir des altérations aux poumons qui peuvent engendrer des complications (9). Dans le type B, l'âge d'apparition des symptômes est variable et certains patients peuvent même vivre jusqu'à l'âge adulte (7). Les maladies de Niemann-Pick type A et B sont causés par les mêmes mécanismes impliquant des mutations dans le gène *SMPD1*. À ce jour, seulement le type de mutation a été trouvé comme pouvant expliquer la différence de phénotype entre le type A et B (7). Elles sont donc généralement regroupées dans la catégorie de déficience à l'acide sphingomyélinase, car le gène *SMPD1* code pour la fabrication de l'enzyme acide sphingomyélinase. Pour leurs parts, les types C et D sont causés par des variants dans les gènes *NPC1* et *NPC2*, encodant le transporteur de cholestérol NPC intracellulaire 1 et 2 respectivement (4). Les types C et D sont identiques au niveau clinique. La particularité du type D est que les gens affectés sont originaires de la Nouvelle-Écosse et sont tous porteurs du même variant dans le gène *NPC1*, p.G992W (10). Aujourd'hui, le type D est intégré dans le type C. Dans le type C, l'évolution de la maladie est généralement plus

lente avec une implication du système nerveux et des symptômes viscéraux plus légers que dans les types A et B (4). Le type C sera étudié plus en profondeur dans ce mémoire.

1.1. La maladie de Niemann-Pick type C

1.1.1. Symptômes

La maladie de Niemann-Pick type C (NP-C) est décrite comme une maladie neuro-viscérale. Les symptômes viscéraux, si présents, précèdent normalement les signes neurologiques ou psychiatriques. Les signe viscéraux impliquent généralement le foie, la rate ou les poumons. L'évolution de la maladie varie selon l'âge d'apparition des premiers symptômes viscéraux. Les signes neurologiques peuvent arriver à différents moment par la suite. Les symptômes principaux présents à chaque groupe d'âge sont résumés dans la Figure 1. Ils peuvent arrivés dès la naissance ou jusqu'à un âge adulte avancé de 60 ans. Chaque groupe sera décrit ci-dessous.

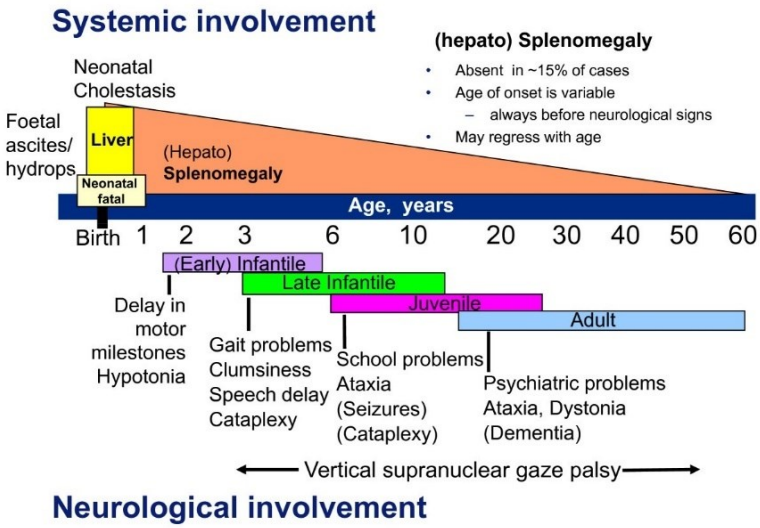

Figure 1. Symptômes neuroviscéraux de NPC (4)

Premièrement, la forme prénatale de NP-C est la plus rare et le pronostic est peu favorable (11). En effet, la majorité des cas meurent dans le premier mois suivant la naissance à la suite d'une hépatosplénomégalie et d'une insuffisance hépatique. Le groupe néonatal (0-2 mois) souffre généralement de jaunisse et de troubles dans le foie, qui sont critiques dans 10% des cas et meurent avant l'âge de 6 mois (4). Ils présentent également de l'hépatosplénomégalie. Le diagnostic n'est souvent pas fait avant que les signes neurologiques apparaissent, dans la phase infantile (12). Dans

cette prochaine phase, l'infantile précoce (2 mois à 2 ans), l'hépatosplénomégalie est toujours présente comme symptôme systémique. Pour ce qui est des signes neurologiques, les premiers à apparaître sont de l'hypotonie et des retards dans les étapes motrices, comme s'asseoir, rouler, attraper des objets ou ramper (4). La majorité des patients atteints de cette forme de NP-C meurent avant l'âge de 5 ans. La forme suivante est l'infantile tardif (2-6 ans). Les symptômes sont très semblables à la précédente. Des symptômes neurologiques s'ajoutent suite à l'hypotonie, entre autres, des problèmes de démarche, de la maladresse et de la cataplexie (4). La cataplexie se produit suite à des émotions fortes de ces patients et elle engendre un affaiblissement soudain des muscles. À ce stade, chez les patients juvéniles (6-15 ans), la splénomégalie est le principal symptôme indicatif. Les enfants commencent à avoir des problème à l'école, particulière pour parler et écrire, souvent un résultat d'ataxie (4). L'ataxie fait référence à des troubles de contrôle musculaire et de coordination. Elle est causée par une neurodégénération cérébrale, principalement des cellules de Purkinje (13). De plus, environ 50% des cas développent des convulsions de degrés variables (4). La catégorie des adolescents ou adultes (15 ans et plus) est plus atypique comme présentation de NP-C. Les premiers signes de la maladies sont majoritairement psychiatriques et peuvent inclure de la psychose, des symptômes bipolaires ou de la schizophrénie (14). Ces symptômes peuvent même se présenter plusieurs années avant les symptômes cognitifs et moteurs normalement associés à la maladie (4). Les deux troubles principaux sont l'ataxie et la paralysie supra nucléaire verticale du regard, présent chez 76% et 75% des patients adultes, respectivement (15). La paralysie supra nucléaire verticale du regard causé par une déficience cérébrale engendre une incapacité à faire des mouvements de yeux verticaux. La ressemblance de NP-C à d'autres maladies neurodégénératives ou psychiatriques peut mener à des grands délais dans le diagnostic de la maladie qui réduit considérablement le temps durant lequel un traitement pourrait être fait (12).

1.1.2. Causes et mécanismes

Comme mentionné plus haut, NP-C est causé par des variants dans deux gènes, *NPC1* (type C1; MIM 257220) ou *NPC2* (type C2; MIM 607625) et qui causent leur perte de fonction. La majorité des cas de NP-C, 95%, sont dus à des variants dans *NPC1* alors que le reste est dû aux variants dans le gène *NPC2* (4). Le variant le plus commun dans *NPC1* est p.I1061T, présent dans environ 20-25% des cas en France et en Angleterre. Sous forme homozygote, le variant engendre beaucoup de trouble de transport du cholestérol et est associé à la forme classique juvénile (4). Sous la forme hétérozygote, les patients développaient des symptômes neurologiques à partir de

l'enfance tardive, de l'âge juvénile ou encore adulte (16). Le second variant le plus commun, p.P1007A, est associé, pour sa part, avec la forme adulte de la maladie. Les variants du gène *NPC2* sont généralement associés avec des symptômes plus sévères. Il y a moins de variants connus, mais p.E20X est le plus fréquent (17).

Le gène *NPC1* code pour la protéine transmembranaire NPC1 située dans les endosomes tardifs et les lysosomes (18). Les endosomes tardifs sont des organites dans les cellules qui servent au transport membranaire. Ils permettent de trier les macromolécules avant qu'elles arrivent aux lysosomes où elles seront dégradées (19). Les lysosomes sont une autre sorte d'organite cellulaire. Les lysosomes digèrent les macromolécules à l'aide d'enzymes digestives (20). NPC1 contient 13 domaines transmembranaire et un domaine de détection du stérol, une sorte de lipide. La protéine NPC2, encodée par le gène *NPC2* est nettement plus petite que NPC1 avec 151 acides aminés comparé aux 1 278 acides aminés de NPC1 (21). NPC2 est située dans les lysosomes et est capable de se lier au cholestérol (21). Des analyses sur des souris mutantes *npc1-npc2*, mutantes *npc1* seulement ou *npc2* seulement ont démontrés que les protéines NPC1 et NPC2 participaient aux mêmes voies métaboliques de transport des lipides dans les endosomes et lysosomes (22). De plus, chaque protéine ne peut pas compenser pour la perte de l'autre. Malgré le fait que les deux protéines peuvent liées le cholestérol, NPC2 a une plus grande affinité et permet également à NPC1 de s'y lier plus rapidement (23). Normalement, ces lipides dans les lysosomes seraient envoyés vers l'appareil de Golgi ou le réticulum endoplasmique pour être transformés en dérivés (Figure 2a) (18). Dans le cas d'une déficience en protéines NPC1 ou NPC2, il y aura une accumulation de lipides dans les endosomes tardifs et les lysosomes ainsi qu'un manque de lipides dans les autres organelles perturbant les prochaines réactions qui permettent au corps de bien fonctionner (Figure 2b) (18).

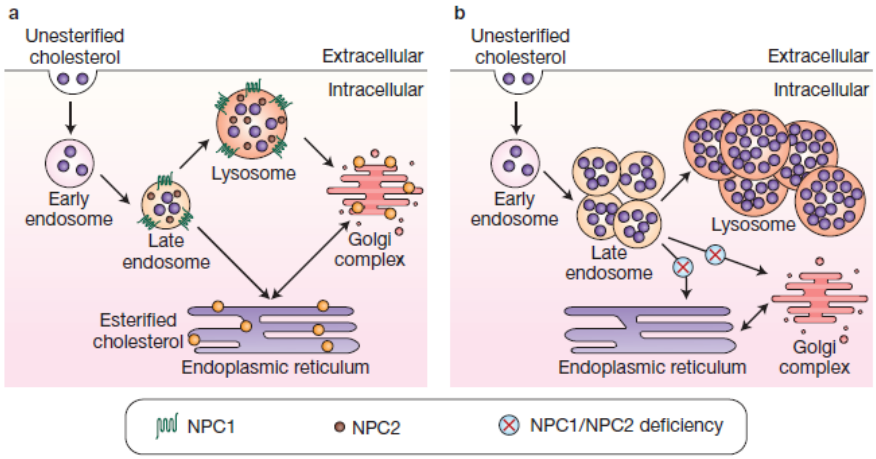

Figure 2. Troubles de transport de lipides dans NP-C (18)

Dans les organes, l'accumulation de cholestérol est plus proéminente dans la rate que dans le foie (4). Pour ce qui est du cerveau, l'accumulation est différente. En effet, les niveaux de cholestérols ne sont pas plus grands dans le cerveau (24). Il y a une autre sorte de lipides, les gangliosides, surtout les GM2 et GM3, qui s'accumulent dans le cerveau et le cortex cérébral (25, 26). Il est possible que NPC1 soit également responsable du transport des gangliosides GM2 et GM3 causant des problèmes de stockage neuronal (27). D'ailleurs, l'accumulation de GM2 est plus prononcée et rapide dans les cas de NP-C comparé à d'autres maladies lysosomales (28). Cela engendre la perte de neurones pouvant entraîner de la dégénération, particulièrement dans les cellules de Purkinje du cervelet (29). Le mécanisme de ce processus serait semblable à celui du transport du cholestérol endosomal et lysosomal qu'on retrouve dans les fibroblastes (30), mais demeure ambigu encore. Une particularité de NP-C se trouve dans les patients avec des variants hétérozygotes. Ils ont généralement des accumulations de cholestérols plus légères, tout de même observables (4).

1.1.3. Diagnostic et traitement

Le test spécifique biochimique qui permet de déterminer s'il s'agit de NP-C, dans la majorité des cas, est le test de la filipine. La filipine est un composé chimique qui se lie au cholestérol non-estérifié. Ainsi, des fibroblastes de patients sont cultivés et mélangés avec de la filipine. Le résultat peut être analysé avec un microscope à fluorescence. En cas de résultat positif, de la fluorescence sera observée dans des vésicules qui sont remplie de cholestérol (4). Il est également possible qu'il y ait un changement moyen pour des patients hétérozygotes. Un autre

option pour détecter la présence de NP-C est à l'aide de biomarqueurs. Cette option est plus rapide, moins invasive et à plus faible coût (31). Un type de biomarqueur qui est très utilisé sont les oxystérols. Les oxystérols sont des dérivés du cholestérol obtenu à l'aide d'une réaction d'oxydation avec des enzymes. Une limite à cette technique est que les niveaux d'oxystérols peuvent également être élevés dans Niemann-Pick de type A et B (32). Ainsi, pour confirmer les résultats de la filipine ou des biomarqueurs, une caractérisation clinique ou un test génétique peut être effectué. Dans la caractérisation clinique, il est important de vérifier les symptômes neurologiques soient la force des muscles ou les réflexes moteurs, pour détecter de l'ataxie, de l'hypotonie ou de la cataplexie (33). Le prochain test est ophtalmologique, pour détecter de la paralysie supra nucléaire verticale du regard, un des plus grands symptômes associé à NP-C (4). Un laboratoire a développé un outil pour faciliter la détection de NP-C dans des patients, l'indice de suspicion (« The Suspicion Index » en anglais) (34). Cette technique permet de trouver un score de risque de prédiction à partir des caractérisations cliniques viscérales, neurologiques et psychiatrique. Le risque peut être bien déterminé pour les enfants en haut de 4 ans, possiblement dû au manque de symptômes neurologiques dans les enfants en bas de 4 ans (35). L'indice de suspicion peut également être élevé pour d'autres maladies neurologiques, tel que la maladie d'Alzheimer et donc d'autres tests doivent être fait pour soutenir le diagnostic (35). L'autre façon de vérifier le diagnostic à la filipine ou le résultat de biomarqueur est à l'aide du séquençage des gènes *NPC1* et *NPC2*. Le séquençage peut par contre être dispendieux considérant que *NPC1* possède 25 exons et est très polymorphique et *NPC2* a 5 exons. Au total, il y a plus de 700 variants pour *NPC1* et 23 pour *NPC2*. À l'aide du séquençage, 90% des cas de NP-C peuvent être diagnostiqué (31). Autrement, il est possible de prioriser les deux variants les plus communs, p.I1061T et p.P1007A, mais cela ne garantit pas que la cause génétique sera identifiée (36). Avec toutes ces méthodes, le délai pour obtenir le bon diagnostic prendrait jusqu'à 6,2 ans chez les adultes (15). Le diagnostic peut donc être difficile particulièrement dans les cas avec présentation psychiatrique ou atypique.

Une fois que le patient est diagnostiqué, il est possible de commencer la prise à charge. Pour gérer les symptômes systémiques comme les convulsions, la cataplexie et la contraction musculaire, différents médicaments existent (4). Par contre, plus la maladie avance, moins les médicaments peuvent s'avérer efficaces. Il existe une thérapie pour les symptômes neurologiques qui est typiquement pour la maladie de Gaucher type 1, mais également approuvé pour NP-C (37).

Il s'agit du miglustat, qui permet de réduire l'accumulation de glycosphingolipides, ce qui permet de ralentir l'avancement des symptômes neurologiques et ainsi prolonger la durée de vie des patients (38). Comme les adultes ont le plus de symptômes neurologiques, ce sont ceux qui réagissent le mieux au traitement au miglustat (39). Dans le même optique, il n'est pas recommandé d'utiliser ce traitement pour les patients ne présentant que des symptômes systémiques, car les effets secondaires sont plus importants que les bénéfices apportés (40). D'autres thérapies alternatives non acceptées encore sont HPβCD, qui permet de retirer le cholestérol accumulé, mais peut mener à une perte de l'ouïe (41) et HDACi, qui est un enzyme qui permettrait de réduire l'accumulation de cholestérol dans les endosomes et les lysosomes (42). Peu importe le moyen choisi, il est important d'avoir un suivi avec les patients durant le traitement.

1.2. Effet fondateur au Québec

Une façon de mieux comprendre la maladie est de mieux comprendre l'historique généalogique de la population étudiée. Particulièrement, l'effet fondateur qui peut modifier la prévalence ou bien mener à différents symptômes selon la population (43). L'effet fondateur est le résultat d'une nouvelle population qui colonise un nouveau territoire. Cela engendre donc un goulot d'étranglement où il y a une diminution de la taille de population initiale et donc une diminution du pool génétique (44). Le Québec est un bon exemple de ce phénomène. Le Québec que nous connaissons aujourd'hui comportant 8 579 350 de personnes (Bilan démographique du Québec Édition 2021, consultation le 19 avril 2022) provient de nombreuses étapes de migration. La première et la plus importante vague d'immigration provient de la France entre 1608 et 1760 (45). En effet, plus de 10 000 immigrants Français se sont installés dans le première ville, Québec, de la province de Québec, jusqu'à l'arrivée des Anglais. Pendant ce temps, il y a également eu plusieurs mouvement à travers même de la province. À la fin des années 1600, une partie de ces immigrants se sont installés dans la région de Charlevoix, au Nord de Québec. La troisième vague d'immigration a été de la région de Charlevoix vers le Saguenay-Lac-Saint-Jean, encore plus au Nord. Cette dernière vague a été faite dans les années 1830 (46). Tous les descendants de ces immigrants sont par la suite connus comme des Canadiens-Français ou des Québécois. Par la suite, la forte fécondation et la continuité de migration interrégionale a mené à l'augmentation des variants fondateurs et ainsi que certaines maladies héréditaires récessives ou dominantes au Québec (47). En effet, les allèles pathogéniques normalement rares dans d'autres populations, se retrouvent

en plus grande proportions dans la population québécoise. Une hypothèse serait que la consanguinité aurait participé à ce phénomène, mais en réalité, cela n'a pas contribué à la concentration d'allèles fondateurs. Par contre, comme le nombre de colonisateur était petit au départ il est évidemment qu'avec le temps des relations consanguines éloignées (plus de cinq générations) soient arrivées (48). Une étude (49) a montré que même si seulement 2% des fondateurs initiaux Français possédaient des variants pathogéniques, se serait suffisant pour expliquer les fortes fréquences de maladies dans les régions de Saguenay-Lac-Saint-Jean et Charlevoix de nos jours. Une autre étude a également démontré que les premiers immigrants auraient contribués à 80% du pool génétique présent (50). Aujourd'hui, les populations québécoises et françaises sont considérées comme indépendantes dues à l'isolation géographique (43). Un exemple de maladie qui est plus fréquente chez les Canadiens-Français que chez les Français est la dystrophie musculaire oculo-pharyngée, qui est de transmission autosomal dominante. La prévalence se trouve à 1 sur 1 000 et 1 sur 200 000 individus pour le Québec et la France, respectivement (51, 52). Un autre exemple est l'ataxie de Friedreich, qui est une maladie bien commune. Sauf qu'au Québec, malgré le fait que la prévalence n'est pas connue, une étude a trouvé qu'un seul couple arrivé en 1634 de France aurait mené à 40 cas de la maladie 12 générations plus tard dans 14 familles non reliées (53). L'avantage d'étudier la population québécoise est que durant toutes les étapes de migration, l'information des naissances, morts, mariages, déplacements étaient bien documentée (54). Ces informations ont été mises dans une base de données informatisée, BALSAC, accessible à tous (55). Des études utilisant BALSAC ont été faites dans le but de mieux comprendre l'origine et l'évolution de maladies dû aux mouvements démographiques et de la généalogie (46). Il est donc plus facile aujourd'hui de faire des études sur la génétique des populations au Québec et beaucoup de choses restent à être explorés pour mieux comprendre les maladies rares (43).

Une autre ressource importante pour des études génétiques sur les québécois est la base de données CARTaGENE (CaG) (56). Il s'agit de la plus grande cohorte prospective au Québec. Elle regroupe des hommes et des femmes âgés entre 40 et 69 ans sélectionnés de façon aléatoire de quatre grandes région du Québec; Saguenay, Québec, Sherbrooke et Montréal. Les premiers participants ont été recrutés en 2009-2010, ils étaient au total 19 996. Des échantillons de sang ont été collectés pour ces individus ainsi que des informations sur la santé et des résultats de tests physiologiques. L'information généalogique de BALSAC est disponible pour 5 125 d'entre eux

(56). Depuis ce temps par contre, aucun suivi médical a été fait. La cohorte CaG comporte maintenant plus de 43 000 individus (57). La cohorte a initialement été conçue pour étudier des facteurs génétiques et environnementaux de maladies et regroupe de individus sains qui se rapproche de la situation socio-démographique du Québec en entier (56). Il y a tout de même des différences qui pourraient affecter le calcul de risque de maladies pour certains groupes. Cette initiative a réussi à améliorer la compréhension de nombreuses maladies dans la population québécoise depuis sa création et plus de 100 études utilisant les données CaG ont été publiés à ce jour (https://cartagene.qc.ca/fr/chercheurs/projets-et-publications, consultation le 20 avril 2022). Nombreuses de ces publications utilisent des approches bio-informatiques pour faciliter leur recherche.

1.3. Techniques de séquençage et outils bio-informatique

La bio-informatique est définie au sens large comme « l'application de la recherche en informatique au progrès des connaissances dans les sciences de la vie » (Larousse, Consultation le 2 mai 2022). Dans le contexte de ce mémoire, la bio-informatique utilise les données de séquençage dans le but de mieux comprendre une maladie, en commençant par les variants qui la cause. Le séquençage de l'ADN de type Sanger a permis de séquencer le tout premier génome humain en 2001 après 13 ans d'efforts, plus de 200 scientifiques et une facture de 2,7 milliards de dollars (58). Par la suite, les méthodes de séquençage de nouvelle génération (NGS) sont arrivées en 2005 et ont continuer d'évoluer par la suite (59). Ainsi, en 2011, séquencer un génome entier coûte moins de 10 000 dollars et ne prend que quelque jours.

1.3.1. Séquençage d'exomes

Le premier type de NGS qui sera utilisé est le séquençage d'exome (exome-seq). Ce type de séquençage se concentre sur la partie codante du génome. Cette partie du génome qui code pour des protéines ne représente qu'environ 2% du génome total mais contient 85% des variants ayant un effet concret sur des maladies (59). L'exome-seq se fait en plusieurs étapes (60). La première constitue à séparer l'ADN en différents fragments (« reads » en anglais) pour construire une libraire. Des adapteurs sont ensuite ajoutés aux morceaux d'ADN. Comme seulement les exons sont importants ici, la librairie est enrichie avec des séquences qui correspondent aux exons à l'aide d'hybridation en phase aqueuse. Cela constitue à hybrider les fragments à des bouts d'ADN qui contiennent la séquence complémentaire aux adapteurs. Les fragments ayant été hybridés sont

conservés et amplifiés pour ensuite être séquencés. Par la suite, l'alignement peut être fait et les variants peuvent être identifiés. L'efficacité de l'hybridation dépend de la capacité à bien choisir les exons (59). Cette information provient de différentes bases de données et peut donc être incomplète car des exons n'ont pas été trouvés encore ou bien certains exons ne sont pas encore annotés. Un avantage de l'exome-seq est qu'il permet d'obtenir beaucoup d'informations pertinentes sans prendre trop d'espace. En effet, pour avoir une couverture moyenne de 30 fois le génome avec le séquençage complet du génome cela prendrait 30Gb de données alors qu'avec 3Gb de données d'exome on peut obtenir une couverture de 75 fois le génome (61, 62). De plus, il a été montré que l'exome-seq peut identifier la cause d'une maladie même en présence d'hétérogénéité phénotypique ou génomique, ce que les études d'association ne sont pas capable de faire (63-65). La stratégie pour identifier des variants pathogéniques à l'aide du exome-seq est donc de trouver tous les variants et ensuite appliquer différents filtres sur la fréquence allélique (« allelic frequency » en anglais; AF), le type de mutation en priorisant ceux qui affectent la protéine : variants non-synonymes, insertions ou délétions (indels) ou variants qui modifie l'épissage (59). D'ailleurs, pour les mutations non-synonymes, un changement dans l'acide aminé d'une protéine a été trouvé à représenter la moitié des causes de maladies génétiques (66). Pour l'AF, les variants à faible fréquence (entre 1% et 5%) et ceux rares (inférieure à 1%) sont supposés avoir un plus grand effet dans le contexte de maladies rares (59). Une autre méthode pour filtrer est à l'aide d'outils d'analyse qui se basent sur la conservation de la séquence et l'effet du changement d'acide aminé. Ces outils seront vus plus en profondeur plus tard. Un autre type de variant qui est moins étudié, qui peut être identifié sont des variants du nombre de copies (« copy number variants » en anglais; CNV). Ces évènements proviennent généralement d'une duplication d'un fragment de chromosome entre 1 kb et 5 Mb (67). Ces variants structuraux ne sont généralement pas inclus dans les pipelines d'analyses bio-informatiques classiques, mais pourraient aider à expliquer le phénotype de certaines maladies, dont des maladies neurodégénératives (68). De nombreux exemples de la littérature prouvent que l'exome-seq est un bon moyen pour identifier des variants associés à des gènes de maladie récessive et aussi dominante dans un contexte clinique pour aider au diagnostic (60).

1.3.2. Séquençage d'ARN

Le séquençage de l'ARN (RNA-seq), permet d'étudier le transcriptome au complet et pas seulement les exons comme c'est le cas de l'exome-seq. Le transcriptome correspond à tous les

types d'ARN présents dans la cellule. Il est d'ailleurs important lors du séquençage de retirer l'ARN ribosomal qui correspond à 90% de l'ARN total et qui est généralement moins important dans les études normales qui s'intéressent à l'ARN messager (69). Ceci peut être faite de deux façon, soit par une capture poly-A qui enrichie pour l'ARN messager ou en éliminant spécifiquement les ARN ribosomaux et donc conserve les ARN messagers ainsi que certains ARN non-codant. Le RNA-seq permet principalement à identifier des nouveaux transcrits ou encore à mesurer l'expression des gènes en utilisant deux méthodes différentes (70). Pour le premier but, la méthode utilisée est d'assembler des transcrits *de novo* du génome de référence pour ensuite y aligner les fragments d'ARN de notre échantillon. Pour le deuxième, la façon la plus simple est d'aligner les fragments à un génome de référence déjà construit et disponible en prenant en compte l'épissage et ensuite de reconstruire les transcrits en utilisant les données de position obtenues lors de l'alignement. Un autre élément à considérer en réfléchissant à l'étude est la longueur des fragments, un séquençage à une extrémité (« single-end » en anglais) ou apparié (« paired-end » en anglais) et la profondeur de séquençage. Par exemple, plus les fragments sont longs, plus la couverture risque d'être meilleure, le séquençage apparié peut aider à mieux aligner des fragments ambigus, mais cela ne veut pas dire que c'est nécessaire pour une simple analyse d'expression différentielle (71, 72). Pour ce qui est de la profondeur, de façon générale plus le séquençage est profond plus il sera facile d'identifier des transcrits ou de les quantifier, mais cela fait aussi que plus de bruit pourrait être détecter rendant l'analyse plus difficile (73). En moyenne, entre 10-30 millions de fragments par échantillons sont nécessaires, mais une autre étude suggère qu'un million serait suffisant pour des analyses d'expression différentielle (69, 74). Une fois ces détails décidés, un contrôle de qualité doit être fait sur les données brutes avant de commencer l'alignement et l'assemblement. Cela peut comprendre la qualité de la séquence, la présence de séquences dupliquées, la valeur de GC, la présence d'adaptateurs ou d'artéfacts de PCR (69). L'outil FastQC est très reconnu pour aider à évaluer la qualité des données provenant du séquençage autant avant l'alignement qu'après (75). Après le séquençage, un fichier FASTQ est fourni pour chaque échantillon et ce fichier contient tous les fragments. Le but est donc d'assigner ces fragments à une position sur le génome (71). L'alignement peut être fait de deux façons différentes comme vu plus haut, avec un génome de référence existant ou avec un transcriptome maison annoté. Durant cette étape, les fragments peuvent être mapper de façon unique, à une seule position ou à plusieurs positions (multi-mappé) (69). Comme les fragments multi-mappés sont plus difficile à analyser au

niveau biologique et peuvent causer des biais, ils sont généralement retiré avant les étapes subséquentes (71). Une autre mesure du contrôle de qualité est faite après l'alignement, pour s'assurer que le pourcentage d'alignement soit suffisamment élevé, idéalement supérieur à 70% indique un bon alignement (69, 76). La prochaine étape est la quantification des transcrits. Une fois que les fragments sont associés à une position sur le génome, il est possible de l'associé à un gène ou un transcrit (71). Une fois que tous les fragments sont associés, un outil peut mesurer le nombre total de fragments qui est mappé à chaque gène ce qui correspond à l'expression. Des études ont démontré que le choix le plus important lors d'une analyse de RNA-seq est celui de l'outil de quantification (77, 78). C'est en effet celui qui aura le plus de répercussion sur les résultats finaux, encore plus que le choix de l'outil pour l'alignement (79). Deux outils fréquemment utilisés pour le compte de fragments sont HTSeq (80) et featureCounts (81). Ces outils donnent le résultat sous la même forme, une matrice de compte pour les gènes ou transcrits sur chaque ligne et les échantillons comme colonnes. Pour une analyse qui compare le même gène entre deux conditions différentes, une normalisation selon la taille du gène n'est pas nécessaire, mais autrement la longueur du gène peut apporter un biais dans la comparaison (69). Pour cette raison, avant de faire des analyses d'expression différentielle, une normalisation pour chaque échantillon est effectuée à l'aide d'une conversion comme le RPKM (« reads per kilobase million » en anglais) qui est normalement pour le séquençage unique ou FPKM pour le séquençage apparié, car il s'agit de fragments à la place de reads. Ceci permet de contrôler pour la longueur des gènes et la longueur de la librairie utilisée. De nos jours, il est recommandé de normaliser en prenant compte des changements entre les échantillons également, comme en prenant la médiane (82). Cette normalisation est effectuée par différents outil bio-informatique offerts et permettent généralement de faire l'expression différentielle par la suite. DESeq2 (83) est un exemple, mais le choix de l'outil pour mesurer l'expression différentielle a moins d'influence sur les résultat que le reste des outils vu plus haut (71). L'expression différentielle correspond à comparer les valeurs d'expression des gènes entre les échantillons, cas versus contrôle par exemple, pour identifier lesquels sont plus exprimés dans quel condition (69). En plus des variants que l'on peut retrouver aussi dans l'exome-seq, il est possible d'identifier des variants d'épissage dans le RNA-seq. L'épissage se produit lorsque les introns de l'ARN pré-messager sont retirés et les exons sont joint ensemble pour donner de l'ARN messager mature (84). Durant ce processus, il est possible qu'il y ait différent transcrits qui se forment pour un gène, ce qui mène par la suite à différentes protéines, nommées isoformes.

Il s'agit donc d'épissage alternatif (85). Il existe différents types d'épissage alternatif ce qui ajoute encore plus à la complexité et la variabilité. Un peu comme les CNVs, l'analyse de l'épissage alternatif font rarement partie du pipeline d'analyse normale, pourtant ils sont très pertinents à étudier dans le contexte de maladies rares (86). De plus, les cellules du système nerveux sont particulièrement sensibles à des évènements d'épissage ce qui peut souvent mener à des maladies neurodégénératives (84). Les variants d'épissage ainsi que ceux mentionnés précédemment peuvent par la suite être annotés et filtrés pour aider dans les analyses.

1.3.3. Outils d'analyse

Malgré l'avancement dans les technologies de séquençage, cela crée énormément de données et il est difficile par la suite d'interpréter les résultats. Prioriser quels variants sont plus importants dans le contexte de chaque maladie devient donc un élément clé pour faciliter les analyses. La priorisation peut être faite à l'aide de différentes méthodes, ou encore mieux, avec une combinaison de ces méthodes. Il est possible d'utiliser l'AF de chaque variant, le type de variation ou encore des scores provenant d'algorithmes informatiques. Différents scores existent dans le but de différencier des variants potentiellement neutres à ceux qui sont potentiellement pathogéniques et délétères. Certains outils se basent sur l'effet du changement de l'acide aminé pour déterminer si le variant d'un seul nucléotide (« single nucleotide variant » en anglais; SNV) est délétère, comme SIFT (87) et Polyphen-2 (88). D'autres se basent sur l'évolution et la conservation de variant pour déterminer leur effet, phastCons (89) et GERP (90) en sont deux exemples.

SIFT est un algorithme qui utilise la séquence de l'alignement pour faire la prédiction du variant selon le type de substitution. Par exemple, si dans l'alignement il y a un acide aminé hydrophobique à une position, SIFT va assumer qu'un changement pour un autre acide aminé hydrophobique sera toléré alors que si le changement est vers un acide aminé avec des propriétés différentes ne sera pas toléré et serait donc possiblement pathogénique (91). Ainsi, la position et le type d'acide aminé sont importants dans l'algorithme pour calculer le score. Le score résultant se trouve entre 1 et 0. SIFT considère qu'un score normalisé inférieur à 0.05 serait délétère alors qu'un score supérieur ou égal à 0.05 serait considéré comme toléré (87). Ceci ne représente qu'une suggestion de seuil et peut être modifié selon le contexte d'analyse. Par exemple, une étude a démontré qu'un seuil de 0.1 serait plus sensible pour détecter des variants délétères (92). PolyPhen-2 utilise quant à lui la structure en plus pour faire sa prédiction de l'effet des substitutions

d'acides aminés (93). L'algorithme se base sur la séquence (huit fonctionnalités prédictives), la phylogénétique et trois fonctionnalités prédictives de la structure (88). À l'inverse de SIFT, PolyPhen-2 donne un score de 0 à 1, correspondant à bénin et pathogénique, respectivement (93).

Autrement, une façon d'identifier des éléments importants est via les séquences conservées à travers les espèces (94). En effet, ce qui explique en grande partie la conservation à travers les espèces est la sélection naturelle qui permet de retirer les allèles délétères aussi appelé sélection négative (89). Les séquences orthologues restantes ont donc plus de chances d'avoir une fonction importante. Une autre explication serait le taux de mutation qui serait différent à travers le génome (95). Au total, il y aurait entre 3% et 8% du génome humain qui serait conservé à travers les mammifères vertébrés (89). PhastCons est un des outils qui permet d'identifier des séquences conservées. L'outil utilise un modèle de Markov caché basé sur la phylogénétique pour prédire quels éléments sont conservés (89). L'algorithme prend en compte le nombre de substitution qui peuvent se produire selon la longueur de la branche phylogénétique ainsi que le type de substitution, par exemple que la transition est plus fréquente et que la transversion est plus dommageable. Le score résultant, entre 0 et 1, représente la probabilité que la position étudiée se trouve dans un élément conservé. Un score maximal de 1 représentant que le variant est très conservé à travers les espèces. GERP est un autre exemple d'outils qui annote des variants selon la conservation de la séquence (90). Comparé à phastCons, GERP utilise une méthode qui détecte les régions où il y a un manque de substitution ou un taux de substitution plus faible que la normale, nommé rejet de substitutions (« rejected substitutions » en anglais). Le rejet de substitutions est donc une mesure de la sélection négative qui a eu lieu dans le passé et se traduit en un score qui varie entre -12.3 et 6.17, le dernier signifiant l'élément le plus conservé.

Plus récemment, un outil nommé CADD a été fait pour intégrer différentes annotations et prioriser les variants (96). Effectivement, pour faire l'annotation, CADD utilise les données du prédicteur d'effet de variant d'Ensembl (« Ensembl Variant Effect Predictor » en anglais) (97), du projet Encode (98), du navigateur de génome UCSC (« USCS Genome Browser » en anglais) (99) ainsi que des scores de conservation ou de séquence comme SIFT, PolyPhen-2, phastCons et GERP précédemment décrits. L'intégration est faite à l'aide d'apprentissage machine supervisé avec une machine à vecteurs de support (« support vector machine » en anglais). Le score résultant varie entre 1 et 99. Dans les 10%, le score normalisé est à 10, à 1% le score est de 20, à 0.1% le score

est de 30 et ainsi de suite. L'équipe de CADD suggère d'établir un seuil de 15 pour détecter des variants délétères classifiés de pathogéniques (96). Ils préviennent aussi qu'en sélectionnant un seuil trop stringent il pourrait y avoir une perte de variants pertinents dans le contexte de l'étude (100). Ils mentionnent également que le seuil dépend de la sévérité du phénotype étudié en plus du mode de transmission de la maladie, récessif ou bien dominant. Un autre moyen existe aussi pour classifier des variants, qui inclut des données d'algorithmes vu plus haut.

Le Collège Américain de Génétique et Génomique Médical (« American College of Medical Genetics » en anglais; ACMG) est une organisation qui a établi un guide de standards et de bonnes pratiques à utiliser pour interpréter des variants. Initialement établies en 2000 (101), puis révisées en 2007 (102) et rerévisées en 2015 (103), ces lignes directrices permettent d'uniformiser la terminologie et faciliter la classification de variants à travers les laboratoires de tests génétiques cliniques. L'évaluation de ces variants est faite pour des patients soupçonnés d'avoir une maladie mendélienne, mais peuvent tout de même être utilisé sur des individus sains (103). Pour les individus sains par contre, il faut faire attention et augmenter les exigences pour classifier un variant de pathogénique. La terminologie suggérée par l'ACMG est d'utiliser les termes « pathogénique », « potentiellement pathogénique », « signification incertaine » (VUS), « potentiellement bénin » et « bénin » pour caractériser un variant (103). Pour les termes potentiellement bénin et potentiellement pathogénique, c'est que la certitude est plus grande que 90%. Ils recommandent aussi d'utiliser le terme variant à la place de mutation et polymorphisme, puisque ceux-ci représentent respectivement un changement permanent à la séquence nucléotidique et un variant avec une AF supérieure à 1%. Pour évaluer un variant, il est important de se baser sur la littérature, les bases de données et des algorithmes pour obtenir plus d'informations. La littérature peut indiquer dans quel contexte un variant a déjà été vu, les bases de données peuvent fournir des informations sur les fréquences du variant dans différentes populations et les algorithmes peuvent aider dans la classification. De façon concrète, l'ACMG utilise trois grilles de critères, la première pour les variants pathogéniques, la deuxième pour les variants bénins et la dernière pour combiner les deux précédentes (103). Sans rentrer dans tous les détails, les critères de classification se basent sur les données populationnelles, sur les scores d'algorithmes informatiques, sur l'effet sur la fonctionnalité, sur les données de ségrégation, *de novo* et alléliques ainsi que sur les bases de données (103). Il peut être pertinent de se baser sur les données cliniques aussi, lorsque disponible, pour améliorer la caractérisation. Une étude de 2017

(104) a été faite pour comparer l'efficacité de l'utilisation de différents algorithmes avec les lignes directrices de l'ACMG. Ils ont trouvé qu'avec 18 algorithmes ils identifient à 39.2% des variants pathogéniques avec un taux de faux positifs de 0.8% alors qu'avec uniquement Polyphen-2, SIFT et CADD la concordance est de 84.9%, mais le taux de faux positifs augmente à 2.1%. Dans le futur, les nouveaux outils seront plus performants et pourront intégrer les anciens algorithmes pour améliorer leur efficacité. Par le fait même, cela permettra de faciliter la classification à l'aide des lignes directrices de l'ACMG. Ainsi, il sera plus facile de reclassifier les VUS identifiés vers un variant pathogénique ou bénin selon les critères. Le développement d'outils reste une priorité pour l'annotation et la classification de variants et permettra de profiter de toutes les informations obtenus par le séquençage, autant d'exome que d'ARN.

1.4. Introduction aux articles

Comme nous avons pu voir, NP-C est une maladie complexe, plus ou moins bien comprise, surtout chez les adultes où les symptômes sont atypiques. Il existe pourtant un traitement efficace pour ce groupe d'âge afin de diminuer les symptômes neurologiques. Pour y avoir accès, le diagnostic doit d'abord se faire rapidement. Ce qui n'est pas le cas, puisqu'il y a généralement un délai d'environ 6 ans avant qu'un patient obtienne le bon diagnostic (15). De plus, avec l'effet fondateur qui se retrouve au Québec, il est possible qu'il y ait des variants fondateurs associés à la maladie ou bien avec une prévalence plus élevée. Pour ces raisons, nous avons voulu mieux comprendre l'état de NP-C au Québec. Cet ouvrage présente deux articles en lien avec cette maladie neurodégénérative rare. Le premier intitulé « Identification et Classification de Variants Rares dans *NPC1* et *NPC2* au Québec » et le second : « Estimation de la prévalence de Niemann-Pick type C au Québec ».

1.4.1. Premier article

Nous savons déjà que de nombreux variants sont associés avec NP-C et que p.I1061T est le plus commun (36). Mais au Québec nous avons aucune idée du pool génétique qui est présent. Comme le titre l'indique la première étape a pour but de mieux comprendre le profil de NP-C au Québec et d'identifier et classifier les variants qui se retrouvent dans la population. Pour nous aider, nous avons utilisé les données de CaG, comprenant 1109 individus (911 RNA-seq et 198 exome-seq). Notre hypothèse est que même si les individus de CaG sont sains, nous allons identifier des variants rares et possiblement pathogéniques dans la population. Notre but est donc de les identifier

à l'aide d'outils bio-informatiques et de les classifier selon leur pathogénicité. Les données CaG brutes ont été analysées à l'aide du pipeline bio-informatique en place dans le laboratoire. Les données d'exome-seq ont été alignées avec l'outil BWA (105) alors que les données de RNA-seq ont été alignées avec STAR (76) et HISAT2 (106). Les variants ont ensuite été identifiés par GATK (107) et VarDict (108) pour exome-seq et RNA-seq, respectivement. L'annotation a finalement été faite avec ANNOVAR (109) où les différentes fréquences alléliques de plusieurs bases de données ont été ajoutées et les scores mentionnés plus haut ont également été ajoutés pour aider à la caractérisation de chacun des variants. Pour faciliter le filtrage des données, nous avons uniquement regardé les variants dans les gènes d'intérêt à NP-C, soit *NPC1* et *NPC2*. Comme il s'agit d'une première étude sur NP-C au Québec, nous nous sommes concentrés sur les variants non-synonymes, les indels et les variants d'épissages pour une idée globale du pool génétique. La classification a ensuite été faite à l'aide du guide de l'ACMG (103) et des filtres personnalisés selon les seuils des scores.

1.4.2. Deuxième article

La deuxième étude se concentre sur la prévalence de NP-C au Québec. Dans les études passées sur la maladie, quelques prévalences ont été ressorties. Dans les années 1990, alors que la maladie n'était pas encore bien comprise et les méthodes de diagnostic n'étaient pas au point, les Pays-Bas et l'Australie trouvaient comme prévalence de la maladie 0,35 cas et 0,47 cas par 100 000 individus, respectivement (110, 111). Au début des années 2000, la France ajuste sa prévalence de NP-C à 0,82 par 100 000 individus (4). Plus tard, le Portugal obtient une prévalence beaucoup plus élevé, de 2,2 cas pour 100 000 individus, alors que le variant p.I1061T n'est pas très commun dans la population (112). De son côté, la République Tchèque a une prévalence de 0,91 cas par 100 000 individus (113). En 2015, le Royaume-Uni estime que la prévalence est de 0,78 cas pour 100 000 individus (16). Aux États-Unis, un calcul a été fait à l'aide de bases de données et ils trouvent une prévalence de 1,12 cas par 100 000 individus (114). Mais plus récemment, dans le même pays, une prévalence aussi basse que 0,29 cas par 100 000 aurait été estimée (115). La détection et l'inclusion de cas prénataux restent un défi qui pourrait modifier le calcul de la prévalence (4). De plus, certains hétérozygotes qui développent des symptômes tardivement ne sont présentement pas inclus mais pourraient faire grimper les prévalences (16). À travers le Canada et au Québec, aucune prévalence n'a été calculée auparavant. Notre hypothèse et que la maladie de NP-C est présentement sous-diagnostiquée dans la population québécoise. En reprenant les mêmes données que dans la première

étude, nous nous sommes concentrés sur deux variants pathogéniques dans le but de faire une première estimation de la prévalence de NP-C au Québec à l'aide de l'équation de Hardy-Weinberg.

À la suite de ces deux études, il sera possible d'avoir une meilleure idée du profil génétique des québécois et québécoises et ainsi suggérer des méthodes pour améliorer le dépistage de nouveau-nés et le diagnostic de NP-C dans la population du Québec qui sont sous l'effet fondateur.

Chapitre 2 – Articles

2.1. Identification and Classification of Rare Variants in *NPC1* and *NPC2* in Quebec

Scientific Reports 11, 10344 (2021). https://doi.org/10.1038/s41598-021-89630-5

Received: 4 March 2021; Accepted: 29 April 2021; Published online: 14 May 2021

Lahoud Touma[1,4,5], Marjorie Labrecque[1,4,5], Martine Tetreault[1,4*] & Antoine Duquette[1,2,3,4*]

[1]Département de Neurosciences, Faculté de Médecine, Université de Montréal, CRCHUM – 900, rue Saint-Denis, Pavillon R, Montréal, QC H2X 0A9, Canada. [2]Service de Neurologie, Département de Médecine, Unité de Troubles du Mouvement André-Barbeau, Centre Hospitalier de L'Université de Montréal (CHUM), Montreal, Canada. [3]Service de Médecine Génique, Département de Médecine, Centre Hospitalier de L'Université de Montréal (CHUM), Montreal, Canada. [4]Centre de Recherche du Centre Hospitalier de L'Université de Montréal (CRCHUM), Montreal, Canada. [5]These authors contributed equally: L. Touma and M. Labrecque. *email: martine.tetreault@umontreal.ca; antoine.duquette@umontreal.ca

Contribution

Martine Tétreault et Antoine Duquette ont participés à la conception de l'étude. Marjorie Labrecque a effectué les analyses bio-informatiques : rouler le pipeline d'alignement et d'annotation, ainsi qu'identifier et filtrer les variants des gènes *NPC1* et *NPC2*. Des scripts maison ont été utilisés pour extraire les données d'intérêts. Lahoud Touma a caractérisé les variants selon les critères de l'ACMG. Marjorie Labrecque a écrit la section Matériel et Méthodes et a fait les figures et tables dans le manuscrit. Lahoud Touma a écrit l'Introduction, les Résultats et la Discussion. Tous les auteurs ont révisé le manuscrit.

2.1.1. Abstract

Niemann–Pick disease type C (NPC) is a treatable autosomal recessive neurodegenerative condition which leads to a variety of progressive manifestations. Despite most cases being diagnosed at a young age, disease prevalence may be underestimated, especially in adults, and

interpretation of *NPC1* and *NPC2* variants can be difficult. This study aims to identify potential pathogenic variants in a large cohort of healthy individuals and classify their risk of pathogenicity to assist with future interpretation of variants. The CARTaGENE (CaG) cohort was used to identify possible variants of *NPC1* and *NPC2*. Nine-hundred and eleven RNA samples and 198 exome sequencing were screened for genetic variants through a bio-informatic pipeline performing alignment and variant calling. The identified variants were analyzed using annotations for allelic frequency, pathogenicity and conservation scores. The ACMG guidelines were used to classify the variants. These were then compared to existing databases and previous studies of NPC prevalence, including the Tübingen NPC database. Thirty-two distinct variants were identified after running the samples in the RNA-sequencing pipeline, two of which were classified as pathogenic and 21 of which were not published previously. Furthermore, 46 variants were both identified in our population and with the Tübingen database, the majority of which were of uncertain significance. Ten additional variants were found in our exome-sequencing sample. This study of a sample from a population living in Quebec demonstrates a variety of rare variants, some of which were already described in the literature as well as some novel variants. Classifying these variants is arduous given the scarcity of available literature, even so in a population of healthy individuals. Yet using this data, we were able to identify two pathogenic variants within our population and several new variants not previously identified.

2.1.2. Introduction

Niemann–Pick disease type C (NPC) is a rare, autosomal recessive neurodegenerative condition which leads to a variety of progressive neurological and non-neurological manifestations. NPC is estimated to affect 1 in 100,000–120,000 live births (116). It is caused by mutations of the *NPC1* gene in 95% of cases while mutations of *NPC2* account for the other 5%. The spectrum of clinical presentation is wide with different onsets which have been classified as: perinatal (shortly before and after birth, 3–12% of cases), early infantile (3 months to < 2 years, 3–37% of cases), late infantile (2 to < 6 years, 21–39% of cases), juvenile (6 to < 15 years, 21–54% of cases), and adult (15 years and greater, 5–27% of cases) (4, 117). In the majority of cases, patients first develop liver, spleen or lung involvement with a high variability of disease severity. These are then followed by progressive neurological symptoms classically presenting with cerebellar ataxia, vertical supranuclear gaze palsy, gelastic cataplexy, seizures and eventually dementia (118). While studies of *NPC2* are easily interpretable, *NPC1* is highly polymorphic where variants of unknown

significance (VUS) are common, with one third of published variants being classified as VUS on ClinVar (117)(ClinVar, 2020).

NPC is one of the few degenerative ataxias for which a treatment is currently available. Miglustat (*Zavesca*) has been approved in several countries for treating progressive neurological complications of NPC (33). The drug has been shown to slow progression of neurological manifestations in children without severe neurological symptoms when initiating therapy, as shown by nonsignificant improvement in horizontal saccadic eye movement velocity in a preliminary open-label randomized controlled trial (119). Miglustat was also associated with improved swallowing function and decreased aspiration risk in observational studies (120, 121).

Despite most cases being diagnosed at a young age, the adult-onset form can be more insidious and often manifests with neuropsychiatric disturbances. The diagnosis is based on clinical evaluation and history with biomarker screening. Several blood-based biomarkers can be used to assist with diagnosis including oxysterols, lysosphingolipids and bile acid metabolites. Consensus guidelines still recommend completing the workup of suspected cases with two additional diagnostic methods: filipin staining of unesterified fibroblasts or molecular testing (33, 122). The former requires a skin biopsy with a specialized laboratory, but can be inconclusive in 15% of cases without molecular testing (123). Molecular testing is more practical but can also be inconclusive in up to 15% of cases mainly due to VUS and the absence of allele segregation studies (4). The assessment of these variants will require additional data input from various laboratories to allow for more specific classification.

CARTaGENE (CaG) is a cohort of healthy individuals living in Quebec (124). The cohort contains a total of 43,000 individuals, including 55% of women, ranging between 40 and 69 years of age. The recruitment for this cohort started in 2010 and participants have been followed up to this date. Genetic data is available for some of these individuals in the form of RNA as well as exome sequencing. This data can thus be used to screen for potential variants in a healthy pool of the population.

The study aim was to identify potential pathogenic variants in *NPC1* and *NPC2* in healthy individuals from the CaG cohort and classify their risk of pathogenicity using the American College of Medical Genetics and Genomics (ACMG) guidelines revised version of 2015 (103) to help assist

with the future interpretation of variants by providing useful additional information derived from large databases.

2.1.3. Materials and methods

2.1.3.1. Initial data

This study was based on a random sample from the CaG cohort, which was representative of the regional distribution of the Quebec population. Data acquisition was made from 911 individuals for the RNA-sequencing (RNA-seq) and 198 individuals for exome-sequencing (exome-seq). 93 of these individuals were in both RNA-seq and exome-seq. A bio-informatic pipeline was therefore used to analyse a total of 1016 individuals. Baseline characteristics as well as screening medical questionnaires were obtained for each participant from the CaG database. Informed consent was previously obtained by CaG researchers for all study participants. The Sample and Data Access Committee (SDAC) of CaG approved the use of the genetic and baseline characteristics for our study. All genetic and bioinformatic analysis were carried out in accordance with relevant guidelines and regulations. Our protocol was approved by our institution's research ethics board (CR-CHUM REB, Project 18.116).

2.1.3.2. Bio-informatic pipeline

The FASTQ files were aligned to the reference genome (Hg19) using BWA for exome sequencing and STAR for RNA-sequencing (76, 105). In both cases, variant calling was performed with GATK and annotated using ANNOVAR and custom scripts (107, 109). The bioinformatic pipeline in place allows the detection of single nucleotide variants (SNV) either non-synonymous, splice junction or synonymous, multi-nucleotide variants (MNV) and indels.

Since *NPC1* (NM_000271) and *NPC2* (NM_006432) genes are located on chromosome 18 or 14 respectively, we only extracted variants on those chromosomes for each sample. We also added information from the NP-C database (NPC-db2) made by the University of Tübingen, using a custom Python script. The database was last searched in July 2019. When a variant was found in the NPC-db2, its pathogenicity classification based on their criteria was added to the resulting file.

2.1.3.3. Analysis

As we were identifying rare variants, common variants (defined as > 1%) found in dbSNP, 1000 Genomes, Exome Variant Server, GnomAD and internal databases were filtered out. Non-synonymous, putative splicing variants and coding indels were prioritized. More specifically, we set a threshold for a CADD score higher than 15, a Polyphen2 score higher than 0.75 with a score of one being very likely pathogenic, a SIFT score, that was reversed to match the Polyphen2 score, with the same criteria as Polyphen2 to filter the variants (89, 91, 93, 100, 125). For the conservation scores, we identified the variants with a score higher than 500 for Phast cons and higher than 5 for GERP. Phast cons ranges between 0 and 1000 and the GERP score between − 12.3 and 6.17.

2.1.3.4. Classification

The ACMG 2015 revised guidelines were used to classify the different variants. The classification uses five distinct categories: benign, likely benign, uncertain significance, likely pathogenic and pathogenic. Each variant was analyzed using the ACMG criteria except the ones requiring segregation and laboratory data which were not available for our dataset. We extracted the NPC-db2 classification for each variant. Thereafter, the variants were searched on ClinVar for previous classification by other groups, using the ACMG criteria. Classifications based on other sets of criteria were not included in our tables. Finally, we searched the largest published study on NPC variants by Wassif et al. for identical variants already identified in their results. Previously unreported variants will be submitted to the ClinVar public database.

2.1.4. Results

2.1.4.1. Baseline Characteristics

The total sample size for both RNA-seq and exome-seq was 1016 patients and 2032 chromosomes. Clinical data was available for 1004 patients (Table 1). Females represented 51% of our population. The highest represented ethnicity was white from European descent (91.5%). The majority of these individuals were employed (64.3%). Age ranged from 44 to 69, with 42% of patients in the 40–49 range. This sample has a similar distribution as the general CaG cohort (Table 1). Additionally, the sample is also representative of the Quebec population based on the most recent epidemiological data (Le bilan démographique du Québec, 2019).

Tableau 1. Baseline characteristics

	Sample	CaG	Quebec
Female	50.6%	55%	50.0%
Age			
40-49	42.3%	42%	31.3%
50-59	29.0%	37%	32.6%
60-69	28.7%	25%	36.1%
Ethnicity			
White (European Descent)	91.5%		87.0%[b]
Arab	2.2%		2.7%[b]
Black (African or Caribbean descent)	1.7%		4.0%[b]
Employed	64.3%	67%	62.4%
Highest level of education			
University	37.1%	45%	31.2%[a]
College	49.8%	32%	45.2%[a]

[a]Data for age range of 35 to 64 in the Quebec population (Banque de données des statistiques officielles sur le Québec, 2015). [b]Data for overall province of Quebec (Statistics Canada 2016)

2.1.4.2. RNA-seq

Our study identified 32 unique rare variants from the 911 RNA samples that were run in the bio-informatic pipeline (Table 2, Fig. 3). Each variant was only present in one chromosome, for an allele frequency of 0.05% in our population. None of the study participants were heterozygous for two rare variants. Twenty of these variants were non-synonymous SNVs while the others were frameshift deletions. Among these variants, two were classified as pathogenic. Indeed, the p.Ile1061Thr is a known protein change that leads to a change from isoleucine to threonine. This variant has been described as causative of NPC in 15–20% of disease alleles in the United States and Europe. Additionally, biological studies have shown that this missense change affects proper protein localization and causes proteasomal degradation in cell culture. Another pathological variant, p.Pro543Leu, has been identified in 1 homozygous and 4 compound heterozygous individuals with symptomatic disease (126). It has previously been reported that this

Tableau 2. Rare variants in CARTaGENE sample, RNA-seq

Variants	Variations	Ref	Alt	Protein Change	Gene	ACMG Classification	Classification in NPC-db2	ClinVar (ACMG)	*Wassif et al.*
chr18:21121118	Nonsynonymous SNV	C	A	p.V810F	NPC1	3	2	3	-
chr18:21140411	Nonsynonymous SNV	T	C	p.N222S	NPC1	2	1	3	Benign
chr18:21136233	Nonsynonymous SNV	G	A	p.P434S	NPC1	2	1	-	Benign
chr18:21140367	Nonsynonymous SNV	G	A	p.P237S	NPC1	2	1	1	Benign
chr18:21113406	Nonsynonymous SNV	T	C	p.I1223V	NPC1	2	2	-	Benign
chr18:21119839	Nonsynonymous SNV	C	T	p.G911S	NPC1	2	1	-	Benign
chr18:21131617	Nonsynonymous SNV	G	A	p.P543L	NPC1	5	4	5	Probably damaging
chr18:21121386	Nonsynonymous SNV	C	T	p.V753M	NPC1	3	2	3	Benign
chr18:21114442	Nonsynonymous SNV	C	T	p.A1187T	NPC1	3	-	3	-
chr18:21140243	Nonsynonymous SNV	G	A	p.A278V	NPC1	3	-	-	-
chr18:21136410	Nonsynonymous SNV	T	C	p.T375A	NPC1	3	-	-	-
chr18:21134806	Nonsynonymous SNV	T	C	p.N490S	NPC1	3	-	-	-
chr18:21134743	Nonsynonymous SNV	G	A	p.T511M	NPC1	2	2	-	Probably damaging
chr18:21118536	Nonsynonymous SNV	G	A	p.S1004L	NPC1	3	2	-	Probably damaging
chr14:74959920	Nonsynonymous SNV	C	T	p.E20K	NPC2	3	-	-	-
chr18:21136422	Nonsynonymous SNV	C	A	p.V371F	NPC1	3	-	-	-
chr18:21140315	Nonsynonymous SNV	G	T	p.P254Q	NPC1	3	-	-	-
chr18:21136439	Nonsynonymous SNV	G	A	p.S365L	NPC1	3	-	-	Possibly damaging
chr18:21121045	Nonsynonymous SNV	A	G	p.M834T	NPC1	3	2	-	Benign
chr18:21116700	Nonsynonymous SNV	A	G	p.I1061T	NPC1	5	5	5	Benign
chr18:21153473	Frameshift deletion	AT	A	p.N41fs	NPC1	3	-	-	-
chr18:21125100	Frameshift deletion	CA	C	p.F590fs	NPC1	3	-	-	-
chr18:21141470	Frameshift deletion	TC	T	p.D162fs	NPC1	3	-	-	-
chr18:21121129	Frameshift deletion	TC	T	p.D806fs	NPC1	3	-	-	-
chr18:21140314	Frameshift deletion	TG	T	p.P254fs	NPC1	3	-	-	-
chr18:21119811	Frameshift deletion	AC	A	p.V920fs	NPC1	3	-	-	-
chr18:21120443	Frameshift deletion	AT	A	p.I858fs	NPC1	3	-	-	-
chr18:21153486	Frameshift deletion	TC	T	p.D37fs	NPC1	3	-	-	-
chr18:21124431	Frameshift deletion	AG	A	p.A669fs	NPC1	3	-	-	-
chr18:21116757	Frameshift deletion	TG	T	p.H1042fs	NPC1	3	-	-	-
chr18:21114436	Frameshift deletion	CTT	C	p.E1188fs	NPC1	3	-	-	-
chr18:21140211	Frameshift insertion	C	CA	p.A289fs	NPC1	3	-	-	-

[a]Classification 1: Benign 2: Likely benign 3:Uncertain significance 4: Likely pathogenic 5: Pathogenic.

mutation leads to early-infantile form of NPC (17). Both participants were heterozygous for these mutations and were asymptomatic according to the baseline medical screening obtained from CaG. The remaining twelve variants were indels and all were classified as VUS. Given the limited coverage, these could represent artifacts and their exact significance is difficult to interpret.

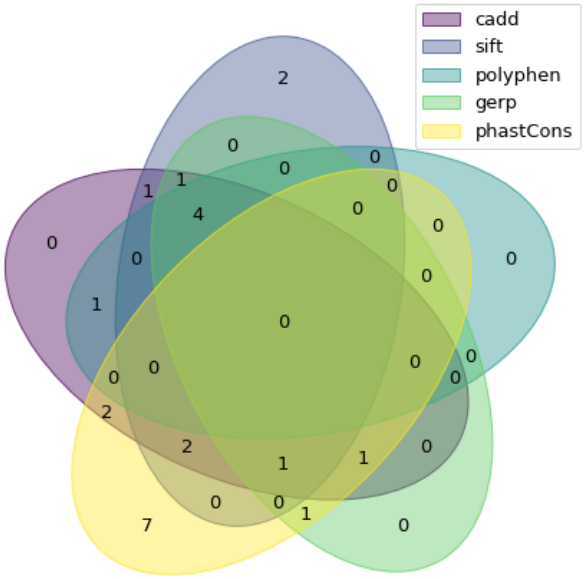

Figure 3. Venn diagram of included variants from RNA-seq for each bioinformatic filter

2.1.4.3. Comparison with other databases

The NPC-db2 database was searched for identified variants which were also present in our population. Twelve out of the 32 variants identified were also present in their sample. Our classification based on the ACMG criteria was overall very similar to their classification for shared variants. Notably, the p.Pro543Leu protein change was marked as potentially pathogenic in NPC-db2, while we were able to classify it as pathogenic based on previous publications and computational/predictive data. Additionally, we identified 13 variants that were filtered out by our bioinformatic pipeline, but that were present both in our population and in the NPC-db2 database (Table 3). These were all previously classified as benign. The main reason for their exclusion in the pipeline was a high allelic frequency (> 1%).

The variants in the study by Wassif et al. were also compared with variants identified in our study (Table 2) (114). Despite not specifically using the ACMG criteria , we were able to compare

Tableau 3. Variants in the CARTaGENE RNA-seq sample excluded from the pipeline but identified in NPC-db2

Variants	Variations	Ref	Alt	Protein Change	Gene	ACMG Classification	Classification in NPC-db2	ClinVar (ACMG)	*Wassif et al.*
chr18:21124945	nonsynonymous SNV	C	G	p.M642I	NPC1	1	1	1	-
chr18:21112206	nonsynonymous SNV	C	T	p.R1266Q	NPC1	1	1	1	Benign
chr18:21120444	nonsynonymous SNV	T	C	p.I858V	NPC1	1	1	1	Benign
chr18:21140432	nonsynonymous SNV	T	C	p.H215R	NPC1	1	1	1	Benign
chr18:21140367	nonsynonymous SNV	G	A	p.P237S	NPC1	1	1	1	Benign
chr18:21136233	nonsynonymous SNV	G	A	p.P434S	NPC1	1	1	-	Benign
chr18:21140411	nonsynonymous SNV	T	C	p.N222S	NPC1	1	1	2, 3	Benign
chr18:21148863	synonymous SNV	A	G	p.Y129Y	NPC1	1	1	-	Benign
chr18:21115579	synonymous SNV	G	A	p.L1111L	NPC1	2	1	-	-
chr18:21124335	synonymous SNV	G	A	p.N701N	NPC1	1	1	-	-
chr18:21134772	synonymous SNV	G	A	p.D501D	NPC1	2	1	-	-
chr18:21114440	synonymous SNV	C	A	p.A1187A	NPC1	2	1	-	-
chr18:21124365	synonymous SNV	C	T	p.P691P	NPC1	2	1	-	-

[a]Classification 1: Benign 2: Likely benign 3: Uncertain significance 4: Likely pathogenic 5: Pathogenic.

their five-level scale of classification to our data. Twelve out of the 32 variants were also classified in their study. One notable difference in classification in the variant p.Ile1061Thr was probably due to a mistake in their table as they present it as benign, while describing it as on of the most common pathogenic variant in their text (114). Moreover, five variants that were in both our databases were excluded by our pipeline. Once again, these were classified as benign and the main reason was a high allelic frequency (> 1%).

2.1.4.4. Exome-seq

Exome sequencing was performed for 198 individuals, composed of 93 individuals for whom we also had the RNA-seq and 105 new individuals for whom we only had exome-seq. Overall, 19 unique variants were identified in the samples, 4 of which were also present in the RNA-seq (Fig. 4). In participants for whom both RNA-seq and exome-seq data were available, the exact same sequence variants were found using both methods. After filtering by the bioinformatic pipeline, 10 variants were identified. Four of those were already found in our RNA-seq, including one classified as pathogenic (Table 4). The six other variants were found in individuals for which we only had exome data. These included two variants in splicing regions, one of which was causative of disease in previous publications (126, 127). The participant was heterozygous for this variant and was asymptomatic according to baseline medical screening obtained from CaG. The RNA-seq for these individuals were not available to confirm the presence of abnormal splicing.

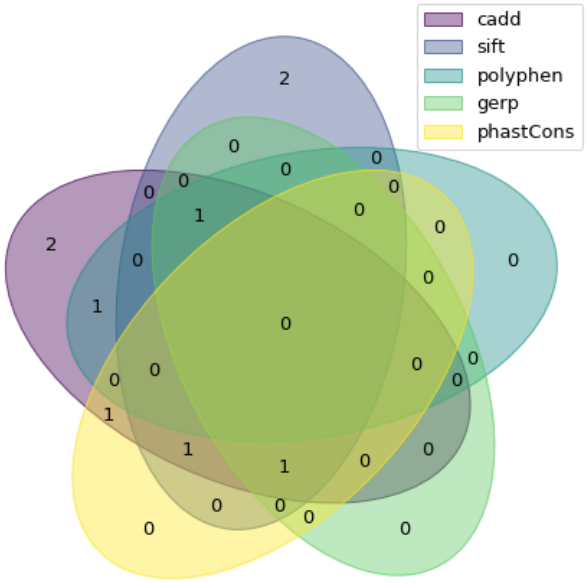

Tableau 4. Venn diagram of included variants from exome-seq for each bioinformatic filter

Tableau 5. Rare variants in CARTaGENE sample, exome-seq

Variants	Variations	Ref	Alt	Protein Change	Gene	ACMG Classification	Classification in NPC-db2	ClinVar (ACMG)	*Wassif et al.*
chr18:21118536	nonsynonymous SNV	G	A	p.S1004L	NPC1	3	2	4	Probably damaging
chr18:21134806	nonsynonymous SNV	T	C	p.N490S	NPC1	3	-	-	-
chr18:21121118	nonsynonymous SNV	C	A	p.V810F	NPC1	3	2	3	-
chr18:21118618	nonsynonymous SNV	C	T	p.V977I	NPC1	3	-	-	-
chr18:21116700	nonsynonymous SNV	A	G	p.I1061T	NPC1	5	5	5	Pathogenic
chr18:21115615	nonsynonymous SNV	T	C	p.I1099V	NPC1	3	-	-	-
chr14:74947404	splicing	C	T		NPC2	3	-	3	Strong negative
chr14:74953027	splicing-extended	C	T		NPC2	3	-	3, 5	-
chr14:74959920	nonsynonymous SNV	C	T	p.E20K	NPC2	3	-	-	-
chr14:74951269	nonsynonymous SNV	T	C	p.K71R	NPC2	3	-	-	Probably damaging

[a]Classification 1: Benign 2: Likely benign 3: Uncertain significance 4: Likely pathogenic 5: Pathogenic.

2.1.5. Discussion

Our study evaluated rare variants in *NPC1* and *NPC2* genes in a sample from the Quebec population in Canada. This population is unique given the important founder effect from French colonisation in the early seventeenth century (45). We had 911 RNA samples and 198 exome samples, with an overlap of 93 individuals. This study presents new variants that have not been previously described in the literature. In addition, known variants were reclassified based on the most recent literature. In fact, we classified each individual variant based on the ACMG 2015 guidelines and compared them to the NPC-db2 database and previously published studies on such variants. The majority of identified variants were of uncertain significance, likely benign or benign. However, we have also identified some likely pathogenic and pathogenic variants in heterozygous individuals.

To select rare variants in our population, we used a pre-specified bioinformatic pipeline. The variants were filtered based on their allelic frequency, with a coverage of at least 15% and where the alternative base was supported at least twice. This ensured that the classification would be applied to the most pertinent variants in our sample. We then focused our analysis on indels and non-synonymous variants, which were more likely to lead to pathogenic mutations. Additional variants were manually identified by comparing all variants present in our sample to those in the NPC-db2 database. These were filtered out from our pipeline according to the aforementioned criteria but were still classified by our laboratory because they were coincidentally present in another study population. The main reason for exclusion was allelic frequency > 1%.

With the increased use of genetic testing and the identification of more variants, it has become essential to apply rigorous classification in clinical genetic testing (128). The set of criteria must be evidence-based, standardized and objective (129). The ACMG 2015 guidelines, used in our study, have been largely used and therefore allow for easier comparison with previous publications. Individual laboratories also share their own classification in large databases (including ClinVar), but it is difficult to compare with their conclusions as the set of criteria is different. Thus, we have only compared our classification with published literature using the same set of criteria.

The CARTaGENE database encompasses a large sample of genomic data, but also baseline information based on detailed questionnaires. Answers included demographics, socioeconomic status, education and medical surveys. Given the possible adult-onset of NPC, we searched for potential symptoms in the questionnaires of patients with pathogenic or likely pathogenic variants. None of the identified individuals presented symptoms suggestive of the disease, as we had expected given the heterozygosity of the alleles. These pathogenic variants will allow us to estimate the carrier frequency in the Quebec population.

Our study has several potential limitations. First, our dataset is based on a relatively small sample size of 1016 individuals. However, given the important founder effect in our Quebec population pool, genetic variation is relatively lower when compared to other populations (130). Second, we did not perform any functional biology experiments which limits our ability to classify some of these mutations based on functional criteria. Third, no segregation data was available in the database, which can often provide strong evidence for a benign variant in a new mutation.

In brief, this study analyzed variants in the *NPC1* and *NPC2* genes from a representative sample of the Quebec population. The results described novel variants that were not previously described in the literature. In addition, known variants were reclassified using the ACMG guidelines. Despite identifying pathogenic or likely pathogenic variants, the individuals were heterozygous and asymptomatic based on baseline questionnaires. Classifying these variants is arduous given the scarcity of available literature, even so in a population of healthy individuals, leading to a large proportion of variants of uncertain significance. Using this data, we were able to identify three pathogenic variants within our population and several new rare variants in *NPC1* and *NPC2* which had not previously been identified. This additional information should help clinicians interpret the pathogenicity of variants identified in these two genes moving forward.

Data availability

Data was provided by the CARTaGENE database from a sample from the Quebec population. The data generated or analyzed during this study are included in this published article and its supplementary information files.

Code availability

We used publicly and freely accessible codes, referenced throughout our method section. Our custom Python scripts are available if necessary.

Acknowledgements

The authors would like to highlight the collaboration with the CARTaGENE database for the genetic and survey data used in this study. The authors would also like to express their appreciation to Dr Eric Bareke and Dr Alina Levtova for their great support in designing the methodology of this study

Author contributions

L.T. and M.L. wrote the main manuscript text and designed the tables. M.L., M.T. and A.D. designed the protocol. M.L. and M.T. ran the data through the bio-informatic pipeline and performed the analysis. L.T. completed the classification. All authors reviewed the manuscript.

Funding

We received an unrestricted educational grant from Actelion Pharmaceutiques Canada, RMGA. M.T. received a Junior 1 salary award from Fond de recherche du Québec—Santé. M.L. received an Excellence bursary from the Bioinformatic program, Université de Montréal.

Competing interests

The authors declare no competing interests.

2.2. Estimated prevalence of Niemann–Pick type C disease in Quebec

Scientific Reports 11, 22621 (2021). https://doi.org/10.1038/s41598-021-01966-0

Received: 11 August 2021; Accepted: 8 November 2021; Published online: 19 November 2021

Marjorie Labrecque[1,2,7], Lahoud Touma[2,3,7], Claude Bhérer[4], Antoine Duquette[2,3,5,6*] & Martine Tétreault[2,3*]

[1]Bioinformatics Program, Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC, Canada. [2]CHUM Research Center, Tour Viger, 900 rue Saint-Denis, R, Montréal, QC H2X 0A9, Canada. [3]Department of Neurosciences, Université de Montréal, Montréal, QC, Canada. [4]Department of Human Genetics, McGill University, Montréal, Canada. [5]Neurology Service, Department of Medecine, André-Barbeau Movement Disorders Unit, CHUM, Montréal, Canada. [6]Genetic Medicine Service, Department of Medecine, CHUM, 1000 rue Saint-Denis, Montréal, QC H2X 0C1, Canada. [7]These authors contributed equally: Marjorie Labrecque and Lahoud Touma. *email: antoine.duquette@umontreal.ca; martine.tetreault@umontreal.ca

Contribution

Martine Tétreault et Antoine Duquette ont participés à la conception de l'étude. Comme les données utilisés viennent de l'article d'identification et caractérisation, Marjorie Labrecque a effectué les analyses bio-informatiques et Lahoud Touma a fait la caractérisation. Marjorie Labrecque a fait l'analyse d'haplotype ainsi que l'enrichissement de variants et le calcul de la prévalence avec l'aide de Claude Bhérer. Martine Tétreault et Antoine Duquette ont aidé à l'interprétation des résultats. Marjorie Labrecque a écrit le manuscrit et fait les tableaux et figures. Tous les auteurs ont révisé le manuscrit.

2.2.1. Abstract

Niemann–Pick type C (NP-C) disease is an autosomal recessive disease caused by variants in the *NPC1* or *NPC2* genes. It has a large range of symptoms depending on age of onset, thus making it difficult to diagnose. In adults, symptoms appear mainly in the form of psychiatric

problems. The prevalence varies from 0.35 to 2.2 per 100,000 births depending on the country. The aim of this study is to calculate the estimated prevalence of NP-C in Quebec to determine if it is underdiagnosed in this population. The CARTaGENE database is a unique database that regroups individuals between 40 and 69 years old from metropolitan regions of Quebec. RNA-sequencing data was available for 911 individuals and exome sequencing for 198 individuals. We used a bioinformatic pipeline on those individuals to extract the variants in the *NPC1*/*2* genes. The prevalence in Quebec was estimated assuming Hardy–Weinberg Equilibrium. Two pathogenic variants were used. The variant p.Pro543Leu was found in three heterozygous individuals that share a common haplotype, which suggests a founder French-Canadian pathogenic variant. The variant p.Ile1061Thr was found in two heterozygous individuals. Both variants have previously been reported and are usually associated with infantile onset. The estimated prevalence calculated using those two variants is 0.61:100,000 births. This study represents the first estimate of NP-C in Quebec. The estimated prevalence for NP-C is likely underestimated due to misdiagnosis or missed cases. It is therefore important to diagnose all NP-C patients to initiate early treatment.

2.2.2. Introduction

Lysosomal storage disorders, like Gaucher's disease, Tay-Sachs disease or Niemann–Pick type C (NP-C) disease are a group of diseases characterized by cholesterol trafficking problems (131). The collective prevalence of lysosomal storage disorders is 1:5.000 births (132). NP-C (MIM 257220 and MIM 607625) are neurodegenerative autosomal recessive and pan-ethnic diseases with a prevalence varying between 0.35 and 2.2 per 100,000 births (110, 112). NP-C has different symptoms depending on the age of onset, which can be infantile, juvenile or in adolescence and adult-onset. The classical phenotype for NP-C is found in the infantile and juvenile populations (131). The clinical presentation is neuro-visceral with hepatosplenomegaly and neurological signs like delay in motor skills, clumsiness, hypotonia and ataxia (4). However, adult-onset subtypes have been reported and often present an atypical phenotype with psychiatric symptoms. In some cases, it can mimic other neurodegenerative diseases such as Alzheimer, Wilson or Parkinson (133-135). The clinical heterogeneity observed in NP-C makes the diagnosis challenging and it is often delayed, especially in atypical subtypes. As a result, the number of misdiagnoses could be high, and this may affect the estimated prevalence.

NP-C is caused by pathogenic variants in the *NPC1* gene (NM_000271) in 95% of cases and the rest is due to pathogenic variants in the *NPC2* gene (NM_006432) (4). In the gnomAD database, there are 2,227 variants in the *NPC1* gene and 467 in the *NPC2* gene (accessed on 7 April 2021) including intronic and UTR variants. An o/e (LOEUF) score above 0.35 for missense variants in both *NPC1* and *NPC2* suggests that these genes are somewhat tolerant to change and may be unlikely to have heterozygote pathogenic variants for the severe pediatric typical NP-C (136). However, these values need to be interpreted with caution in the context of atypical or adult-onset NP-C. The NPC-db2 database (see "Patients and methods" section) is also often used to assess the pathogenicity of variants. It regroups 692 variants in the *NPC1* gene and their classification. Amongst all those variants, 200 in *NPC1* and 5 in *NPC2* have been reported as pathogenic (137). Furthermore, there is growing evidence that heterozygote carriers of recessive mutations can increase the risk of developing a disease and even lead to milder and adult-onset phenotypes (138, 139) including for neurodegenerative disorders such as Parkinson and NP-C (140). Regarding NP-C specifically, one study described hepatosplenomegaly in 71% of heterozygotes carriers and several individuals had impaired cognitive functions (141). The same study also suggests that NP-C heterozygosity may lead to late-onset neurodegeneration. Thus, the late-onset form of NP-C, due to heterozygous mutations or simply the high heterogeneity of symptoms, can be misdiagnosed and result in NP-C being underdiagnosed.

The prevalence of the disease is variable worldwide due to population differences, diagnostic awareness and diagnostic methodology (4, 113). Some populations may also have higher disease prevalence because of a founder effect, leading some rare variants to become more frequent. French-Canadians from Quebec are a well-known example of such founder population, with more than 30 monogenic conditions showing an increased prevalence and/or particular variants/phenotypes (43, 46). Among the French-Canadian population, we also observe regional founder effects such as in the Saguenay and Gaspesia regions (142). A founder effect was previously reported in Canada for Niemann–Pick disease. Originally, Niemann–Pick type D (NP-D), had been reported clinically in Nova Scotia in four individuals of Acadian ancestry (143). The founder pathogenic variant, p.Gly992Trp in *NPC1*, has a carrier frequency between 10 and 26% (10). The high frequency of this variant is the result of a homogeneous genetic pool in the community and is unknown outside of Nova Scotia (143). Since the identification of the underlying genetic abnormality, NP-D is no longer recognized as a different disease and it is now included in

the NP-C spectrum (4). Considering the founder effect in the French-Canadian population, we want to explore how it may affect the prevalence of rare variants in *NPC1* and *NPC2* and, consequently, of the NP-C disease. In Quebec, CARTaGENE (CaG) is a unique prospective cohort of healthy individuals (56). It was developed to help study diseases and the genetics of the Quebec population.

Therefore, our goal is to measure the prevalence of NP-C in the Quebec population using population genetics data from CaG. Our results will inform on genetic testing approaches to be applied in a clinical setting in cases presenting clinical features reminiscent of NP-C and, more importantly, when an atypical form is encountered.

2.2.3. Patients and methods

CaG contains genomic data and health information for 20,000 residents of Quebec between the age of 40 and 69 (56). The data used in our analysis comes from the RNA-sequencing (RNA-seq) of 911 people and from the exome-sequencing (exome-seq) of 198 people in the CaG cohort (56). Among the 1,109 data sets, 93 people are in both the RNA-seq and exome-seq groups, which gives us sequencing information on 1,016 distinct individuals. The cohort used in this study is representative of Quebec's population based on sex, age and ethnicity (144). Individuals in our cohort had no known neurological diseases. The population structure was previously verified with a principal component analysis to confirm the French Canadians ancestry of the individuals of CaG (145).

The use of RNA-sequencing to identify variants has been proven to be reliable and enable the identification of high-quality variants (146-149). The pipeline for identification and classification of NP-C variants was used as described before (144). Briefly, the sequences from the 1016 individuals were aligned to the human reference genome version GRCh37. The RNA-seq was aligned with HISAT2 (106) and the variants were identified with VarDict (108). The exome-seq data was aligned with BWA (105) and the variants were identified using GATK (107). Both datasets were annotated with ANNOVAR (109) as well as custom scripts. Annotations include different pathogenicity scores (CADD (100), SIFT (91), Polyphen2 (93)), conservation scores (phastCons (89) and GERP (125)) and allele frequencies of different databases (GnomAD (136), NPC-db2 (https://medgen.medizin.uni-tuebingen.de/NPC-db2/ accessed on 18 July 2019)). We then kept the rare missense and indel variants with the following filters: allelic frequency (AF) $< 1\%$, a pathogenicity score CADD > 15 and inversed SIFT or Polyphen2 > 0.75 and for

conservation a GERP > 5 or phastCons > 500. The pathogenicity of variants identified in *NPC1* and *NPC2* was assessed using the ACMG guidelines (103). Using this approach, two pathogenic variants were selected for further analysis. No evidence of allele degradation through nonsense mediated decay was observed in our cohort.

For each variant, we estimated the allele frequency in our cohort as the count of alternate allele divided by 2,032, the total number of alleles. The allele frequencies observed in our cohort were compared to that observed in exome and genome sequencing data from 55,852 Non-Finnish Europeans from the gnomAD v2.1 database (136). To estimate which variants are found at significantly higher frequencies in the CaG cohort compared to Non-Finnish Europeans, we applied Fisher's Exact test one-sided alternative as implemented in R version 4.1.0 to obtain the p-value and the odds ratio.

Summing allele frequencies over all pathogenic variants we obtained the cumulative frequency of pathogenic variants in NPC1/2 genes. The prevalence was estimated using Hardy–Weinberg equation $p^2 + 2pq + q^2 = 1$, where q is the cumulative frequency of pathogenic variants. We can estimate q^2, the frequency of homozygotes, as the expected birth incidence assuming that it is at Hardy–Weinberg equilibrium and that the penetrance is complete. We were then able to calculate the number of births that could be affected by NP-C every year in Quebec by multiplying q^2 by the number of births in Quebec in 2019. By dividing the number of births in Quebec by the number of births that could be affected by NP-C we can measure the prevalence and report it per 100,000 births.

Ethics approval

The Sample and Data Access Committee (SDAC) of CaG approved the use of the genetic and baseline characteristics for our study. All genetic and bioinformatic analysis were carried out in accordance with relevant guidelines and regulations. Our protocol was approved by our institution's research ethics board (CR-CHUM REB, Project 18.116).

Consent to participate

Informed consent was obtained from all individual participants included in the study.

2.2.4. Results

2.2.4.1. Identification of two *NPC1/2* pathogenic variants in the CaG cohort

In the CaG cohort, we found two variants classified as pathogenic in the NPC1 gene (Table 5). The first one is in exon 10, c.1628C > T:p.Pro543Leu (P543L). This variant was found once before in a homozygote individual of French-Canadian origin (150). In our cohort, all three individuals with the P543L variant were found in the RNA-seq data and were heterozygote carriers of the variant. The first person identified with the P543L variant is a 46-year-old Caucasian woman. She has migraine occurrences but no known neurodegenerative disease. The second individual is a 51-year-old Caucasian male, also without known neurodegenerative disease. The third individual carrying the P543L variant, is a 62-year-old Caucasian woman also without neurodegenerative disease. All three individuals were sampled in the region of Quebec City, a region previously associated with founder mutations or diseases (43, 151). Thus, considering that all carriers of the P543L variant came from the same region in Quebec, we wanted to explore if they shared a haplotype. We found a haplotype of 9.4 Mb surrounding the variant that was shared among all carriers, suggesting a recent common ancestor, putatively of French-Canadian origin (Table 6).

Tableau 6. NP-C variants classified as pathogenic.

cDNA	Protein	Exon	Variant type	rsID	Allele number detected
c.1628C>T	p.P543L	10	missense	rs369368181	3
c.3182T>C	p.I1061T	21	missense	rs80358259	2

The second pathogenic variant found is in exon 21, c.3182T > C:p.Ile1061Thr (I1061T). This variant is the most common associated with NP-C (36). The first I1061T variant in CaG was found in the RNA-seq data, in a 62-year-old Caucasian female with no known neurological condition. The second I1061T variant comes from the exome-seq data and was found in a 54-year-old Caucasian, who did not mention any neurodegenerative problems. The two individuals with the I1061T variant were also heterozygous. Next, we wanted to see if both the variants identified as pathogenic in our cohort were more frequent in the Quebec population compared to the European population.

Tableau 7. Haplotypes of individuals with the P543L variant (in red).

Position	21109250	21111680	21120444	21131617	21140432	21148863	21166545	21713456	21714228	21714791	21714934	21715084	21715162	21715264	21742586	21743039	21891975	22020543	22032913	22033261	22056970	22057738
P1	T	G	C	A	C	A	G	T	T	A	T	C	T	CT	A	AT	G	C	G	G	A	C
	T	G	T	G	T	G	G	T	T	A	T	C	T	C	A	A	A	G	GT	G	A	C
P2	T	G	C	A	C	A	G	T	T	A	T	C	T	CT	A	AT	G	C	G	G	A	C
	T	G	C	G	C	A	G	T	T	A	T	C	T	C	A	A	A	G	GT	A	A	C
P3	C	G	C	A	C	A	G	T	T	A	T	C	T	CT	A	AT	G	C	G	G	A	C
	C	G	C	G	C	A	G	T	T	A	T	C	T	C	AT	A	A	G	G	G	A	T

2.2.4.2. A higher frequency of pathogenic variants in Quebec compared to Europeans

In total, we had a cohort of 1,016 individuals or 2032 alleles. The P543L variant found in three heterozygotes individuals has an allele frequency (AF) of 1.48e−03 in the CaG cohort and in Non-Finnish Europeans (NFE) from the gnomAD database, the AF is 1.79e−05. The I1061T is observed in two heterozygotes which means an AF of 9.84e−04 and 3.94e−04 in CaG and gnomAD NFE, respectively. Both variants show enrichment in allele frequency in our cohort compared to gnomAD NFE as shown in Fig. 5. The P543L variant is a lot more common in our population (one-sided Fisher exact test p-value = 5.577e−5; Odds Ratio (OR) = 82.3) and the I1061T variant is also enriched although not significantly (p-value = 0.1985; OR = 2.5).

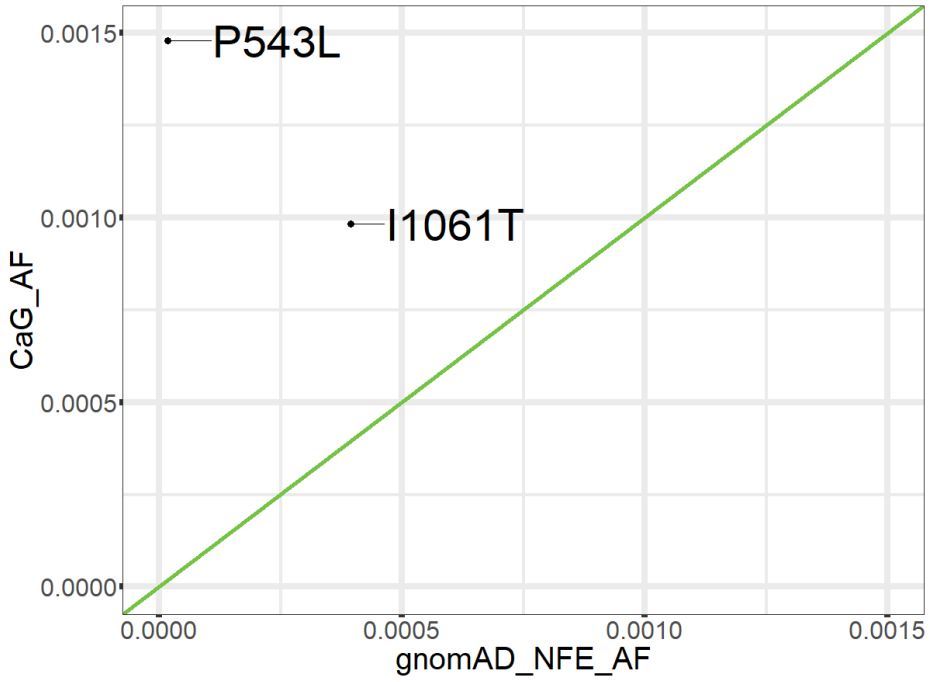

Figure 4. Comparison of variant allele frequencies (AF) in the CARTaGENE cohort compared to gnomAD NFE.

The AF of the two identified pathogenic variants, P543L and I1061T is plotted in the CaG cohort (CaG_AF) as a function of the allele frequency in gnomAD NFE (gnomAD_NFE_AF). Made with the ggplot package in R (version 4.1.0).

2.2.4.3. Estimated prevalence of NP-C in the Quebec population

Using the allele frequency from both pathogenic variants, P543L and I1061T, we estimated the prevalence of NP-C in the Quebec population. The frequency of allele mutant (q) was 0.0025 (5/2032) and the frequency of the disease (q^2) is 6.055e−6. Assuming Hardy–Weinberg equilibrium, there should be 0.51 new case of NP-C every year in Quebec based on the 84,200 births in 2019 (Le bilan démographique du Québec, 2020). We estimate a prevalence of 0.61 per 100,000 births in Quebec.

2.2.5. Discussion

In this study, we defined the prevalence of NP-C in the Quebec population. Analysis of RNA- and exome-sequencing from the CaG cohort led to the identification of two pathogenic variants P543L and I1061T. The allele frequencies of these variants were used to calculate the prevalence as well as establish if the variants were enriched in our population.

2.2.5.1. Pathogenic variants

The P543L variant was previously seen in one infantile homozygote French-Canadian with severe neurological symptoms reminiscent of classical infantile NPC phenotype and passed away at age 6 (150). Although the study described genetic and clinical findings in 35 patients of different ethnicities, this variant was only found in the French-Canadian patient, suggesting that the P543L variant could be more prevalent in Quebec. In our cohort, P543L was found in three heterozygotes individuals all originating from the same region in Quebec. Those individuals shared a large haplotype that suggests a common French-Canadian ancestor. Since the genotype of the parents of these individuals is not available, phasing of the haplotype is not possible, and thus a direct link cannot be established. However, the French-Canadian population, and specifically the Quebec-Charlevoix region, is well known for founder effects and several diseases or variants have been described more frequently in this region (43, 151). In addition, the P543L variant was significantly enriched in our cohort when compared to Non-Finnish Europeans. Thus, the suggestive haplotype, the enrichment in our population, and the report of this variant in a French-Canadian patient support P543L as a founder mutation. The association of this variant with a regional founder effect is of great importance as it will inform future genetic testing, where this variant should be prioritized in patients presenting a NP-C phenotype in Quebec.

The second pathogenic variant identified, I1061T, is the most prevalent variant associated with NP-C to date. Its prevalence varies depending on the population and the highest, 20%, is found in Western Europe (mainly in France and the UK), followed by Spain (10%), Portugal (6%) and Italy and Germany (5%) (36, 112, 152). This variant was also very prevalent in the Hispanic population of the Upper Rio Grande in the United States. This high prevalence is due to a founder effect from the Spanish settlers in Mexico at the beginning of the eighteenth century (36). Since Quebec was founded by French immigrants in 1600–1700 and English settlers later on, it would explain why this variant is also prevalent in our cohort. On the other hand, the prevalence of I1061T does not significantly differ from that of Non-Finnish Europeans.

2.2.5.2. Prevalence of NP-C

In the world, we observe an heterogenous estimation of NP-C prevalence, depending on the population. In Canada, and more specifically in Quebec, no prevalence had been calculated or estimated for NP-C. Elsewhere in the world, the prevalence of NP-C for 100,000 live births is 0.47 in Australia (111), 0.91 in Czech Republic (113), 2.2 in Northern Portugal (112), 0.35 in the Netherlands (110), 0.82 for France (4), 0.78 for the United Kingdom (16) and 1.12 for the United States of America (USA) (114). By combining these different prevalences, we obtain an average of 0.95 cases per 100,000 births. In all studies except for the USA, prevalence was measured using the number of patients diagnosed with NP-C divided by the number of births in the same period, reported on 100,000 births. For the USA, the prevalence is derived from an estimate based on four databases. The low prevalence observed in Australia and the Netherlands (0.47 and 0.35 per 100,000 respectively) could be explained by an underdiagnosis in the 1990s, when the phenotype spectrum was not clearly defined. More recent prevalence estimates were not found for those countries. Several factors, such as the inclusion of prenatal cases or of heterozygotes can influence prevalence. For example, in France, adding prenatal cases increases prevalence from 0.82 per 100,000 births to 0.96 per 100,000 births (4). In the UK, many heterozygote individuals for the I1061T variant had neurological symptoms in all age groups (16). NP-C in the heterozygous form may predispose patients to a late-onset form of the disease with symptoms of dementia, tremors similar to Parkinson's disease or psychosis (140). Individuals who are heterozygous for pathogenic variants and who present with atypical symptoms are likely to be misdiagnosed and this could underestimate the true prevalence of NP-C.

As there is no registry of NP-C cases in Quebec and it is impossible to measure the prevalence using the classical method, we used the Hardy–Weinberg equation. We also calculated the prevalence if one NP-C case was born in Quebec every year, every two years, every three years, or every four years. Using this method, the prevalence at birth would be, respectively, 1.19:100,000, 0.59:100,000, 0.39:100,000 or 0.30:100,000. Our prevalence estimate based on the two pathogenic variants found in the CaG cohort is 0.51 case every year or 0.61 case per 100,000 births. Based on clinical data, less than one case appears to be diagnosed in Quebec each year. If one NP-C case was diagnosed every two years, the disease would probably not be underdiagnosed in the province. Otherwise, it would suggest that NP-C could be underdiagnosed clinically in the Quebec population. Since our estimation is only based on heterozygote asymptomatic carriers of two pathogenic variants, it could explain the low estimated prevalence in Quebec. The death of early infantile cases may also play a role in an underestimation. Considering the founder effect, it would be interesting to measure the prevalence of NP-C using data only from individuals from French-Canadian origin in the CaG cohort. It could help us reinforce the role of the P543L variant as a founder mutation in French-Canadians. In the future, establishing a registry of NP-C cases in Quebec is of great importance since it would allow a more precise estimation of the prevalence. It is also important to ensure that all NP-C cases are properly diagnosed in order to start treatment as soon as possible. It is particularly true for adult-onset cases who are more difficult to diagnose but for whom therapy has proven helpful to reduce neurological symptoms (14). Identifying atypical cases is also a priority to ensure a better understanding of the causes and mechanisms underlying this rare disease.

Genetic testing is readily available for NP-C, for both NPC1 and NPC2 genes. In clinical practice, it is generally used as a confirmatory test after initial screening by measuring levels of oxysterols (153). The level of oxysterols is only done when clinical features suggest the diagnosis of NP-C. Some clinicians use a Suspicion Index tool based on visceral, neurologic and psychiatric symptoms to evaluate the suspicion of NP-C (154). In Quebec, newborn screening includes cystic fibrosis, congenital hypothyroidism, hemoglobinopathies and several metabolic disorders (155). The disease prevalence of NP-C remains relatively low but is more prevalent than some disorders that are systematically screened. Given the founder effect in Quebec, the difficult diagnosis and the potential treatment strategies that can improve outcomes, it would be interesting to consider adding oxysterols to blood screening in newborns. One important pitfall to screening with

oxysterols is the risk of false positive in patients with neonatal cholestasis (156). Thus, the optimal screening method for NP-C in this population requires further investigation. There is currently a pilot study of newborn screening assays in New York, including bile acids for NP-C, which will inform us on the clinical validity of screening newborns for complex disorders (157).

2.2.6. Conclusion

Certain diseases can be misdiagnosed for a variety of reasons including lack of awareness and heterogeneous clinical presentation. Being able to measure disease prevalence is very important to increase awareness among clinicians. Unfortunately, this can be difficult for rare diseases, where one misdiagnosis can have an important effect on prevalence. For hereditary diseases, estimating prevalence genetically can help compensate for these difficulties and NP-C is an excellent example of a condition for which symptom heterogeneity can make the diagnosis challenging. By identifying a founder variant in the Quebec City region and obtaining data suggesting that the disease is probably underdiagnosed, we are able to encourage clinicians of that area to consider the NP-C diagnosis more readily. This will hopefully lead to earlier detection and treatment of the disease.

Acknowledgements

The authors would like to highlight the collaboration with the CARTaGENE database for the genetic and survey data used in this study and Compute Canada for computational resources. The authors would also like to express their appreciation to Dr Eric Bareke and Dr Alina Levtova for their great support in designing the methodology of this study.

Author contributions

M.L. performed the bioinformatic analysis and prevalence calculation and wrote the first draft of the manuscript. L.T. performed the characterization and classification of the variants and revised the manuscript. C.B. participated in the prevalence calculation and revised the manuscript. A.D. participated in study design, interpretation of data and revised the manuscript. M.T. participated in study design, interpretation of data, supervised M.L. in bioinformatic analysis and revised the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by an unrestricted educational grant from Actelion Pharmaceutics Canada, RMGA. M.T. received a Junior 1 salary award from the Fond de recherche du Québec—Santé. M.L. received an Excellence bursary from the Bioinformatic program, Université de Montréal.

Competing interests

The authors declare no competing interests.

Chapitre 3 – Discussion

NP-C est une maladie neurodégénérative rare et pan-ethnique très peu étudiée au Québec. Nous avons utilisé une approche bio-informatique dans le but d'identifier des variants rares dans les gènes *NPC1* et *NPC2* associés à NP-C. Nous avons classifié des nouveaux variants et reclassifier d'autres selon notre population à l'aide des critères de l'ACMG. Par la suite, comme nous avons trouvé deux variants pathogéniques, nous les avons utilisés pour estimer pour la première fois la prévalence de NP-C au sein de la population québécoise.

3.1. Retour sur le premier article

Tout d'abord, les données cliniques disponibles de la cohorte CaG provenant des 1109 individus uniques séquencés initialement montrent que notre échantillon représente généralement bien la population québécoise. De plus, tous les participants étaient sains dans CaG selon les données cliniques disponibles, plus particulièrement dans le contexte de notre étude, ils n'avaient pas de maladie neurodégénératives au moment du recrutement. À l'aide de l'exome-seq, nous avons identifié 15 variants uniques dans exome seulement et quatre qui se trouvaient également dans le RNA-seq. Tous les individus avec ces variants étaient uniquement porteurs. Après les filtres, il en restait dix, dont les quatre également dans le RNA-seq (Tableau 4). Pour les six seulement dans l'exome, il y en avait deux qui étaient dans une région d'épissage. Tous les variants ont été classifiés comme ayant une signification incertaine, sauf le variant p.I1061T qui a été classifié comme pathogénique avec la méthode ACMG. Dans le résultat de RNA-seq, 32 variants uniques ont été identifiés dont quatre mentionnés précédemment qui étaient également dans l'exome-seq (Tableau 2). Comme pour l'exome, les variants identifiés dans le RNA-seq étaient uniquement présent de façon hétérozygote. 20 variants non-synonymes, 11 délétions et une insertion. Tous les indels ont été classifiés comme ayant une signification incertaine. Dans les non-synonymes, il y a six potentiellement bénins, 12 signification incertaine et deux pathogéniques. Dans les deux pathogéniques on retrouve le variant p.I1061T qui était dans l'exome-seq, mais chez un individu différent et le second variant est le p.P543L présent chez trois participants. Un seul variant trouvé dans *NPC2* autant pour exome-seq et RNA-seq est p.E20K. Dans le but de vérifier nos classifications, nous avons comparé nos résultats avec les bases de données ClinVar et NPC-db2 (Consultées en juillet 2019). Cette dernière est faite spécifiquement pour des variants de *NPC1*. Nous avons aussi vérifié avec les variants trouvé dans l'article de Wassif *et al.* (114). Dans la base

de données NPC-db2, 12 des 32 variants RNA-seq ont été retrouvés et classifiés de façon similaire. Le variant p.P543L était classifié potentiellement pathogénique alors que nous avons pu le classifier de pathogénique dans notre population Canadienne-Française. Nous avons exclu 13 autres variants qui se trouvaient dans NPC-db2, mais qui avait une AF supérieure à 1%. Pour l'article de Wassif *et al.*, 12 variants ont également été identifiés dans notre étude (114). De plus, il y en avait cinq présents dans leur étude que nous avons exclue dû à une trop haute AF chez nos participants.

Voyons un peu plus en profondeur les variants de *NPC1* qui ont été identifiés dans notre cohorte dans le RNA-seq. Le variant p.N222S a été classifié comme possiblement bénin, car il est prédit comme non-pathogénique par des outils de prédiction, il a tout même été identifié dans deux cas adultes comme hétérozygote composé (« compound heterozygote », en anglais) avec p.I1061T ou p.C468G (114, 158). Ainsi, seul, il n'est possiblement pas délétère pour l'individu. Le variant p.P237S a également été marqué comme potentiellement bénin avec l'ACMG. Une étude a démontré que ce variant n'affectait pas le mécanisme de transport de cholestérol non-estérifié qui cause normalement la maladie (159). Cela est confirmé par NPC-db2 et ClinVar qui l'ont identifié comme bénin. Un hétérozygote composé a précédemment été identifié (160) avec les mutations p.P237S et p.S1004L, qui ont également été identifiés dans notre cohorte dans RNA-seq et exome-seq. Cela montre qu'il est possible que ce soit le variant p.S1004L qui a engendré le phénotype non-classique de NP-C chez ce patient. De plus, Wassif *et al.* a également classifié ce variant comme probablement délétère (114). Le variant p.T511M est également classifié comme potentiellement bénin. Les informations dans la littérature sont contradictoires. Une étude a levé l'hypothèse que ce variant serait plutôt commun et pourrait mener à une forme adulte de NP-C (161). Alors qu'une autre étude suggère que le manque d'informations pourrait signifier que p.T511M mène à une forme mortelle prénatale (114). D'autres analyses pourraient être nécessaire pour mieux comprendre l'effet de cette mutation. Les autres variants non-synonymes n'ont pas d'information pertinente sur des patients dans la littérature et sont donc classifiés comme possiblement bénin ou de signification incertaine.

Le premier variant identifié dans *NPC2* dans le RNA-seq et l'exome-seq, est p.E20K, un nouveau variant, pour cette raison il est classifié comme signification incertaine. Par contre, une mutation non-sens au même endroit (p.E20X) est la plus commune de ce gène. Elle a d'ailleurs

bien été caractérisée chez quatre patients homozygotes avec une forme infantile qui présentent tous de l'hépatosplénomégalie (126). Il serait intéressant de mieux comprendre l'effet de p.E20K, surtout en prenant en compte que le changement protéique se fait d'un acide aminé polaire négativement chargé à un acide aminé polaire positivement chargé. Cela pourrait avoir un effet sur la fonction de la protéine (162). Dans l'exome-seq, deux variants d'épissage ont été identifié, c.441+1G>A dans 2 individus hétérozygotes et c.190+5G>A chez un hétérozygote. Le variant d'épissage c.441+1G>A a été identifié dans un patient hétérozygote d'une étude qui portait sur la démence (163). Le patient de 70 ans a un syndrome corticobasal qui entraîne la dégénération de cellules nerveuses. Il a aussi une histoire familiale de symptômes psychiatriques et a développé des troubles moteurs et de la dystonie à 67 ans. L'effet de ce variant d'épissage pourrait mener à une protéine incomplète. Le lien avec NP-C n'est pas direct, mais le patient a également une accumulation de cholestérol non-estérifié. Quant à lui, le variant d'épissage, c.190+5G>A ou IVS2+5G>A, a été vu dans plusieurs patients, dont deux frères et sœurs juvéniles homozygotes avec des symptômes légers et une chance de survie élevée (126). Un patient grec (164) et un patient chinois (165) ont également été trouvés avec ce variant. L'épissage de c.190+5G>A produit différent transcrits qui dans certains cas peut produire une protéine fonctionnelle ce qui pourrait expliquer que l'évolution de la maladie est plus subtile. Malgré le fait que les variants d'épissage c.441+1G>A et c.190+5G>A ont été trouvé dans des patients NP-C, il n'a pas été possible pour nous de classifier ces variants autrement que signification incertaine. Les VUS constituent un défi majeur lors de l'interprétation de variants liés à une maladie. Il est tout de même important de les reporter lors de publications. Il faut par la suite s'assurer de ne pas trop leurs donner d'importance, sans non plus les ignorer lors d'un diagnostic (166). L'utilisation de d'autres méthodes peuvent aider à la classification de VUS, comme par exemple le RNA-seq. Une étude sur des données RNA-seq a permis de reclassifier 34% de 250 VUS puisque ces variants avaient un effet sur l'épissage (167). Cette étude a également pu identifier des évènements d'épissage grâce au RNA-seq alors qu'ils n'étaient pas trouvés par le séquençage de Sanger, par exemple. Ainsi, les auteurs recommandent des analyses d'épissage à l'aide du RNA-seq pour augmenter le taux de diagnostic lors d'analyses bio-informatiques. Dans notre cas, les porteurs des variants d'épissage dans CaG n'avaient pas de données RNA-seq, alors des analyses d'épissage n'ont pas été possibles. Plus d'études sur ces variants et l'effet de l'épissage sur NP-C permettront de mieux comprendre les mécanismes qui mènent à différents phénotypes. Malgré le fait que la majorité des variants ne sont

pas pathogéniques, il est intéressant de voir que dans notre population saine, certains des variants identifiés chez des hétérozygotes, lorsque présent sous la forme homozygote ou composé hétérozygote, peuvent mener à une forme de NP-C. Nous avons réussi à identifier des nouveaux variants et en reclassifier certains qui pourront aider à améliorer la compréhension de la maladie. Autrement, les variants pathogéniques p.P543L et p.I1061T seront discutés plus en profondeur dans le deuxième article.

3.2. Retour sur le deuxième article

Dans le second article nous nous concentrons sur les variants pathogéniques identifiés précédemment pour estimer la prévalence de NP-C au Québec. Le premier variant pathogénique identifié est le p.P543L. Dans notre cohorte il est présent chez trois hétérozygotes. La première porteuse est une femme caucasienne de 46 ans, le deuxième est un caucasien de 51 ans et la dernière personne est une femme caucasienne de 62 ans. Ils n'ont tous pas reporté de troubles neurodégénératifs. Ces trois participants viennent également tous de la région de Québec. Cette région est connue pour avoir des mutations fondatrices et donc nous avons voulu vérifier si un haplotype était partagé. Une région de 9,4Mb autour du variant était partagée par les trois individus (Tableau 6) ce qui suggère un ancêtre commun possiblement d'origine Canadienne-Française. Le second variant est le plus commun parmi les européens, p.I1061T. Nous avons trouvé un porteur caucasien de 54 ans dans les données exome-seq et une femme de 62 ans également porteuse dans le RNA-seq. Les deux individus n'ont mentionné aucun symptômes neurodégénératifs. Pour vérifier si les variants étaient plus fréquents en Europe que dans la cohorte CaG, nous avons comparé l'AF des deux variants avec les AF de gnomAD chez les européens non finlandais (« Non-Finnish Europeans » en anglais; NFE). Pour p.P543L, l'AF est de 1,48e-03 (intervalle de confiance de 95% = [4.16e-04, 3.93e-03]) dans CaG et de 1,79e-05 [3.73e-06, 5.76e-05] chez les NFE. Pour le variant p.I1061T, l'AF dans CaG est de 9,84e-04 [2.05e-04, 3.15e-03] et celle dans NFE est de 3,94e-04 [2.90e-04, 5.24e-04]. Les deux variants sont enrichis dans la cohorte québécoise CaG (Figure 5). Des tests exacts de Fisher ont été faits pour démontrés le niveau d'enrichissement. Le variant p.P543L est enrichi significativement (valeur-P = 5,577e-05; rapport de côte (« odds-ratio » en anglais; OR) = 82,3). Cela est suggestif que p.P543L est une mutation fondatrice. De plus, la fréquence du variant p.P543L dans gnomAD NFE ne rentre pas dans l'intervalle de confiance de CaG, ce qui renforci que ce variant est très enrichi dans la population québécoise. Le variant

p.I1061T est également enrichi, mais de façon non significative (valeur-P = 0,199; OR = 2,5). La fréquence de NFE est également incluse dans l'intervalle de confiance de CaG pour p.I1061T. Cela n'est pas surprenant considérant que le variant p.I1061T est le plus commun en Europe. L'équation d'équilibre d'Hardy-Weinberg a été utilisée pour le calcul de prévalence de NP-C au Québec. La fréquence de l'allèle mutant (q, dans l'équation) a été calculée avec les AF des variants p.P543L et p.I1061T ce qui équivaut à 0,0025 ou 5 allèles mutantes divisées par les 2032 allèles totales de la cohorte. Il y aurait donc 0,51 nouveau cas de NP-C par année assumant qu'il y a eu 84 200 naissances au Québec en 2019. Ainsi, en assumant l'équilibre de Hardy Weinberg, notre estimation de la prévalence de NP-C au Québec est de 0,61 cas par 100 000 naissances.

Le variant p.P543L a été découvert en 2004 et a été identifié comme étant dans la boucle C de la protéine NPC1 (168). Cette boucle est d'ailleurs celle qui possède le moins d'importance au niveau structural et fonctionnel. Néanmoins, un patient homozygote pour p.P543L d'origine Canadienne-Françoise a été diagnostiqué avec NP-C forme infantile (150). Il est mort à l'âge de 6 ans, ce qui suggère que le variant mène à une forme sévère de la maladie avec symptômes neurologiques. Ce variant a par la suite été détecté dans plusieurs hétérozygote composé. Le premier, en combinaison avec le variant p.E612D, qui a été diagnostiqué avec une forme sévère, l'âge des premiers symptômes n'est pas mentionné (169). Le second patient a les variants p.P543L et p.V1212L (170). Ces premiers symptômes, troubles de déglutition (dysphagie) et hépatosplénomégalie, sont apparues à 4 ans et il a reçu le diagnostic à 9 ans. La dernière étude a identifié deux patients hétérozygote composé, un accompagné d'un variant d'épissage et l'autre avec un variant qui décale le cadre de lecture (171). Les deux ont de l'hépatosplénomégalie, des troubles pulmonaires et cognitifs. Le premier a également du VSPG et le deuxième a de la dysphagie. Ils ont également une forme infantile précoce. Le variant p.P543L est donc sévère chez les homozygotes ou accompagné d'un autre type de mutation. De plus, comme ce variant a également été trouvé dans une autre personne d'origine Canadienne-Française, il est possible qu'il s'agisse d'une mutation fondatrice. L'haplotype de 9,4Mb autour du variant partagé par les trois individus provenant de la région de Québec, suggère un possible effet fondateur et un ancêtre commun. En effet, un gros haplotype partagé est signe de sélection positive (172). Le variant p.P543L est également enrichi de façon significative dans la population comparé aux NFE et avec un OR de 82,3, l'association est forte. Tous ces éléments combinés et le fait que la région de Québec

est bien connue pour les mutations fondatrices (43, 138) sont indicatif que le variant p.P543L est une mutation fondatrice.

Le second variant pathogénique identifié est p.I1061T. Il s'agit du variant associé à NP-C le plus commun et il est donc très bien caractérisé dans la littérature. Ce variant mène à une forme juvénile classique, il est le plus fréquent en France, en Angleterre et dans une population fondatrice aux États-Unis (36). Un modèle de souris a également été fait pour étudier les effets du variant et développer des traitements (173). Le variant n'a pas été vu dans des contrôles ou des enfants atteints de la maladie (36). Il se pourrait donc que chez un hétérozygote, une forme plus douce ou tardive de la maladie se développe. Comme cela pourrait être le cas des personnes identifiées dans notre cohorte. Dans les maladies à transmission récessive, le concept que des hétérozygotes peuvent développer certains symptômes ou avoir un risque plus accru de développer la maladie plus tard a gagné un intérêt croissant dans la littérature (140). C'est d'ailleurs le cas dans la maladie de Parkinson, pour des variants hétérozygotes dans le gène *Parkin* (174). De plus, il existe des patients qui possèdent des variants NP-C qui se font diagnostiquer avec une forme de parkinsonisme. Un patient a été diagnostiqué avec la maladie de Parkinson à 65 ans, à 71 ans il a développé un peu de démence et son petit-fils a été diagnostiqué avec NP-C comme hétérozygote composé avec p.I1061T et p.R1186G à l'âge d'un an (175). Ils ont ensuite découvert que le patient était porteur de p.I1061T. Il s'agit donc d'un exemple où un patient hétérozygote a développé une maladie neurodégénérative tardivement. D'autres études ont montré que des hétérozygotes pouvaient également avoir un test à la filipine positif, qui montre une accumulation plus légère de cholestérol non-estérifié ou des symptômes légers de maladies neurodégénératives (163, 176, 177). Cela mène à la conclusion que les cas hétérozygotes qui ont une présentation atypique ou tardive sont possiblement sous-diagnostiqués, influençant la vraie valeur de la prévalence. Considérant ces informations, il serait intéressant de voir si les cinq participants hétérozygotes pour des variants pathogéniques *NPC1* dans CaG ont des symptômes maintenant, surtout que cela fait plus de 10 ans qu'ils ont été recrutés, mais qu'il n'y a pas eu de suivi (56). S'ils ont des symptômes semblables à ceux de NP-C, cela pourrait solidifier l'argument que les patients hétérozygotes peuvent développer la maladie tardivement.

À l'aide des deux variants décrits ci-haut, la prévalence de NP-C dans la population québécoise a été estimée à 0,61 cas par 100 000 naissances. Ce résultat se trouve sous la moyenne

calculée selon la littérature. Ce calcul se base sur les prévalences trouvés dans la littérature pour les Pays-Bas, l'Australie, le Royaume-Uni, la France, la République Tchèque, les États-Unis et le nord du Portugal. Leurs prévalences respectives sont de 0,35, 0,47, 0,78, 0,82, 0,91, 1,12 et 2,20 cas par 100 000 naissances (4, 16, 110-114). Les deux plus basses prévalences proviennent d'études faites dans les années 1990, un moment où la maladie était moins comprise et des cas ont probablement été exclus. Par exemple, en Australie, le plus vieux patient diagnostiqué avait seulement 37 ans (111) Alors que nous savons maintenant que des patients avec apparition de symptômes plus tardifs, pouvant aller jusqu'à 60 ans, existent (4). De plus, l'Australie a été colonisé par l'Angleterre et donc il serait normal que la prévalence ressemble à celle de ces colonisateurs ou encore celle des États-Unis qui ont été colonisé par l'Angleterre également (111). La très haute prévalence obtenue au nord du Portugal peut être expliqué par le fait que cette région est isolée géographiquement ce qui a mené à un effet fondateur (112). Considérant cet élément, il est surprenant que notre estimation de prévalence soit aussi basse, puisqu'un effet fondateur est également présent dans notre population. Par contre, notre prévalence a été mesurée à partir de participants sains, il s'agit donc de l'estimation la plus minimale pour le Québec. La prévalence de 0,61 cas par 100 000 naissances signifie qu'il devrait y avoir 0,51 cas de diagnostiqué par année. Ceci combiné au fait que moins d'un patient est réellement diagnostiqué avec NP-C par année, il est très probablement que la maladie soit sous-diagnostiqué dans la population québécoise. Les cas avec symptômes légers, soit parce qu'ils sont atypiques ou qu'ils sont hétérozygotes, accentuent davantage le taux d'erreur de diagnostic, menant à une sous-estimation de la prévalence. En effet, dans la cohorte du Royaume-Uni, 42 patients avec mutations hétérozygotes de p.I106T ont éventuellement développer des symptômes neurologiques (16). Leur inclusion dans le calcul de prévalence ferait beaucoup augmenter cette valeur, surtout considérant que 146 patients ont été utilisé initialement. L'inclusion de cas prénataux, provenant de grossesses terminées a également un effet sur la prévalence comme l'a démontré une étude en France (4). La prévalence a passé de 0,82 cas à 0,96 cas par 100 000 naissances en incluant les cas prénataux. Ainsi, pour obtenir une meilleure estimation, il sera nécessaire de mieux comprendre les mécanismes qui expliquent la variabilité phénotypique et améliorer la détection de cas de NP-C dans la population générale et les cas prénataux. La sensibilisation à NP-C a d'ailleurs permis d'identifier de plus en plus de cas au Royaume-Uni. Ils avaient en moyenne 3,5 nouveaux cas par année avant les années 2000 et entre 2011 et 2015, ils ont environ 6 nouveaux cas par années (16).

3.3. Limitations

3.3.1. Taille de l'échantillon et choix de cohorte

L'étude présentée dans ce mémoire ne représente pas une étude de type cas-contrôle. La raison principale est que la maladie étudiée est une maladie rare, avec variabilité phénotypique et diagnostic difficile. Pour cette raison, nous avons décidé d'utiliser la cohorte CaG pour en apprendre plus sur le profil génétique de la maladie dans la population québécoise. Ainsi, nous avions une cohorte de 1109 individus sains entre 40 et 69 ans, dont les données de séquençage étaient disponibles. Cela pourrait être qualifiée de relativement petit, considérant que le Québec a une population de plus de 8,6 millions d'habitants aujourd'hui (Institut de la statistique du Québec, consultation le 15 juin 2022). Malgré le fait que l'étude regroupe des individus sains, ceux-ci peuvent tout de même nous informer sur la composition génétique de la population. Par exemple, ils peuvent être porteurs de variants pathogéniques pour des maladies dont seulement les homozygotes ont des symptômes ou encore avoir un variant pour une maladie dont l'âge de début est plus tardif (178).

La cohorte CaG a été conçue pour représenter les grandes régions métropolitaines qui correspondent à 55,7% du Québec (56). Alors, les individus des régions plus rurales ne sont pas représentés dans cette étude. La moins bonne représentation de la cohorte a peut-être un effet sur la précision du calcul de prévalence. Il a tout de même été possible de faire une première estimation de la prévalence qui facilitera le recrutement pour de futurs études. Par contre, l'effet fondateur présent au Québec a pour résultat que le pool génétique soit moins diversifié. De plus, les individus de la cohorte de CaG ont été choisis de façon aléatoire et ils n'ont pas été pris parce qu'ils avaient un phénotype précis. Cela à l'avantage de représenter plus globalement le Québec. Une autre contrainte est que les informations familiales des participants ne sont pas disponibles ce qui fait qu'il n'est pas possible de phaser l'haplotype pour avoir une idée plus claire de l'effet fondateur. Finalement, la base de données de CaG représente un avantage particulier pour étudier la génétique de la population québécoise.

3.3.2. Critères de l'ACMG

Pour la classification avec ACMG, plusieurs limitations sont présentes dans notre étude. D'une part, un des critères se base sur l'AF pour confirmer qu'un variant est bénin, celle-ci devant

être supérieur à 5% pour être classifié comme bénin (103). Par contre, ce n'est généralement pas ajusté pour les variants rares associés à des maladies rares. En effet, dans les études sur les maladies rares, l'AF utilisée est généralement de 1% ou même inférieur, ce qui fait que certains variants pourraient être faussement identifiés comme bénin dans notre étude (179).

Un autre critère de l'ACMG est l'utilisation de données d'analyses fonctionnelles. Les tests fonctionnels permettent d'identifier l'effet du variant sur la conformation et la fonction de la protéine et ainsi l'effet sur le phénotype (180). Ces données permettent d'aider à la classification d'un variant de signification incertaine vers un variant soit bénin ou pathogénique. Une méthode pour confirmer si un variant est délétère est d'utiliser un western blot ou un test de réaction de polymérisation en chaîne. Par contre, en utilisant la base données CaG, nous n'avons pas accès aux échantillons biologiques des participants, alors de tests biochimiques sont impossibles dans notre cas. Considérant que NP-C est une maladie rare dont peu d'études fonctionnelles ont été faites et qu'il y a un biais de publication pour les études qui identifient des variants pathogéniques (179), cela rend difficile l'analyse fonctionnelle à l'aide de la littérature. Une étude de Scott *et al.* (168) résume l'emplacement de 110 mutations sur la protéine NPC1, qui se retrouvent sur les boucles et les régions transmembranaire de la protéine. Cela permet de voir quelles régions sont plus tolérantes aux mutations, mais ne permet pas de savoir si les régions qui possèdent plus de mutations qu'estimé ont plus d'impacts sur la fonction (168). Pour vérifier cela, des modèles de souris possédant la mutation que l'on veut tester est une bonne méthode pour vérifier la pathogénicité. Un modèle de la mutation la plus commune, p.I1061T, a été faite et ressemble énormément au phénotype chez l'humain (173).

Le dernier critère ACMG qui permet de faire une classification de variant est l'utilisation de données de ségrégation. Dans notre cas, avec la cohorte CaG, ces informations étaient tout simplement indisponibles. La ségrégation utilise les données génétique de la famille pour confirmer si un variant est bénin ou pathogénique. En effet, un manque de ségrégation d'un variant peut indiquer que le variant n'est pas pathogénique, car le lien entre le variant et le phénotype observé dans la famille n'est pas confirmé (103, 179). Dans une étude brésilienne sur la ségrégation dans des maladies rares, sur 321 variants, ils ont réussi à reclassifier 51 variants comme bénin et 211 comme potentiellement bénin (179). Le même processus peut également être fait pour identifier des variants pathogéniques. Dans les deux cas, il est possible d'avoir une meilleure compréhension

de l'effet du variant et ainsi aider au diagnostic. Deux conclusions importantes sont ressorties par l'étude de Quaio *et al.* (179). La première est que les analyses de ségrégation sur les maladies avec début tardif sont plus complexes. La seconde est que les SNVs rares sont spécifiques au populations et donc que les résultats bénins du critère de ségrégation ne s'appliquent pas nécessairement aux autres ethnicités. Dans CaG, les individus proviennent tous de familles différentes et donc le critère de ségrégation n'a simplement pas pu être utilisé dans notre étude (56). Cela a comme conséquence de réduire la précision de la classification.

3.3.3. Hardy-Weinberg

Pour le deuxième article, l'utilisation de l'équilibre de Hardy-Weinberg pour le calcul de la prévalence contient des conditions pour assurer l'équilibre. Nous avons en effet assumé que la population était en équilibre avec différentes conditions pour pouvoir utiliser la formule de calcul de prévalence. Ces conditions incluent que la population doit être de taille infinie, que les couples soient formés par hasard, que les générations doivent être distinctes, qu'il n'y a pas de sélection naturelle, de mutation ou de migration. Pour la population de taille infinie, il est évidemment impossible de remplir cette condition. il est possible de remplacer cette notion pour avoir un échantillon assez grand pour que lorsqu'il y a des remplacements ou non nous arrivions au même résultat (181). Pour la sélection naturelle et la migration, il est également très difficile de justifier qu'il n'y en a pas. En effet, avec l'immigration et l'émigration des populations à travers le monde, la migration est bien présente. Ces conditions servent normalement à assurer que les fréquences alléliques restent les mêmes chez un individu jusqu'à l'âge de reproduction. Il est également impossible que la condition de mutation soit respectée. Un article de 2012 a démontré qu'il y aurait environ 74 SNV *de novo*, trois indels et 0,02 CNV de nouveau par individu à chaque génération (182). L'effet de chacune de ces conditions sur la variation des fréquences génotypiques est donc difficile à quantifier lors d'études populationnelles, mais pourrait être négligeable selon le contexte. De nos jours, l'équation de Hardy-Weinberg est principalement utilisé pour tester quels conditions font qu'une population n'est pas en équilibre lors de tests d'associations (183). Il est donc très justifiable, dans notre cas, d'assumer que l'équilibre d'Hardy-Weinberg est atteinte dans la population québécoise pour obtenir une estimation de la prévalence de NP-C.

Chapitre 4 – Conclusion

Les articles présentés dans ce mémoire ont permis d'identifier et caractériser des variants rares chez les québécois dans les gènes *NPC1* et *NPC2* associés à la maladie de NP-C. Avec les variants pathogéniques identifiés, nous avons pu faire la toute première estimation de cette maladie au Québec. Nous avons réussi à identifier 37 variants parmi une cohorte de 1109 individus uniques de la cohorte CaG. Parmi ces variants, il y avait 12 indels, deux variants d'épissage et 23 SNVs non-synonymes. Deux de cette dernière catégorie ont été classifiés comme pathogéniques, p.P543L et p.I1061T. Avec ces résultats, nous avons obtenu une prévalence de 0,61 cas par 100 000 naissances au Québec. Selon les données cliniques disponibles, il est possible que NP-C soit sous-diagnostiquée dans la population québécoise.

Les maladies rares, particulièrement lorsqu'étudié au sein d'une population fondatrice, comme celle Canadienne-Française, demandent une attention particulière considérant leur pool génétique différent de la population initiale. Une population fondatrice qui est très connue est la Finlande (184). En effet, la Finlande a été colonisée par des européens et la taille de la population a beaucoup augmentée depuis 1750. La population aurait vécu un goulot d'étranglement très serré ayant comme résultat que les variants communs sont restés aussi communs que le reste de l'Europe et que la fréquence des variants rares a beaucoup augmentée menant à une fréquence élevée de maladies récessives (184, 185). Une autre population très connue pour avoir une fréquence de variants délétères supérieure au reste du monde sont les juifs Ashkénaze. Ils sont des descendants des juifs de l'Europe il y a plus de 1 000 ans (186). Comme ces populations fondatrices savent qu'ils ont des particularités, ils peuvent ajuster en conséquence les méthodes de diagnostic.

Ainsi les études dans les populations fondatrices et l'identification de mutation fondatrice apportent une vision importante pour la compréhension de différentes maladies. Cela mènera à un meilleur diagnostic selon la population et une meilleure gestion clinique par la suite. Comme c'est le cas du variant p.P543L identifié ici qui devrait être inclus à partir de maintenant lors du dépistage de maladies neurodégénératives au Québec. Cela permettra possiblement de réduire le délai nécessaire pour le diagnostic de NP-C qui peut atteindre six ans chez les adultes (15). C'est d'ailleurs le cas des populations juives où le diagnostic prénatal de la maladie lysosomal de Tay-Sachs a permis de réduire significativement le nombre de cas diagnostiqué chaque année, passant de 40 à quatre environ (186). Le dépistage est important autant chez les familles qui souhaitent

avoir un enfant que chez les femmes déjà enceintes. Une étude de 1992 présentait déjà une façon de détecter des cas de la forme classique de NP-C chez les femmes enceintes (187), il faudrait améliorer cette technique pour inclure les formes des cas atypiques qui ont été découvertes depuis. Le dépistage de NP-C est particulière important puisqu'un traitement, le miglustat, est disponible pour réduire les symptômes neurologiques (33). Un groupe de patients a pris ce traitement pendant 12 mois et une amélioration ou stabilisation a été vu chez 70,5% des gens (188). Cette étude est la plus grande cohorte de NP-C utilisée pour des recherches cliniques, regroupant 163 patients de l'Europe, le Brésil, l'Australie et le Canada. Les résultats du traitement sont encourageant puisqu'ils sont reproductibles à travers le monde, dans différentes populations.

Le ministre de la Santé et des Services sociaux du Gouvernement du Québec a fait l'annonce récemment d'une politique pour améliorer la sensibilisation et le diagnostic de maladies rares au Québec (189). Cela permettra de sensibiliser la population, offrir une meilleur formation au personnel de santé pour prendre en main les patients atteints de maladies rares et donner accès à un meilleur diagnostic pour la population. Il serait alors possible d'utiliser la méthode présentée ici pour identifier plus de mutations fondatrices pour augmenter la compréhension de maladies rares qui touchent la population québécoise. De plus, la construction d'un registre pour chacune des maladies rares identifiées au Québec serait idéale. Cela permettra de regrouper les patients et de collecter les informations pertinentes sur ceux-ci pour faciliter la recherche par la suite. La difficulté, comme nous avons pu le voir dans ce mémoire est de recruter des patients. Ainsi, rassembler les informations à un même endroit et à long terme permettra de mieux comprendre les maladies avec chaque nouveau patient qui est identifié. Le registre est également un bon moyen pour regrouper des patients sans diagnostic. Leurs résultats de test clinique pourront être accessibles et les VUS identifiés pourront être reclassifiés au fur et à mesure que des nouveaux résultats de recherche s'ajoutent. Une emphase particulière pourra être mise sur ces patients dans le but d'identifier la cause de leurs symptômes. Des techniques comme le RNA-seq, des modèles cellulaires ou des modèles d'animaux simples pourront alors être utilisés pour améliorer la compréhension. Comme nous avons vu plus haute, le RNA-seq permet de reclassifier des variants en vérifiant leur effet sur l'épissage (167). Des animaux modèles comme la drosophile et le *C. elegans* sont souvent utilisés comme modèles fonctionnelles pour aider à l'interprétation de la pathogénicité d'un variant (190). Cette étude a conclu que les modèles animaux sont plus précis que les outils de prédictions bio-informatiques comme SIFT et CADD. Ainsi, ils apportent

beaucoup de validation au niveau fonctionnel, mais ils sont plus difficiles à faire. Pour les cultures cellulaires, l'utilisation de cellule souches pluripotentes induites est un autre moyen d'étudier l'effet de VUS sur des patients (191). Avec cette technique, différentes maladies cardiovasculaires ont été étudiés et mieux comprises. Combiné avec la technologie de CRISPR, il est possible d'étudier différentes maladies et de simuler différents variants. Toutes ces techniques pourront éventuellement améliorer la compréhension de VUS et ainsi le diagnostic de maladies rares.

Références bibliographiques

1. Sun A. Lysosomal storage disease overview. Annals of translational medicine. 2018;6(24).

2. Vellodi A. Lysosomal storage disorders. British journal of haematology. 2005;128(4):413-31.

3. Mehta A. Epidemiology and natural history of Gaucher's disease. European Journal of Internal Medicine. 2006;17:S2-S5.

4. Vanier MT. Niemann-Pick disease type C. Orphanet journal of rare diseases. 2010;5(1):1-18.

5. Crocker AC, Farber S. Niemann-Pick disease: a review of eighteen patients. Medicine. 1958;37(1):1.

6. Crocker AC. The cerebral defect in Tay-Sachs disease and Niemann-Pick disease. Journal of neurochemistry. 1961;7(1):69-80.

7. Schuchman EH, Desnick RJ. Types a and B Niemann-pick disease. Molecular genetics and metabolism. 2017;120(1-2):27-33.

8. McGovern MM, Aron A, Brodie SE, Desnick RJ, Wasserstein MP. Natural history of Type A Niemann-Pick disease: possible endpoints for therapeutic trials. Neurology. 2006;66(2):228-32.

9. McGovern MM, Lippa N, Bagiella E, Schuchman EH, Desnick RJ, Wasserstein MP. Morbidity and mortality in type B Niemann–Pick disease. Genetics in Medicine. 2013;15(8):618-23.

10. Greer WL, Riddell DC, Gillan TL, Girouard GS, Sparrow SM, Byers DM, et al. The Nova Scotia (type D) form of Niemann-Pick disease is caused by a G3097-->T transversion in NPC1. Am J Hum Genet. 1998;63(1):52-4.

11. Spiegel R, Raas-Rothschild A, Reish O, Regev M, Meiner V, Bargal R, et al. The clinical spectrum of fetal Niemann–Pick type C. American Journal of Medical Genetics Part A. 2009;149(3):446-50.

12. Baxter LL, Watkins-Chow DE, Johnson NL, Farhat NY, Platt FM, Dale RK, et al. Correlation of age of onset and clinical severity in Niemann–Pick disease type C1 with lysosomal abnormalities and gene expression. Scientific reports. 2022;12(1):1-13.

13. Geberhiwot T, Moro A, Dardis A, Ramaswami U, Sirrs S, Marfa MP, et al. Consensus clinical management guidelines for Niemann-Pick disease type C. Orphanet journal of rare diseases. 2018;13(1):1-19.

14. Wraith JE, Vecchio D, Jacklin E, Abel L, Chadha-Boreham H, Luzy C, et al. Miglustat in adult and juvenile patients with Niemann-Pick disease type C: long-term data from a clinical trial. Mol Genet Metab. 2010;99(4):351-7.

15. Sévin M, Lesca G, Baumann N, Millat G, Lyon-Caen O, Vanier MT, et al. The adult form of Niemann–Pick disease type C. Brain. 2007;130(1):120-33.

16. Imrie J, Heptinstall L, Knight S, Strong K. Observational cohort study of the natural history of Niemann-Pick disease type C in the UK: a 5-year update from the UK clinical database. BMC Neurol. 2015;15:257.

17. Verot L, Chikh K, Freydiere E, Honore R, Vanier MT, Millat G. Niemann–Pick C disease: functional characterization of three NPC2 mutations and clinical and molecular update on patients with NPC2. Clinical genetics. 2007;71(4):320-30.

18. Pacheco CD, Lieberman AP. The pathogenesis of Niemann–Pick type C disease: a role for autophagy? Expert reviews in molecular medicine. 2008;10.

19. Mellman I. ENDOCYTOSIS AND MOLECULAR SORTING. Annual Review of Cell and Developmental Biology. 1996;12(1):575-625.

20. Zhang L, Sheng R, Qin Z. The lysosome and neurodegenerative diseases. Acta Biochimica et Biophysica Sinica. 2009;41(6):437-45.

21. Vance JE. Lipid imbalance in the neurological disorder, Niemann-Pick C disease. FEBS letters. 2006;580(23):5518-24.

22. Sleat DE, Wiseman JA, El-Banna M, Price SM, Verot L, Shen MM, et al. Genetic evidence for nonredundant functional cooperativity between NPC1 and NPC2 in lipid transport. Proceedings of the National Academy of Sciences. 2004;101(16):5886-91.

23. Infante RE, Wang ML, Radhakrishnan A, Kwon HJ, Brown MS, Goldstein JL. NPC2 facilitates bidirectional transfer of cholesterol between NPC1 and lipid bilayers, a step in cholesterol egress from lysosomes. Proceedings of the National Academy of Sciences. 2008;105(40):15287-92.

24. Erickson RP, Garver WS, Camargo F, Hossian GS, Heidenreich RA. Pharmacological and genetic modifications of somatic cholesterol do not substantially alter the course of CNS disease in Niemann–Pick C mice. Journal of inherited metabolic disease. 2000;23(1):54-62.

25. Pentchev PG, Gal AE, Booth AD, Omodeo-Sale F, Fours J, Neumeyer BA, et al. A lysosomal storage disorder in mice characterized by a dual deficiency of sphingomyelinase and glucocerebrosidase. Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism. 1980;619(3):669-79.

26. Siegel DA, Walkley SU. Growth of ectopic dendrites on cortical pyramidal neurons in neuronal storage diseases correlates with abnormal accumulation of GM2 ganglioside. Journal of neurochemistry. 1994;62(5):1852-62.

27. Zervas M, Dobrenis K, Walkley SU. Neurons in Niemann-Pick disease type C accumulate gangliosides as well as unesterified cholesterol and undergo dendritic and axonal alterations. Journal of Neuropathology & Experimental Neurology. 2001;60(1):49-64.

28. Abi-Mosleh L, Infante RE, Radhakrishnan A, Goldstein JL, Brown MS. Cyclodextrin overcomes deficient lysosome-to-endoplasmic reticulum transport of cholesterol in Niemann-Pick type C cells. Proceedings of the National Academy of Sciences. 2009;106(46):19316-21.

29. Walkley SU, Suzuki K. Consequences of NPC1 and NPC2 loss of function in mammalian neurons. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 2004;1685(1-3):48-62.

30. Higgins ME, Davies JP, Chen FW, Ioannou YA. Niemann–Pick C1 is a late endosome-resident protein that transiently associates with lysosomes and the trans-Golgi network. Molecular genetics and metabolism. 1999;68(1):1-13.

31. Patterson MC, Clayton P, Gissen P, Anheim M, Bauer P, Bonnot O, et al. Recommendations for the detection and diagnosis of Niemann-Pick disease type C: An update. Neurology: Clinical Practice. 2017;7(6):499-511.

32. Vanier MT, Gissen P, Bauer P, Coll MJ, Burlina A, Hendriksz CJ, et al. Diagnostic tests for Niemann-Pick disease type C (NP-C): A critical review. Molecular genetics and metabolism. 2016;118(4):244-54.

33. Patterson MC, Hendriksz CJ, Walterfang M, Sedel F, Vanier MT, Wijburg F, et al. Recommendations for the diagnosis and management of Niemann–Pick disease type C: an update. Molecular genetics and metabolism. 2012;106(3):330-44.

34. Wijburg FA, Sedel F, Pineda M, Hendriksz CJ, Fahey M, Walterfang M, et al. Development of a suspicion index to aid diagnosis of Niemann-Pick disease type C. Neurology. 2012;78(20):1560-7.

35. Wraith JE, Sedel F, Pineda M, Wijburg FA, Hendriksz CJ, Fahey M, et al. Niemann-Pick type C Suspicion Index tool: analyses by age and association of manifestations. Journal of inherited metabolic disease. 2014;37(1):93-101.

36. Millat G, Marçais C, Rafi MA, Yamamoto T, Morris JA, Pentchev PG, et al. Niemann-Pick C1 disease: the I1061T substitution is a frequent mutant allele in patients of Western European descent and correlates with a classic juvenile phenotype. Am J Hum Genet. 1999;65(5):1321-9.

37. Butters TD, Dwek RA, Platt FM. Inhibition of glycosphingolipid biosynthesis: application to lysosomal storage disorders. Chemical reviews. 2000;100(12):4683-96.

38. Zervas M, Somers KL, Thrall MA, Walkley SU. Critical role for glycosphingolipids in Niemann-Pick disease type C. Current Biology. 2001;11(16):1283-7.

39. Iturriaga C, Pineda M, Fernandez-Valero EM, Vanier MT, Coll MJ. Niemann–Pick C disease in Spain: clinical spectrum and development of a disability scale. Journal of the neurological sciences. 2006;249(1):1-6.

40. Wraith JE, Baumgartner MR, Bembi B, Covanis A, Levade T, Mengel E, et al. Recommendations on the diagnosis and management of Niemann-Pick disease type C. Molecular genetics and metabolism. 2009;98(1-2):152-65.

41. Crumling MA, Liu L, Thomas PV, Benson J, Kanicki A, Kabara L, et al. Hearing loss and hair cell death in mice given the cholesterol-chelating agent hydroxypropyl-β-cyclodextrin. PloS one. 2012;7(12):e53280.

42. Maceyka M, Milstien S, Spiegel S. The potential of histone deacetylase inhibitors in N iemann–P ick type C disease. The FEBS journal. 2013;280(24):6367-72.

43. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, et al. Population history and its impact on medical genetics in Quebec. Clin Genet. 2005;68(4):287-301.

44. Mayr E. Animal species and evolution. Animal species and evolution: Harvard University Press; 2013.

45. Moreau C, Vézina H, Labuda D. Founder effects and genetic variability in Quebec. Medecine Sciences: M/S. 2007;23(11):1008-13.

46. Bchetnia M, Bouchard L, Mathieu J, Campeau PM, Morin C, Brisson D, et al. Genetic burden linked to founder effects in Saguenay–Lac-Saint-Jean illustrates the importance of genetic screening test availability. Journal of Medical Genetics. 2021;58(10):653-65.

47. De Braekeleer M. Hereditary disorders in Saguenay-Lac-St-Jean (Quebec, Canada). Human heredity. 1991;41(3):141-6.

48. Bouchard G, De Braekeleer M. Histoire d'un genome: Population et genetique dans l'est du Quebec: Sillery, Québec: Presses de l'Université du Québec; 1991.

49. Heyer E. Genetic consequences of differential demographic behaviour in the Saguenay region, Quebec. American Journal of Physical Anthropology. 1995;98(1):1-11.

50. Heyer E, Tremblay M. Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. American journal of human genetics. 1995;56(4):970.

51. Brais B, Bouchard J-P, Xie Y-G, Rochefort DL, Chrétien N, Tomé F, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nature genetics. 1998;18(2):164-7.

52. Brunet G, Tome FM, Samson F, Robert JM, Fardeau M. Oculopharyngeal muscular dystrophy. A census of French families and genealogic study. Revue Neurologique. 1990;146(6-7):425-9.

53. Barbeau A, Sadibelouiz M, Roy M, Lemieux B, Bouchard JP, Geoffroy G. Origin of Friedreich's disease in Quebec. Canadian Journal of Neurological Sciences. 1984;11(S4):506-9.

54. Scriver CR. Human genetics: lessons from Quebec populations. Annual review of genomics and human genetics. 2001;2(1):69-101.

55. Bouchard G, Roy R, Casgrain B, Hubert M. Computer in human sciences: from family reconstitution to population reconstruction. From information to knowledge Conceptual and content analysis by computer. 1995:201-26.

56. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol. 2013;42(5):1285-99.

57. Jantzen R, Payette Y, de Malliard T, Labbé C, Noisel N, Broët P. Validation of breast cancer risk assessment tools on a French-Canadian population-based cohort. BMJ open. 2021;11(4):e045078.

58. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. 2001.

59. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? Journal of medical genetics. 2011;48(9):580-9.

60. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics. 2011;12(11):745-55.

61. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, et al. Whole exome capture in solution with 3 Gbp of data. Genome biology. 2010;11(6):1-8.

62. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clinical chemistry. 2009;55(4):641-58.

63. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics. 2010;42(9):790-3.

64. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, et al. Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. The American Journal of Human Genetics. 2010;87(3):418-23.

65. Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie FV, et al. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nature genetics. 2011;43(4):303-5.

66. Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, et al. Human gene mutation database—a biomedical information and research resource. Human mutation. 2000;15(1):45-51.

67. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome research. 2006;16(8):949-61.

68. Dilliott AA, Zhang KK, Wang J, Abrahao A, Binns MA, Black SE, et al. Targeted copy number variant identification across the neurodegenerative disease spectrum. Molecular Genetics & Genomic Medicine. 2022:e1986.

69. 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 biology. 2016;17(1):1-19.

70. Haas BJ, Zody MC. Advancing RNA-seq analysis. Nature biotechnology. 2010;28(5):421-3.

71. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature Reviews Genetics. 2019;20(11):631-56.

72. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature methods. 2011;8(6):469-77.

73. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome research. 2011;21(12):2213-23.

74. Lei R, Ye K, Gu Z, Sun X. Diminishing returns in next-generation sequencing (NGS) transcriptome data. Gene. 2015;557(1):82-7.

75. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

76. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

77. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC bioinformatics. 2017;18(1):1-12.

78. Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome biology. 2015;16(1):1-16.

79. Yang C, Wu P-Y, Tong L, Phan J, Wang M, editors. The impact of RNA-seq aligners on gene expression estimation2015 2015.

80. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. bioinformatics. 2015;31(2):166-9.

81. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.

82. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics. 2013;14(6):671-83.

83. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):1-21.

84. Nik S, Bowman TV. Splicing and neurodegeneration: Insights and mechanisms. Wiley Interdisciplinary Reviews: RNA. 2019;10(4):e1532.

85. Mills JD, Janitz M. Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. Neurobiology of aging. 2012;33(5):1012-e11.

86. Lord J, Baralle D. Splicing in the diagnosis of rare disease: advances and challenges. Frontiers in Genetics. 2021;12:1146.

87. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research. 2001;11(5):863-74.

88. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.

89. 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 research. 2005;15(8):1034-50.

90. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome research. 2005;15(7):901-13.

91. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31(13):3812-4.

92. Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de La Cruz M, et al. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proceedings of the National Academy of Sciences. 2003;100(10):5896-901.

93. 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.

94. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS biology. 2005;3(1):e7.

95. Ellegren H, Smith NGC, Webster MT. Mutation rate variation in the mammalian genome. Current opinion in genetics & development. 2003;13(6):562-8.

96. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014;46(3):310-5.

97. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.

98. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

99. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic acids research. 2012;41(D1):D64-D9.

100. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research. 2019;47(D1):D886-D94.

101. Kazazian, Jr., Boehm CD, Seltzer WK. ACMG recommendations for standards for interpretation of sequence variations. Genetics in Medicine. 2000;2(5):302-3.

102. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genetics in Medicine. 2008;10(4):294-300.

103. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine. 2015;17(5):405-23.

104. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome biology. 2017;18(1):1-12.

105. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics. 2009;25(14):1754-60.

106. 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.

107. 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 research. 2010;20(9):1297-303.

108. 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.

109. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164-e.

110. Poorthuis BJHM, Wevers RA, Kleijer WJ, Groener JEM, de Jong JGN, Van Weely S, et al. The frequency of lysosomal storage diseases in The Netherlands. Human genetics. 1999;105(1):151-6.

111. Meikle PJ, Hopwood JJ, Clague AE, Carey WF. Prevalence of lysosomal storage disorders. Jama. 1999;281(3):249-54.

112. Pinto R, Caseiro C, Lemos M, Lopes L, Fontes A, Ribeiro H, et al. Prevalence of lysosomal storage diseases in Portugal. European Journal of Human Genetics. 2004;12(2):87-92.

113. Poupetová H, Ledvinová J, Berná L, Dvoráková L, Kozich V, Elleder M. The birth prevalence of lysosomal storage disorders in the Czech Republic: comparison with data in different populations. J Inherit Metab Dis. 2010;33(4):387-96.

114. Wassif CA, Cross JL, Iben J, Sanchez-Pulido L, Cougnoux A, Platt FM, et al. High incidence of unrecognized visceral/neurological late-onset Niemann-Pick disease, type C1, predicted by analysis of massively parallel sequencing data sets. Genetics in Medicine. 2016;18(1):41-8.

115. Burton BK, Ellis AG, Orr B, Chatlani S, Yoon K, Shoaff JR, et al. Estimating the prevalence of Niemann-Pick disease type C (NPC) in the United States. Molecular genetics and metabolism. 2021;134(1-2):182-7.

116. Garver WS, Francis GA, Jelinek D, Shepherd G, Flynn J, Castro G, et al. The National Niemann–Pick C1 disease database: report of clinical features and health problems. American journal of medical genetics Part A. 2007;143(11):1204-11.

117. Burlina AP. Neurometabolic Hereditary Diseases of Adults: Springer; 2018.

118. Yanjanin NM, Vélez JI, Gropman A, King K, Bianconi SE, Conley SK, et al. Linear clinical progression, independent of age of onset, in Niemann–Pick disease, type C. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2010;153(1):132-40.

119. Patterson MC, Vecchio D, Prady H, Abel L, Wraith JE. Miglustat for treatment of Niemann-Pick C disease: a randomised controlled study. The Lancet Neurology. 2007;6(9):765-72.

120. Fecarotta S, Amitrano M, Romano A, Della Casa R, Bruschini D, Astarita L, et al. The videofluoroscopic swallowing study shows a sustained improvement of dysphagia in children with Niemann–Pick disease type C after therapy with miglustat. American Journal of Medical Genetics Part A. 2011;155(3):540-7.

121. Solomon BI, Smith AC, Sinaii N, Farhat N, King MC, Machielse L, et al. Association of miglustat with swallowing outcomes in Niemann-Pick disease, type C1. JAMA neurology. 2020;77(12):1564-8.

122. Porter FD, Scherrer DE, Lanier MH, Langmade SJ, Molugu V, Gale SE, et al. Cholesterol oxidation products are sensitive and specific blood-based biomarkers for Niemann-Pick C1 disease. Science translational medicine. 2010;2(56):56ra81-56ra81.

123. Vanier MT, Latour P. Laboratory diagnosis of Niemann–Pick disease type C: the filipin staining test. Methods in cell biology. 126: Elsevier; 2015. p. 357-75.

124. Godard B, Marshall J, Laberge C. Community engagement in genetic research: results of the first public consultation for the Quebec CARTaGENE project. Public Health Genomics. 2007;10(3):147-58.

125. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS computational biology. 2010;6(12):e1001025.

126. Millat G, Chikh K, Naureckiene S, Sleat DE, Fensom AH, Higaki K, et al. Niemann-Pick disease type C: spectrum of HE1 mutations and genotype/phenotype correlations in the NPC2 group. The American Journal of Human Genetics. 2001;69(5):1013-21.

127. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, et al. Molecular findings among patients referred for clinical whole-exome sequencing. Jama. 2014;312(18):1870-9.

128. Kumar D. From evidence-based medicine to genomic medicine. Genomic medicine. 2007;1(3):95-104.

129. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen—the clinical genome resource. New England Journal of Medicine. 2015;372(23):2235-42.

130. Tremblay M, Vézina H, Desjardins B, Bengtsson T, Mineau GP. Kinship and Demographic Behavior in the Past. 2008.

131. Evans WRH, Hendriksz CJ. Niemann-Pick type C disease–the tip of the iceberg? A review of neuropsychiatric presentation, diagnosis and treatment. BJPsych Bulletin. 2017;41(2):109-14.

132. Wraith JE, editor Lysosomal disorders2002 2002: Elsevier.

133. Sakiyama Y, Narita A, Osawa S, Nanba E, Ohno K, Otsuka M. Abnormal copper metabolism in Niemann–Pick disease type C mimicking Wilson's disease. Neurology and Clinical Neuroscience. 2014;2(6):193-200.

134. Malnar M, Hecimovic S, Mattsson N, Zetterberg H. Bidirectional links between Alzheimer's disease and Niemann–Pick type C disease. Neurobiology of disease. 2014;72:37-47.

135. Zavala L, Garretto NS, Arakaki T, Quiroga SR, Moron DG, Vega P, et al. Niemann Pick Type C as Presentation of Huntington-Like Syndrome (P4.043). Neurology. 2018;90(15 Supplement):P4.043.

136. 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.

137. Patterson M. Niemann-Pick disease type C. GeneReviews®[Internet]. 2020.

138. Tétreault M, Gonzalez M, Dicaire M-J, Allard P, Gehring K, Leblanc D, et al. Adult-onset painful axonal polyneuropathy caused by a dominant NAGLU mutation. Brain. 2015;138(6):1477-83.

139. Bras J, Guerreiro R, Hardy J. Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease. Nat Rev Neurosci. 2012;13(7):453-64.

140. Schneider SA, Tahirovic S, Hardy J, Strupp M, Bremova-Ertl T. Do heterozygous mutations of Niemann-Pick type C predispose to late-onset neurodegeneration: a review of the literature. J Neurol. 2021;268(6):2055-64.

141. Bremova-Ertl T, Sztatecsny C, Brendel M, Moser M, Möller B, Clevert DA, et al. Clinical, ocular motor, and imaging profile of Niemann-Pick type C heterozygosity. Neurology. 2020;94(16):e1702-e15.

142. Roy-Gagnon MH, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, et al. Genomic and genealogical investigation of the French Canadian founder population structure. Hum Genet. 2011;129(5):521-31.

143. Winsor EJ, Welch JP. Genetic and demographic aspects of Nova Scotia Niemann-Pick disease (type D). Am J Hum Genet. 1978;30(5):530-8.

144. Touma L, Labrecque M, Tetreault M, Duquette A. Identification and Classification of Rare Variants in NPC1 and NPC2 in Quebec. Sci Rep. 2021;11(1):10344.

145. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat Genet. 2015;47(4):400-4.

146. Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25(6):911-9.

147. 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. Am J Hum Genet. 2019;104(3):466-83.

148. Nicolau S, Choquet K, Bareke E, Shao YH, Brais B, O'Ferrall EK, et al. A Molecular Diagnosis of LGMDR1 Established by RNA Sequencing. Can J Neurol Sci. 2021;48(2):293-6.

149. Vasli N, Harris E, Karamchandani J, Bareke E, Majewski J, Romero NB, et al. Recessive mutations in the kinase ZAK cause a congenital myopathy with fibre type disproportion. Brain. 2017;140(1):37-48.

150. Millat G, Baïlo N, Molinero S, Rodriguez C, Chikh K, Vanier MT. Niemann-Pick C disease: use of denaturing high performance liquid chromatography for the detection of NPC1 and NPC2 genetic variations and impact on management of patients and families. Mol Genet Metab. 2005;86(1-2):220-32.

151. Thiffault I, Dicaire MJ, Tetreault M, Huang KN, Demers-Lamarche J, Bernard G, et al. Diversity of ARSACS mutations in French-Canadians. Can J Neurol Sci. 2013;40(1):61-6.

152. Fernandez-Valero EM, Ballart A, Iturriaga C, Lluch M, Macias J, Vanier MT, et al. Identification of 25 new mutations in 40 unrelated Spanish Niemann-Pick type C patients: genotype-phenotype correlations. Clin Genet. 2005;68(3):245-54.

153. Jiang X, Sidhu R, Porter FD, Yanjanin NM, Speak AO, te Vruchte DT, et al. A sensitive and specific LC-MS/MS method for rapid diagnosis of Niemann-Pick C1 disease from human plasma. J Lipid Res. 2011;52(7):1435-45.

154. Wijburg FA, Sedel F, Pineda M, Hendriksz CJ, Fahey M, Walterfang M, et al. Development of a suspicion index to aid diagnosis of Niemann-Pick disease type C. Neurology. 2012;78(20):1560-7.

155. Quebec Go. Blood and Urine Screening in Newborns 2021 [updated July 2020. Available from: https://www.quebec.ca/en/health/advice-and-prevention/screening-and-carrier-testing-offer/blood-and-urine-screening-in-newborns/diseases-screened.

156. Polo G, Burlina A, Furlan F, Kolamunnage T, Cananzi M, Giordano L, et al. High level of oxysterols in neonatal cholestasis: a pitfall in analysis of biochemical markers for Niemann-Pick type C disease. Clin Chem Lab Med. 2016;54(7):1221-9.

157. Wasserstein MP, Caggana M, Bailey SM, Desnick RJ, Edelmann L, Estrella L, et al. The New York pilot newborn screening program for lysosomal storage diseases: Report of the First 65,000 Infants. Genet Med. 2019;21(3):631-40.

158. Tängemo C, Weber D, Theiss S, Mengel E, Runz H. Niemann-Pick Type C disease: characterizing lipid levels in patients with variant lysosomal cholesterol storage [S]. Journal of lipid research. 2011;52(4):813-25.

159. Blom TS, Linder MD, Snow K, Pihko H, Hess MW, Jokitalo E, et al. Defective endocytic trafficking of NPC1 and NPC2 underlying infantile Niemann–Pick type C disease. Human Molecular Genetics. 2003;12(3):257-72.

160. Sun X, Marks DL, Park WD, Wheatley CL, Puri V, O'Brien JF, et al. Niemann-Pick C Variant Detection by Altered Sphingolipid Trafficking and Correlation with Mutations within a Specific Domain of NPC1. The American Journal of Human Genetics. 2001;68(6):1361-72.

161. Ługowska A. Chapter 21 - Niemann-Pick type C disease (NPC). In: Bukiya AN, Dopico AM, editors. Cholesterol: Academic Press; 2022. p. 525-51.

162. Chimenti MS, Khangulov VS, Robinson AC, Heroux A, Majumdar A, Schlessman JL, et al. Structural reorganization triggered by charging of Lys residues in the hydrophobic interior of a protein. Structure. 2012;20(6):1071-85.

163. Cupidi C, Frangipane F, Gallo M, Clodomiro A, Colao R, Bernardi L, et al. Role of Niemann-Pick Type C Disease Mutations in Dementia. Journal of Alzheimer's Disease. 2017;55:1249-59.

164. Mavridou I, Dimitriou E, Vanier MT, Vilageliu L, Grinberg D, Latour P, et al. The Spectrum of Niemann-Pick Type C Disease in Greece. In: Morava E, Baumgartner M, Patterson M, Rahman S, Zschocke J, Peters V, editors. JIMD Reports, Volume 36. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 41-8.

165. Zhang H, Wang Y, Lin N, Yang R, Qiu W, Han L, et al. Diagnosis of Niemann-Pick disease type C with 7-ketocholesterol screening followed by NPC1/NPC2 gene mutation confirmation in Chinese patients. Orphanet Journal of Rare Diseases. 2014;9(1):82.

166. Timmermans S, Tietbohl C, Skaperdas E. Narrating uncertainty: variants of uncertain significance (VUS) in clinical exome sequencing. BioSocieties. 2017;12(3):439-58.

167. 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. 2020;22(6):1005-14.

168. Scott C, Ioannou YA. The NPC1 protein: structure implies function. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 2004;1685(1-3):8-13.

169. Garver WS, Jelinek D, Meaney FJ, Flynn J, Pettit KM, Shepherd G, et al. The National Niemann-Pick Type C1 Disease Database: correlation of lipid profiles, mutations, and biochemical phenotypes. Journal of lipid research. 2010;51(2):406-15.

170. Chien YH, Peng SF, Yang CC, Lee NC, Tsai LK, Huang AC, et al. Long-term efficacy of miglustat in paediatric patients with Niemann-pick disease type C. Journal of Inherited Metabolic Disease: Official Journal of the Society for the Study of Inborn Errors of Metabolism. 2013;36(1):129-37.

171. Héron B, Valayannopoulos V, Baruteau J, Chabrol B, Ogier H, Latour P, et al. Miglustat therapy in the French cohort of paediatric patients with Niemann-Pick disease type C. Orphanet journal of rare diseases. 2012;7:36-.

172. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics. 2011;12(10):703-14.

173. Praggastis M, Tortelli B, Zhang J, Fujiwara H, Sidhu R, Chacko A, et al. A murine Niemann-Pick C1 I1061T knock-in model recapitulates the pathological features of the most prevalent human disease allele. Journal of Neuroscience. 2015;35(21):8091-106.

174. Schneider SA, Talelli P, Cheeran BJ, Khan NL, Wood NW, Rothwell JC, et al. Motor cortical physiology in patients and asymptomatic carriers of parkin gene mutations. Movement disorders. 2008;23(13):1812-9.

175. Kluenemann HH, Nutt JG, Davis MY, Bird TD. Parkinsonism syndrome in heterozygotes for Niemann–Pick C1. Journal of the Neurological Sciences. 2013;335(1):219-20.

176. Harzer K, Beck-Wödl S, Bauer P. Niemann-Pick disease type C: new aspects in a long published family–partial manifestations in heterozygotes. JIMD Reports-Volume 12: Springer; 2013. p. 25-9.

177. Bremova-Ertl T, Sztatecsny C, Brendel M, Moser M, Möller B, Clevert DA, et al. Clinical, ocular motor, and imaging profile of Niemann-Pick type C heterozygosity. Neurology. 2020;94(16):e1702-e15.

178. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, et al. Deleterious-and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. The American Journal of Human Genetics. 2012;91(6):1022-32.

179. Quaio CRD, Ceroni JRM, Cervato MC, Thurow HS, Moreira CM, Trindade ACG, et al. Parental segregation study reveals rare benign and likely benign variants in a Brazilian cohort of rare diseases. Scientific Reports. 2022;12(1):1-9.

180. Millot GA, Carvalho MA, Caputo SM, Vreeswijk MPG, Brown MA, Webb M, et al. A guide for functional analysis of BRCA1 variants of uncertain significance. Human mutation. 2012;33(11):1526-37.

181. Smith MU, Baldwin JT. Making sense of Hardy-Weinberg equilibrium. The American Biology Teacher. 2015;77(8):577-82.

182. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nature Reviews Genetics. 2012;13(8):565-75.

183. Yu C, Zhang S, Zhou C, Sile S. A likelihood ratio test of population Hardy-Weinberg equilibrium for case-control studies. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2009;33(3):275-80.

184. Kere J. Human population genetics: lessons from Finland. Annual review of genomics and human genetics. 2001;2(1):103-28.

185. Nevanlinna HR. The Finnish population structure A genetic and genealogical study. Hereditas. 1972;71(2):195-235.

186. Ostrer H, Skorecki K. The population genetics of the Jewish people. Human genetics. 2013;132(2):119-27.

187. Vanier MT, Rodriguez-Lafrasse C, Rousson R, Mandon G, Boué J, Choiset A, et al. Prenatal diagnosis of Niemann-Pick type C disease: current strategy from an experience of 37 pregnancies at risk. American journal of human genetics. 1992;51(1):111.

188. Patterson MC, Mengel E, Vanier MT, Moneuse P, Rosenberg D, Pineda M. Treatment outcomes following continuous miglustat therapy in patients with Niemann-Pick disease Type C: a final report of the NPC Registry. Orphanet journal of rare diseases. 2020;15(1):1-10.

189. sociaux CdmdlSedS. Déploiement de la première politique nationale sur les maladies rares 2022 [cited 2022.

190. Lange KI, Best S, Tsiropoulou S, Berry I, Johnson CA, Blacque OE. Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in Caenorhabditis elegans. Human molecular genetics. 2022;31(10):1574-87.

191. Guo H, Liu L, Nishiga M, Cong L, Wu JC. Deciphering pathogenicity of variants of uncertain significance with CRISPR-edited iPSCs. Trends in Genetics. 2021;37(12):1109-23.