UQAC
Université du Québec
à Chicoutimi

Découverte de l'histoire évolutive de l'ARSACS dans la population du Saguenay

Par Farnaz Farahmand Fard

Mémoire présenté à l'Université du Québec à Chicoutimi en vue de l'obtention du grade de Maîtrise ès sciences (M. Sc.) en Santé durable

Quebec, Canada

Résumé

L'ataxie spastique autosomique récessive de Charlevoix–Saguenay (ARSACS) est particulièrement fréquente dans la région du Saguenay–Lac-Saint-Jean (SLSJ), au Québec (Canada), en raison d'un fort effet fondateur. La principale variante causale du gène SACS (c.8844delT) est présente chez 94 % des patients, bien que plusieurs introductions indépendantes de variantes distinctes puissent avoir eu lieu. L'objectif de cette étude était d'identifier les points d'entrée potentiels de la mutation et de reconstituer l'histoire évolutive des haplotypes associés à l'ARSACS.

Des données généalogiques ont été recueillies pour 166 patients et 207 porteurs, dont 67 patients génotypés. À l'aide de la bibliothèque genlib dans R, nous avons évalué la complétude des généalogies, identifié les fondateurs parmi l'ensemble des patients et reconstruit les structures familiales liées à l'ARSACS à partir des coefficients de parenté entre porteurs. Pour l'analyse de génotypage, les segments d'identité par descendance (IBD) ont été générés à l'aide de RefinedIBD, ce qui nous a permis d'évaluer la proportion de segments partagés à chaque position génomique. Les haplotypes de différentes longueurs ont ensuite été définis à l'aide de geneHapR, et les patrons de déséquilibre de liaison (LD) ont été analysés et visualisés avec Haploview, permettant la reconstitution des blocs de LD. De plus, des arbres phylogénétiques d'haplotypes ont été inférés avec RAxML afin de retracer la transmission des haplotypes des fondateurs jusqu'aux patients contemporains. Enfin, une approche d'alignement des pedigrees a été appliquée pour intégrer les phylogénies basées sur le séquençage aux généalogies, dans le but d'identifier le fondateur original de la variante c.8844delT.

Notre analyse généalogique a permis d'identifier les couples ayant la plus forte contribution génétique au bassin de gènes des porteurs. Le partage d'IBD a révélé un signal marqué sur le chromosome 13 correspondant au locus SACS. Des haplotypes de différentes longueurs (0,8, 2,2, 4 et 8 Mb) entourant la variante c.8844delT ont été reconstitués, le plus commun (0,8 Mb) étant observé chez 86 % des patients, tandis que les autres présentaient des haplotypes plus rares. En nous concentrant sur les haplotypes plus longs (2,2 Mb) chez les patients homozygotes pour c.8844delT, nous avons affiné les relations phylogénétiques et, grâce à l'alignement des pedigrees, identifié un couple ayant introduit la variante dans la population du SLSJ il y a plus de 400 ans.

Cette approche intégrée, combinant analyses généalogiques et génomiques, met en évidence à la fois le fort effet fondateur et la dynamique évolutive des haplotypes de l'ARSACS dans le SLSJ, offrant ainsi une meilleure compréhension de l'histoire démographique de ce désordre neurogénétique rare.

Summary

The autosomal recessive spastic ataxia of Charlevoix–Saguenay (ARSACS) is particularly frequent in the Saguenay–Lac-Saint-Jean (SLSJ) region of Quebec, Canada, due to a strong founder effect. The main causal variant of the SACS gene (*c.8844delT*) is present in 94% of patients, although several independent introductions of distinct variants may have occurred. The objective of this study was to identify the potential entry points of the mutation and to reconstruct the evolutionary history of the haplotypes associated with ARSACS.

Genealogical data were collected for 166 patients and 207 carriers, including 67 genotyped patients. Using the GENLIB library in R, we assessed the completeness of the genealogies, identified the founders among all patients, and reconstructed the family structures related to ARSACS based on kinship coefficients between carriers. For the genotyping analysis, segments of identity by descent (IBD) were generated using RefinedIBD, allowing us to evaluate the proportion of shared genomic segments at each position. Haplotypes of various lengths were then defined using geneHapR, and linkage disequilibrium (LD) patterns were analyzed and visualized with Haploview, enabling the reconstruction of LD blocks. In addition, phylogenetic trees of haplotypes were inferred with RAxML to trace the transmission of haplotypes from founders to contemporary patients. Finally, a pedigree alignment approach was applied to integrate sequencing-based phylogenies with genealogies, in order to identify the original founder of the *c.8844delT* variant.

Our genealogical analysis identified the couples with the greatest genetic contribution to the carrier gene pool. IBD sharing revealed a strong signal on chromosome 13 corresponding to the SACS locus. Haplotypes of different lengths (0.8, 2.2, 4, and 8 Mb) surrounding the *c.8844delT* variant were reconstructed, with the most common one (0.8 Mb) observed in 86% of patients, while others showed rarer haplotypes. Focusing on the longer haplotypes (2.2 Mb) among patients homozygous for *c.8844delT*, we refined phylogenetic relationships and, through pedigree alignment, identified a couple who introduced the variant into the SLSJ population more than 400 years ago.

This integrated approach, combining genealogical and genomic analyses, highlights both the strong founder effect and the evolutionary dynamics of ARSACS haplotypes in the SLSJ region, providing a deeper understanding of the demographic history of this rare neurogenetic disorder.

Contents

Thanks

I would like to express my heartfelt gratitude to all the individuals and institutions who contributed to the completion of this thesis. I am especially grateful to my research director, Professor Simon Girard, for his unwavering support, availability, and invaluable guidance throughout this project. Thank you for welcoming me into the Genopop laboratory and for fostering an environment that encourages both professional and personal growth.

I am deeply thankful to Claudia Moreau for her constant help, her readiness to answer my questions, and her ongoing support. Beyond that, I truly appreciate her critical mindset, which challenged me to push my limits and seek ever more rigorous solutions. Thank you for your contagious enthusiasm for population genetics and for your relentless curiosity in the search for answers.

I am also grateful to my colleagues and friends in the Genopop laboratory, past and present, not only for their inspiration and support but also for the wonderful memories we shared—the unforgettable anecdotes, the profound conversations, and even the most delightfully absurd discussions.

I would like to extend my sincere thanks to the Fondation de l'Ataxie Charlevoix-Saguenay for their financial support during my master's studies.

Finally, I am profoundly grateful to all the participants of the Groupe de recherche interdisciplinaire sur les maladies neuromusculaires (GRIMN) clinic, whose contributions made this research possible. I would also like to acknowledge the BALSAC group for providing the genealogical data and the Centre d'expertise et de services Génome Québec for supplying the genetic data.

List of Abbreviations

DNA	Deoxyribonucleic acid
ARSACS	Autosomal Recessive Spastic Ataxia of Charlevix-Saguenay
CaG	CARTAGENE
IBD	Identity-by- descent segment
INDEL	Insertions and deletions
MAF	Minor allele frequency
PCA	Principal components analysis
SLSJ	Saguenay–Lac-St-Jean
SNP	Single nucleotide polymorphism
RAxML	Randomized Axelerated Maximum Likelihood
UMAP	Uniform manifold approximation and projection
UrbanQc	Urban regions of Quebec
WGS	Whole genome sequencing
CSLSJ	Charlevoix–Saguenay–Lac-Saint-Jean
HEPN	Higher Eukaryotes and Prokaryotes Nucleotide-binding
IF	Intermediate Filaments
NF	Neurofilaments
HDF	Human Dermal Fibroblasts
GTP	Guanosine triphosphate
ATP	Adenosine Triphosphate
KO cells	Knockout cells
HSP	Heat Shock Protein
FAK	Focal Adhesion Kinase
PTEN	Phosphatase and TENSIN homolog
LD	Linkage Disequilibrium
GTR	General time-reversible
VCF	Variant Call Format
cM	Centimorgan
Mb	Megabase
FAK	Focal Adhesion Kinase
MRCA	Most Recent Common Ancestor
CPU	Central Processing Unit

List of Figures

1 Introduction

1.1 ARSACS

Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) is clinically characterized by a progressive cerebellar ataxia, peripheral neuropathy, and spasticity. ARSACS is inherited in an autosomal recessive manner. The first manifestation of ataxia can include: Most commonly, a slowly progressive gait disorder that appears unsteady and predisposes to unexpected falls; Disequilibrium ("dizziness"), which may lead to an evaluation for peripheral vestibular dysfunction; Hand and finger clumsiness or tremor, which may raise the possibility of essential tremor or even parkinsonism; Rarely, double vision, which could lead to an evaluation by an optometrist or ophthalmologist. At disease onset, these manifestations may be intermittent or evident only at certain times (e.g., later in the day, when tired, after consuming alcohol). The manifestations typically become constant and slowly worsen.

Disease onset of classic ARSACS is often in early childhood, leading to delayed walking because of gait unsteadiness in very young toddlers, while an increasing number of individuals with disease onset in teenage or early adult years are now being described. Typically, the ataxia is followed by lower-limb spasticity and later by peripheral neuropathy – although pronounced peripheral neuropathy has been observed as a first sign of ARSACS. Oculomotor disturbances, dysarthria, and upper-limb ataxia develop with slower progression than the other findings. Brain imaging demonstrates atrophy of the superior vermis and the cerebellar hemisphere with additional findings on MRI, such as linear hypointensities in the pons and hyperintense rims around the

thalami. Many affected individuals have yellow streaks of hypermyelinated fibers radiating from the edges of the optic disc noted on ophthalmologic exam, and thickened retinal fibers can be demonstrated by optical coherence tomography. Mild intellectual disability, hearing loss, and urinary urgency and incontinence have been reported in some individuals [1].

1.1.1 Prevalence of ARSACS

More than 200 mutations have been discovered in the SACS gene around the world. Besides French Canadians, emerging studies discovered SACS mutations in several other countries. However, several cases of ARSACS may be unreported [2].ARSACS, an early onset cerebellarataxia inherited in an autosomal recessive manner, arises from mutations in SACS that impair sacsin function. The disorder was first characterized in the Charlevoix–Saguenay–Lac-Saint-Jean (CSLSJ) population of Quebec through linkage disequilibrium studies in1978[2] ,[3].

For the period 1941-1985, the ARSACS incidence at birth in SLSJ (Saguenay-Lac-Saint-Jean) were estimated at 1/1,932[4] but is now declining because of voluntary carrier screening liveborn infants. Now, the carrier rate is 1/22 inhabitants [5]. These numbers should be considered conservative, since it cannot be excluded that a few cases may not have been ascertained. However, by the age of 30 years old, most, if not all, of the cases are ascertained. Since no differences in the incidence at birth were observed between those born in the 1940s and 1950s and those born between 1971 and 1985, it is likely that the underestimation is very low.

1.2 Autosomal recessive inheritance pattern of ARSACS

Genetics is described as the study of genes and heredity, therefore what is transmitted to us by our parents [6]. This transmission of characteristics from an ancestor to his descendant is done by the arrangement of four nucleotides A, T, C and G which make up our deoxyribonucleic acid (DNA) [4],[5]. Heredity is the transmission of characteristics from a living being to the next generation through DNA and this occurs during reproduction. These observable characteristics are called phenotype. Thus, each individual is composed of 50% of its mother's genome and 50% of its father's genome; and each position of the genome (for autosomes) is found in two copies called alleles.[7] This brings the notion of homozygosity, where the 2 alleles present are identical, and that of heterozygosity, where the 2 alleles present are different.

In diploids, most chromosomes exist in pairs (same length, centromere location, and banding pattern), with one set coming from each parent. These chromosomes are called **autosomes**. However, many species have an additional pair of chromosomes that do not look alike. These are **sex chromosomes** because they differ between the sexes. In humans, males have one of each while females have two X chromosomes. Autosomes are those chromosomes present in the same number in males and females, while sex chromosomes are those that are not [8].

Diseases transmitted in an autosomal recessive pattern are present when an individual inherits two disease alleles. Following Mendel’s Law of Segregation, this means that both parents must have at least one disease allele that they can pass on to their child.

The most common situation of an autosomal recessive disease occurs when the parents are each carrier or heterozygous (Gg), illustrated in Figure1. Children of carrier parents have a 25% chance of inheriting the disorder. This value is obtained by using the Punnet square model used in genetics[9]

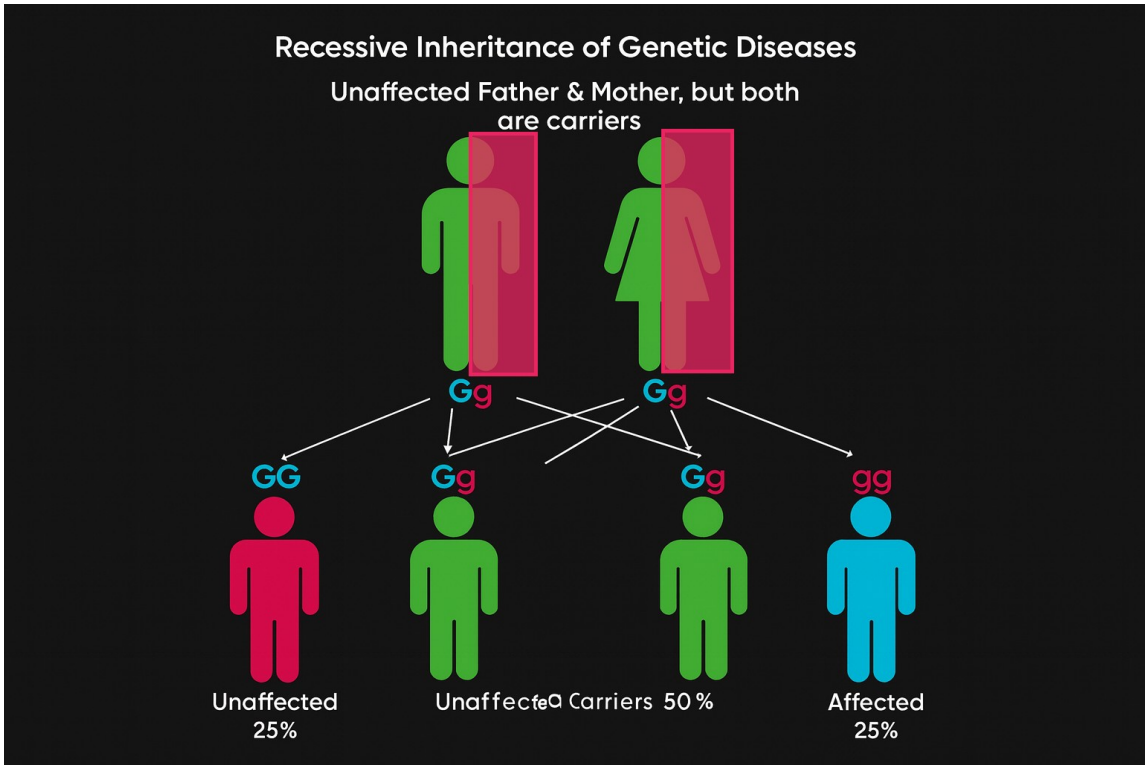

Figure 1 Autosomal recessive inheritance.
Each parent has a 50% chance of passing on the disease allele. Using the multiplication rule of probability, there is a 50% chance that the father passes on his disease allele and a 50% chance that the mother passes on her disease allele; 50% x 50% = 25% .So with the mating of carrier parents, there is a 25% chance that the child will be affected, a 50% chance that the child would be a carrier, and 25% chance that they would be homozygous dominant and unaffected. Image courtesy O.Chaigasame.

1.3 Sacsin (SACS) Gene

The sacsin gene (SACS) is located on chromosome 13, with 145,075 bases, and is oriented in the minus strand of DNA. SACS comprises 10 exons, with nine coding exons, and the 10th exon contains 11,487 base pairs, notable as the longest exon among vertebrates [10],[11]. The SACS gene encodes a large 520-kDa multidomain protein of 4579 amino acids, called sacsin [12]. Sacsin is expressed in several different tissues, with higher expression in the central nervous system or skin and lower expression in the pancreas and skeletal muscle. In the brain, sacsin expression is the highest in the motor system, including the cerebellum, granular system and in Purkinje cells[11].

1.3.1 Sacsin Protein Domains and Functions

Sacsin has five domains: ubiquitin-like domain (Ubl), three large sacsin internal repeats (SIRPT1, SIRPT2, SIRPT3), xeroderma pigmentosum C-binding domain (XPCB), J-domain (DNAJ) and higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain. These domains could have different functions and interacting partners (Figure 2). However, they could also cooperate in different pathways[13].

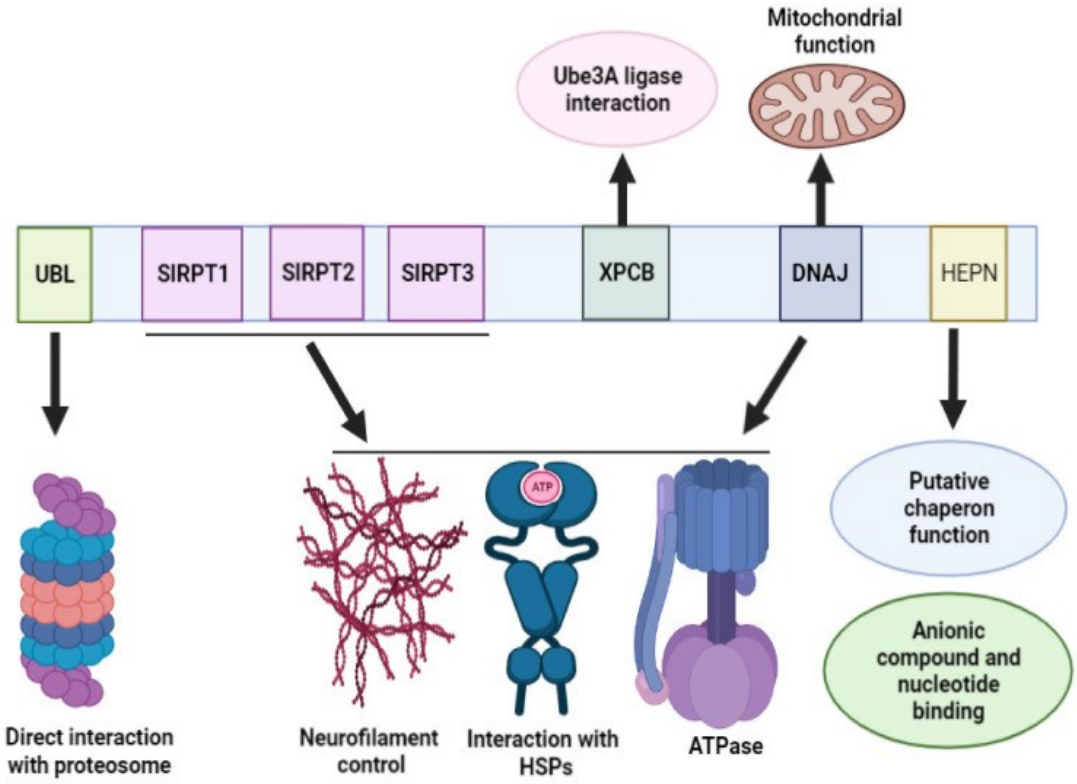

Figure 2 Schematic showing the domain structure of the of sacsin protein and the potential functions of different domains.

It contains several different domains, including the ubiquitin-like (UBL) domain in the N-terminal region, three sacsin internal repeat (SIRPT or SRR) domains, the helical XPC-binding domain, a sacsin J-domain and the higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain in the C-terminal region.

The N-terminal region of sacsin could interact with proteasomes. Additionally, it may play a role in regulating protein folding by interacting with heat shock proteins. The Ubl domain could interact directly with the proteasome system through the 19S cap and 26S proteasomes and be involved in the degradation pathway [11],[14]. The SIRPT domains contain a homologous region, the Hsp90 chaperone [3],[15],[16],[17], which was suggested to be

involved in ATPase activities. ATP hydrolysis is important for proper sacsin function. Dysfunctions of the SIRPT domain were associated with reduced or lost ATP hydrolysis, along with the J-domain, SIRPT domains could stimulate the ATPase activity of Hsp70 [17],[18] . The XPC-binding domain (XPCB) could bind the Ube3A ubiquitin ligase. It may be possible that dysfunctions in Ube3A–sacsin interaction could impact the onset of Angelman syndrome-related ataxia [19]. Both ARSACS and Angelman syndrome include ataxia as a clinical feature; however, their molecular causes are distinct. Angelman syndrome results from dysfunction of the *UBE3A* gene, whereas ARSACS is caused by mutations in the *SACS* gene, which encodes sacsin.[20],[21]

The J-domains can enhance the protein–protein interactions and regulate the activity of heat shock proteins, including Hsp70. A J-domain contains Hsp40 homologous sequences and may impact homeostasis. An interaction between sacsin, Hsp70 and the ubiquitin proteosome system may be involved in defensive mechanisms against abnormal protein aggregation [11]. Sacsin was suggested to play a role in controlling the homeostasis between intermediate filaments (IFs) and neurofilaments (NFs)[12]. If sacsin is missing, nerve cells could contain abnormal bundles of NFs. Patient-derived HDF cells showed abnormal IF (vimentin) distribution and broken microtubule organization. In ARSACS cells, the misfolded IFs aggregated and formed a cage-like structure around the microtubule organization center. These aggregates may result in abnormal clearance and autophagy [22]. Adding the SIPRT and J-domain into motor neurons from sacsin knockout mice resulted in a reduced amount of NF bundles. Both domains could prevent the assembly of NFs [12]. Treatment with SacsJ-myc-TAT in SACS+/+ motor neurons resulted in the induction of IF and NF disassembly.

SacsJ-myc-TAT is a synthetic peptide based on the J domain of the sacsin protein, and it's been used experimentally to study (and partially rescue) cellular defects in ARSACS[23]. In Sacs–/– motor neurons, the NF bundles were resolved, and the NF network was restored [24]. These data suggest that the J-domain could impact the regulation of IF-NF assembly and disassembly directly [12],[22],[23].

The function of the HEPN domain was initially unclear[16]. In eukaryotes, the HEPN domain may have nucleotide binding activity, but it could also bind anionic compounds in neurons[25],[26] ,[27],[28]. HEPN domains could dimerize and form a high-affinity site for GTP binding, and potentially impact the chaperon activity of the sacsin protein. The HEPN domain may be involved in the elevation of the ATP/GTP concentration, and in the sacsin–Hsp70 interaction. The HEPN domain may also co-operate with the J-domain, and they contribute to nucleotide binding. Mutations in the HEPN region were suggested to disrupt the nucleotide-binding activity and result in abnormal folding/oligomerization of sacsin [12],[16], [27],[28].

A recent study by Romano et al[29]. revealed that, besides controlling the filament architecture, sacsin could play a crucial role in cell adhesion, microtubule organization and trafficking proteins. The mentioned authors provided an extensive study on the possible functions of sacsin. Knockout of the SACS gene in SH-SY5Y cell lines revealed several disease-associated changes. SH-SY5Y cells are human neuroblastoma-derived neuronal cells commonly used to model neurodegenerative mechanisms. They express sacsin and can be genetically modified to mimic ARSACS cellular phenotypes[30],[21]. Microtubule

organization and dynamics were altered in KO cells, microtubule polymerization was enhanced, and the movement of tubules became abnormal. Sacsin may also regulate Tau phosphorylation by interacting with tyrosine kinase enzymes. Several kinesin proteins were hyperphosphorylated, resulting in disturbances in mitochondrial movements. Tau and STMN1 pS16 were also hyperphosphorylated. These findings revealed that sacsin could possibly interact with tyrosine kinases. Additionally, sacsin interactions with HSP proteins could be critical in microtubule organization and stabilization. Non-functional sacsin also resulted in disrupted focal adhesion and dynamics, which could result in disturbances in axonal growth, synaptic formation and balance in the brain. Focal adhesion kinases were suppressed in SACS KO cells, and modulation of FAK-PTEN pathways may be beneficial in the case of cellular deficits. Disturbances were found in the adhesion mechanisms, since the membrane-bound adhesion molecules and cell adhesion molecules were both mis-localized. These findings suggest that, besides abnormal trafficking and cellular interactions, localizations could contribute to the reduced interaction between heat shock proteins and sacsin, leading to ARSACS pathology. Additionally, sacsin may interact with exosomes, which could also impact several chaperons or microtubule related mechanisms. Figure 3 demonstrates that sacsin may serve as a critical regulator of different cellular processes, including chaperon activity, the transport of vesicles and microtubule and filament organization [31]

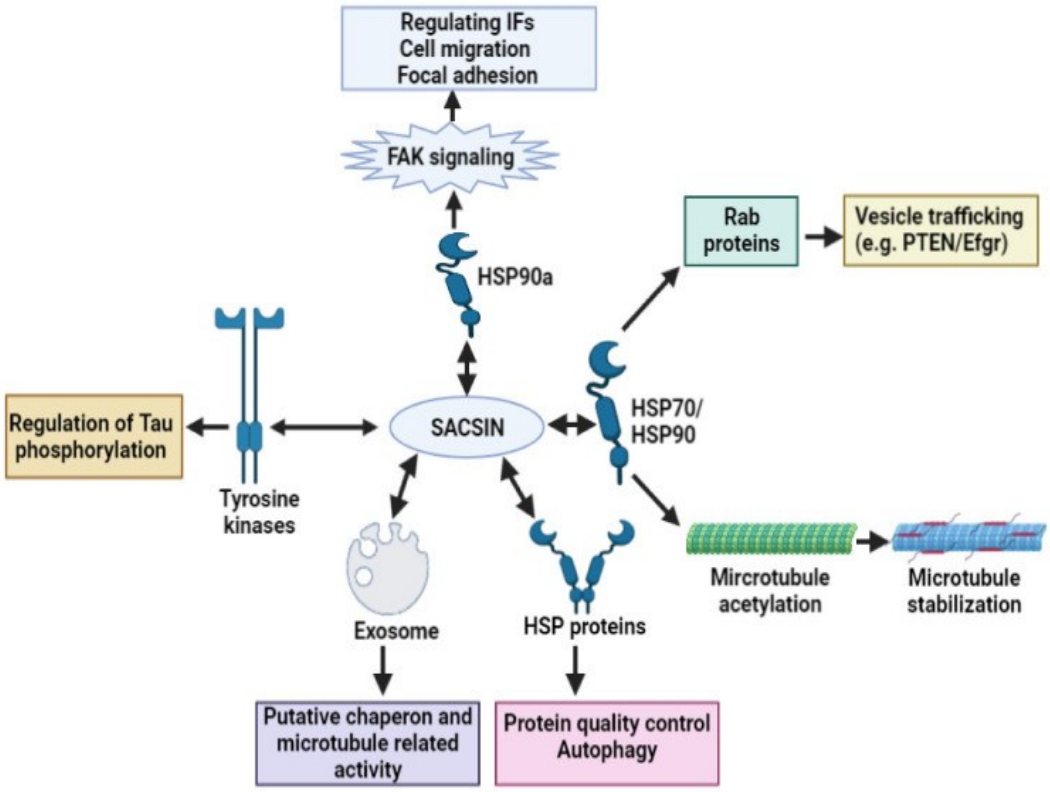

Figure 3 Possible impact of sacsin in different cellular processes

Saccin protein is involved in many processes, including chaperon functions, microtubule organization and vesicle trafficking.

1.3.2 SACS genetics and mutations

More than 200 pathogenic mutations have been reported in the SACS gene all around the world. The majority of mutations were found in exon 10, which was verified as the longest exon among vertebrates [32]. The majority of mutations may result in non-functional SACS or reduced sacsin expression. Disease-associated variants can be either homozygous or compound heterozygous. Compound heterozygotes (CH) in classical genetics refer to two

different mutations at a particular gene, one on each chromosome, both typically rare and deleterious together causing an autosomal recessive trait[33]. The phenotypes of mutations may be diverse; besides the classical phenotypes (ataxia, spasticity), additional atypical symptoms may also be present (intellectual disability, memory dysfunctions). Mutations can be either missense mutations, STOP codon mutations or frameshift variants (Figure 4). The missense mutations could result in lower stability and abnormal conformation of sacsin. Normally, the misfolded sacsin goes through co-translational ubiquitination and degradation by the proteosome system. If degradation does not happen, mutant sacsin could potentially aggregate in cytosol. Based on other neurodegenerative diseases, the potential aggregation of sacsin may result in putative additional gain-of-function toxicity. The frameshifts or nonsense variants may repress the translation, resulting in the degradation of the SACS transcript. All of the above could be associated with missing sacsin expression or insufficient amounts of sacsin protein [34].

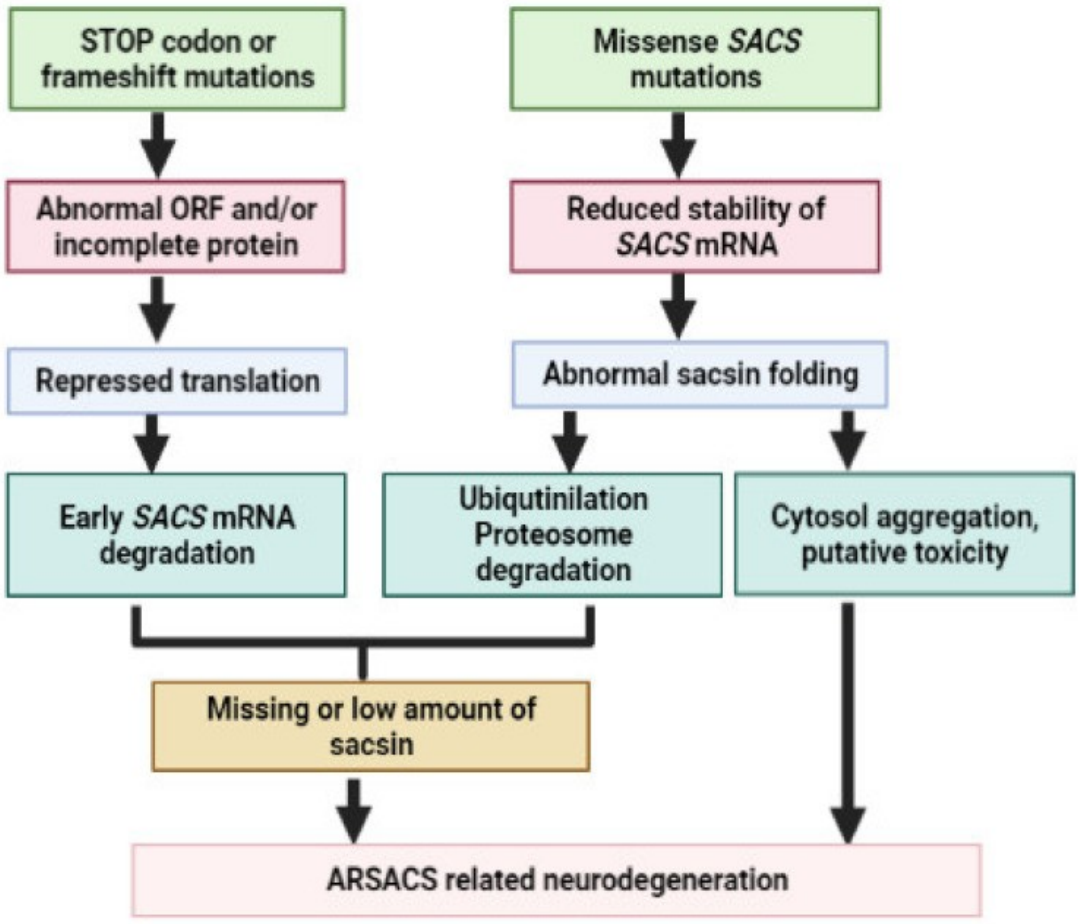

Figure 4 Potential effects of mutations in SACS gene.

Schematic representation of the causal ARSACS mutation pattern in the French-Canadian population. Both mutation types ultimately result in absent or reduced sacsin protein, leading to progressive neurodegeneration.

The first SACS mutations were reported in a CSLSJ family. In this population, more than 300 patients were observed who could have been the descendants of a single founder [3],[32]. Several mutations were found in the affected population in the Quebec province. Initially, two pathogenic mutations were identified, c.6594delT (mutation location on transcript) or

p.I2949Ffs*4 (mutation location on protein) and c.5254C>T (p.Q1752X, 3%). Both mutations could be related to the truncation of the sacsin protein [3],[32],[34]. Later, additional variants were observed in the Quebec population. The indel mutation c.8844delT (p.I2949FfsX2952) is the most frequent in French Canadian population. c.8844delT in SACS (HGVS: NM_014363.6:c.8844del; protein consequence p.Ile2949Phefs*4) is a single-base deletion that causes a frameshift and premature stop codon, leading to loss of functional sacsin protein and is a pathogenic, founder mutation frequently observed in ARSACS patients from Québec[35], [31].

However, additional variants also appeared in several ARSACS cases in this region, such as c.7504C>T (p.Arg2502Ter) and c.814C>T (p.R272C). The c.814C>T (p.R272C) mutation in SACS is a missense variant that likely destabilizes sacsin, reducing its normal function. Though some mutant protein may still be produced, its altered stability or folding contributes to the pathogenesis of ARSACS by impairing sacsin's molecular chaperone roles[36],[31],[15] . The c.7504C>T mutation in *SACS* results in a premature stop (p.Arg2502Ter), truncating much of the sacsin protein, and is classified as a pathogenic, founder variant in ARSACS due to its severe loss-of-function impact[37],[1],[36].

Large deletions in French Canadians may not be either common or a lethal variant. The diverse phenotypes of the disease may appear due to the partial loss of function in the sacsin protein [36],[38],[39].

1.4 The settlement of Quebec

The large European settlement movements of the 17th century led to the formation of many founder populations[40]. This includes, among others, the French-Canadian population of Quebec. Before the arrival of Europeans, the territory was inhabited by the First Peoples. However, their population rapidly declined after the first contacts with Europeans [41]. The French-speaking population of Quebec currently comprises approximately 8.9 million individuals, of whom approximately 74% are descended from the 8,000 to 10,000 French immigrants who arrived in the province between 1608 and 1759[42],[43],[44]. During this period, approximately 25,000 European immigrants arrived in the province, but less than half settled in the region permanently [44],[45]. The first occupied territories were concentrated between the regions of Quebec City and Montreal, which were founded in 1608 and 1642 respectively [44]. These regions now correspond to major urban centers that received a large number of immigrants of various nationalities who mixed with the population[46],[47]. This initial settlement laid the foundations for the founding effect of the Quebec population of French-Canadian origin [48].

Subsequently, the development primarily continued through natural population increase [49]. The European settlement of Quebec was then supported by immigrants from France, specifically from the central and western provinces [50]. Immigration was mainly for fur trade and land clearing work, so men outnumbered women by three to one[51],[52]. About two-thirds of the pioneer women arrived between 1663 and 1673 with the contingents of the Filles du Roi [52],[53]. The arrival of these women marked the beginning of significant

population growth due to natural increase, as women had an average of ten children before the age of 50 [53],[54].

The province, then called New France, was conquered by the English in 1759, thus ending French immigration. At that time, the population was around 65,000 inhabitants [55]. After the British Conquest, immigrants mainly came from England, Scotland, and Ireland and settled in the regions of Quebec and Montreal [49],[55],[56]. Despite France losing control of the province, the population retained its French-speaking heritage and Catholic religion. Therefore, a limited number of unions were formed with Anglo-Protestant immigrants [57],[49] . Subsequent waves of migration did not significantly contribute to the diversification of the population's genetic pool [49].

In the early 19th century, a series of economic, social, and cultural factors favored settlement on new lands [50],[55]. The high population growth led to overpopulation in inhabited areas, making territorial expansion necessary to prevent young people from emigrating to the United States [50]. Individuals dispersed to successively establish new regions, thus forming distinct communities. The settlement gradually extended across the territory, continuing its expansion into new communities after the first third of the 19th century [58],[50],[59]. As a result, stratification exists within the Quebec genetic pool due to geographical isolation, successive migrations, and genetic drift [60],[61],[62],[63]. This is particularly true in remote or sparsely populated regions [64]. Recent studies on the structure of the Quebec population have observed regional structures that differ from metropolitan areas [61]. Notably, a distinction has been observed in the Acadian population of Gaspésie, a

population founded by a small number of founding families and having remained relatively isolated [61],[62],[63]. Another regional population that has significantly differentiated from the rest of Quebec is that of Saguenay–Lac-St-Jean (SLSJ), a region that has experienced a significant regional founder effect that has been well characterized [61],[62],[65] .

1.4.1 Saguenay–Lac-St-Jean

The Saguenay–Lac-St-Jean (SLSJ) region opened up to settlement around the mid-19th century. Around 1830-1840, due to population expansion, families from Charlevoix began struggling to provide good land for their children, which influenced migration [50]. Migration occurred with entire families, ensuring the establishment of their children [66]. The expansion of the territory occurred without definitive borders, and the population of SLSJ experienced significant growth from a sample primarily composed of families from the Charlevoix region [50],[67]. Ten years after the settlement opened, in 1850, the population was around 5,000 inhabitants [67]. Until the 20th century, there was low mortality and extremely high fertility. Between 1861 and 1961, the population of SLSJ increased 25 times compared to a fivefold increase for the entire province [50],[68]. Over time, the availability of land began to decrease, and from 1920, a voluntary limitation of the number of births started to be observed [66].

Historical demographic research has traced the contribution of Quebec regions that contributed to the initial settlement of SLSJ. Slightly more than 80% of the genetic pool of the SLSJ population comes from about 2,600 individuals who settled in the province in the 17th century [69],[70]. Some of these individuals carried rare mutations associated with

genetic diseases [71]. The promotion of demographic expansion of a population relatively homogenized by family ties has left traces in the genetics of current individuals [50],[66],[69]. Thus, there is an increased prevalence of several rare genetic diseases in the SLSJ region, while they are relatively rare in the rest of the province [58],[57],[72],[32]. Numerous studies based on the region's demography and genetics have led to the establishment of a genetic screening program. Carrier tests are offered to individuals who wish to have children and whose grandparents are from SLSJ, Charlevoix, or the Côte-Nord. They currently cover four genetic diseases that are more frequent in individuals from these regions [73]. These four common hereditary diseases are congenital lactic acidosis, Charlevoix-Saguenay recessive spastic ataxia, hereditary sensory and motor neuropathy with or without agenesis of the corpus callosum, and hereditary tyrosinemia type 1 [74],[73]. However, these diseases are not the only ones with a higher prevalence in the region. A recent study identified 80 rare pathogenic variants that are more frequent in SLSJ than in the more urban regions of Quebec due to the regional founder effect [75].

1.5 Founder effect

A founder effect occurs when a portion of the population becomes isolated from the rest of the group by establishing itself in a new territory and forming the basis of a new population[76] . Following this bottleneck, the genetic pool of the new population will be constituted by the genetic material of each of its founders. Over time, the new sub-population will display a different genetic profile from that of the initial group. Significant genetic and phenotypic differences are present among the different populations [77]. In populations that have experienced a recent founder effect, a reduced allelic diversity is observed. The same variants

are found more frequently, and this can include pathogenic variants. This explains why certain hereditary diseases may be more common in populations that have undergone a founder effect [78]. Indeed, the fixation or loss of alleles in the new population occurs randomly, reducing the impact of fitness [79]. Genetic drift reduces the efficiency of natural selection in eliminating deleterious variants [77]. In populations affected by the founder effect, the focus is not on a higher occurrence of rare genetic variants but rather on how these variants are distributed differently.

The founder effect is directly related to the demographic split between the source population and the new population, making it essential to understand the demographic history of populations to interpret observable genetic patterns in current populations [80]. In demography, several variables help describe and analyze populations. These include changes such as birth and death rates, as well as inbound and outbound migration movements [81]. These factors influence not only the size of populations but also their genetic composition. Indeed, most founder-effect populations are not completely isolated. Migration flows can enrich the genetic pool of the population and increase its diversity. Large migrations can even prevent allele fixation provided there is sufficient genetic flow to compensate for the initial loss of diversity [82],[83]. For allelic diversity to increase, new individuals must integrate into the local population by reproducing with individuals already established in the territory [82].

To claim that a disease is more frequent due to a founder effect, it is necessary that several unrelated patients are affected by the same disease. This disease must also be caused by the same genetic mutation, whose origin can be traced to a common ancestor [84]. However, it is

not essential to precisely identify this ancestor. Its existence can be inferred from the genetic context surrounding the mutation. Indeed, in carriers of a founder mutation, it should be contained within a segment of DNA shared among the carriers [84].

In isolated populations or those that have experienced a founder effect, individuals may be related by more or less distant common ancestors. Thus, they will display identical-by-descent (IBD) segments that allow inferring the nature of the relationships between individuals. DNA segments are considered IBD if they are identical and originate from a common ancestor [85]. The more recent the common ancestor between two individuals, the longer and more numerous the IBD segments will be. Indeed, with each meiosis, the segments are cut by the process of recombination, causing the number and length of segments to decrease with each generation [86]. Meiosis is a type of cell division in sexually reproducing organisms that reduces the number of chromosomes in gametes (the sex cells, or egg and sperm)[87].Since founder variants originate from a common ancestor, they should be transmitted with an IBD segment [84],[88]. If carriers display an IBD segment around the variant's position, it indicates introduction by a founder ancestor, followed by transmission in the population through genetic drift [84].

1.5.1 Founder effect expansion in Saguenay–Lac-St-Jean

Founder variants have been observed in various populations around the world [89]. The genetic structure of Saguenay–Lac-St-Jean population is considered to be the product of successive migration waves corresponding to a successive founder effect (figure 5): (a) the first founder event took place during the French regime (1608–1760) when approximately 10

000 immigrants settled in the Saint Lawrence valley, in the west of the Province of Quebec. They account for the major part of the contemporary French-Canadian gene pool[44]; (b) the second founder event started at the end of the 17th century, when inhabitants from Quebec city and Côte-de-Beaupré (on the north shore of the Saint Lawrence river) moved to the Charlevoix region where 600 individuals settled between 1675 and 1840[50]; (c) the third founder event corresponds to the settlement of the SLSJ region. It started in the 1830's with the arrival of inhabitants coming first mostly from the nearby Charlevoix region, and afterwards from other regions of the Saint Lawrence valley[68]. From 1838 to 1911, almost 30 000 individuals migrated to the SLSJ, 70% of them from Charlevoix [90],[91]. Thus, SLSJ provides a great example of a founder population [74].

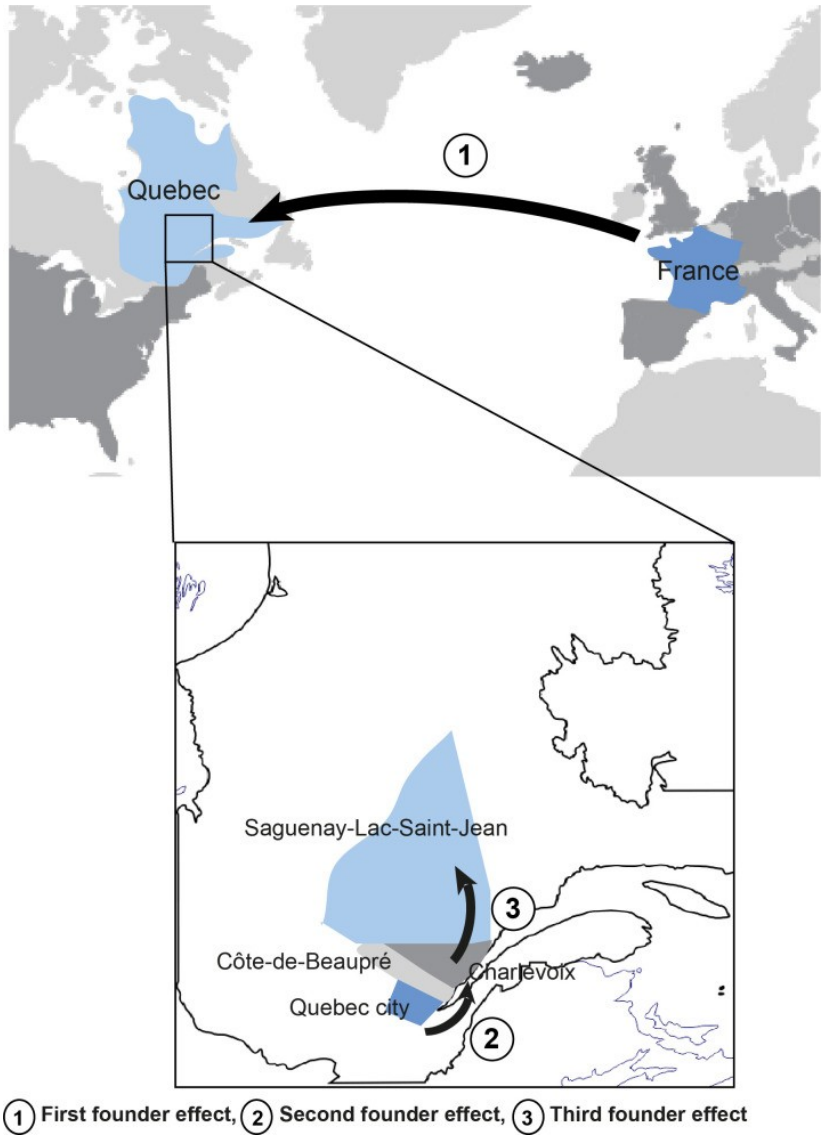

Figure 5. Three main migratory events contributing to the founder effect in Saguenay–Lac-Saint-Jean (SLSJ) region.

During the 17th and 18th centuries, between 10 000 and 12 000 immigrants, mainly from France, settled in the Saint Lawrence Valley (first founder effect). From the end of the 17th century, inhabitants of the Saint-Lawrence Valley, more particularly from Quebec City and the Côte-de-Beaupré area, settled in the Charlevoix region (second founder effect). Finally, settlers from Charlevoix moved to the SLSJ region from the 1830s (third founder effect). They were later followed by settlers from other Quebec regions, but they represent most of the founders of the SLSJ population (Re-use permitted under CC BY-NC liscese

2 Objectives

Autosomal recessive spastic ataxia of Charlevoix–Saguenay (ARSACS) is one of the most prominent examples of a rare hereditary disease that has become frequent in a geographically restricted region due to a strong founder effect. The Saguenay–Lac-Saint-Jean (SLSJ) region of Quebec, Canada, has a unique population history shaped by the settlement of a limited number of French immigrants in the 17th and 18th centuries, followed by relative geographic and cultural isolation. This historical context created a bottleneck and allowed particular genetic variants to increase in frequency.

We hypothesize that the primary ARSACS variant, c.8844delT in the SACS gene, was introduced into the SLSJ population by one or a few individuals who immigrated from France to New France (French Canada) in the 17th century, and that this variant persisted and expanded as the descendants of these early settlers formed much of the contemporary SLSJ population. The presence of a few additional, less frequent pathogenic variants (e.g., p. Cys814Tyr and p. Gln7504Ter) raises the possibility that more than one founder lineage may have contributed to ARSACS in the region.

The overarching goal of this project is therefore to trace the historical and genetic origins of ARSACS in the SLSJ population by integrating large-scale genealogical records with modern genomic data. We aim to identify not only the contemporary distribution of disease-associated haplotypes but also the intermediate ancestors and ultimately the original founder(s) responsible for introducing the pathogenic variant(s) into the population.

3 Study population

Participants were recruited from a large ARSACS natural history study ongoing at the Groupe de recherche interdisciplinaire sur les maladies neuromusculaires (GRIMN) of the Centre Intégré Universitaire de Santé et de Services Sociaux (CIUSSS) of the SLSJ. Inclusion criteria were: (1) ≥ 18 years-old, (2) genetically confirmed diagnosis of ARSACS homozygous for the SACS c.8844delT mutation, (3) able to provide informed consent and (4) blood sample already collected or available for rapid collection. This project builds on a longitudinal study that began in 2006, through which cohorts of various neuromuscular disorders—such as myotonic dystrophy type 1 (DM1), ARSACS, and oculopharyngeal muscular dystrophy (OPMD)—have been established. Over the years, these cohorts have supported research on natural history, patient registries, biomarkers, clinical trials, and the development of care guidelines.[92]. Each participant was assigned a unique identification code to facilitate data management and to protect confidentiality by avoiding the use of patient names. Genealogical data were available for 166 patients, with each individual's ID linked to a corresponding genealogy identifier in the BALSAC database[93]. In addition, sequencing data were accessible for 71 patients through the Centre d'expertise et de services Génome Québec. Overall, both genealogical and sequencing data were available for 67 patients

4 Chapter 1 (Genealogical analysis)

4.1 Genealogical data

Genealogical data are a mirror of regional demography; they provide information on phenomena that have a direct influence on the genetic structure of populations [62],[65] , [67] . The study of genealogies allows us to demonstrate the choices of reproductive partners and to inform us about fertility and migration patterns. However, this type of data is generally rare or covers only a few generations[94]. The population of Quebec is one of the rare populations that has a set of genealogical data covering its entire territory over its entire post-colonial history. This data is available through the BALSAC population file, which has been under development since 1971 at the Université du Québec à Chicoutimi [93] . The file was constructed from Quebec civil status records, including birth, marriage and death certificates. In all these documents, there are dates, names and places that allow the researcher to make connections and connect individuals together. These records were digitized and then paired to reconstruct family lines from the beginning of the 17th century to the second half of the 20th century[95]. The available data ranges from the beginning of the European settlement in the early 1600s to 1965. Unlike genetics where it is only possible to make hypotheses about the past from contemporary data, genealogical data allow to have a real portrait of the evolution of a population

4.2 Methods

4.2.1 Completeness

Genealogies are used to calculate various coefficients. We performed a quality control based on the completeness of the genealogies. Completeness refers to the proportion of ancestors who are present in a given generation compared to the maximum number of ancestors

expected, the expected number of ancestors being equal to 2^x where x corresponds to the number of generations[96]. This formula arises from the fact that each individual, or ancestor, has two different parents. Thus, the total number of ancestors increases exponentially. Low completeness is explained by a lack of available data on ancestors and can bias genealogical analyses. Therefore, to ensure the possibility of tracing the founder of the ARSACS variant in the Saguenay population, the first step was to calculate the completeness of the genealogical data of people who were married in the SLSJ between 1935-1960 in order to confirm that sufficient information was available. The R library GENLIB was used to calculate several coefficients from genealogical data such as Completeness[96].

4.2.2 Kinship coefficient

Due to the recessive inheritance pattern of the disease and the involvement of both parents in its transmission to their children, we focused our analyses primarily on parents. Among the 166 patients for whom genealogy data were available, we identified 207 corresponding parents. Using the GENLIB library[97] in R, we calculated the kinship coefficients between pairs of parents to investigate the structure of ARSACS families and their genetic relatedness. In genetic analysis, the kinship coefficient (often denoted as Φ) is a measure of the genetic relationship between two individuals, defined as the probability that a randomly selected allele from one individual is identical by descent (IBD) to a randomly selected allele from another individual at the same locus. In simpler terms, it quantifies the chance of inheriting the same gene copy from a shared ancestor[98]. The GENLIB R package calculatess kinship coefficients to different familial relationships, with closer relatives showing higher values. For example, siblings typically have a coefficient of 0.25, while cousins have about 0.0625.

4.2.3 MRCAs and Founders

Given the recessive inheritance pattern of the disease, which requires the involvement of both parents, we first calculated the number of parents associated with these patients. In total, 207 parents were identified for the 166 patients because many of the patients were siblings.

We identified the most recent common ancestors (MRCAs) of the 207 parents at different genealogical distances, measured in meioses (for example, two cousins are separated by four meioses).

Using GENLIB[97], we then calculated the minimal genealogical distance, defined as the shortest path through which two individuals are related via an MRCA. A pair of parents may have more than one MRCA, as long as none of the ancestors in the MRCA set also appears as a descendant who is an ancestor of the pair. Using the same package in R, we also identified the founders of the parents within the genealogy data. Founders are defined as the individuals without parents in the genealogy. Importantly, these founders, derived exclusively from genealogy, represent the earliest carriers of the whole genetic material observed in ARSACS patients not only the carrier of the causal allele.

4.2.4 Genetic Contribution

To determine which individuals had the greatest influence on transmitting genetic material to their descendants, we calculated genetic contribution. Genetic contribution is the proportion of genetic material that an ancestor (or founder) is expected to transmit to a given individual or population, based on their position in the genealogy. It is calculated from the probability that alleles from a particular ancestor are inherited by a descendant, and it reflects the ancestor's

relative influence on the genetic makeup of the studied individuals [96]. The more often an ancestor appears in the genealogy of probands from the same region, the greater the probability that its alleles have been transmitted to the current population. The ancestors of a population do not contribute to its gene pool in an equitable manner. Certain factors may cause ancestors to have a much higher occurrence and therefore a greater genetic contribution, thereby reducing diversity in the contemporary population [60],[49]. In particular, specific ancestors who arrived early in the regional settlement process may have left a large number of descendants and contributed greatly to the population genetics[49],[99]. This implies that even when new arrivals come later to diversify the gene pool of a population, their genetic contribution will remain lower even if they produce a large number of descendants.

In this study, **GENLIB** package in R is used to calculate the genetic contribution of founders by summing contributions across all genealogical paths from the ancestor to the parents' gene pool [97].

4.3 Results

4.3.1 How much our genealogy data is complete?

The BALSAC database provides genealogical information for the populations of Quebec and the Saguenay–Lac-Saint-Jean (SLSJ) region. Using this resource, we examined the completeness of the records related to ARSACS patients, their families, and all their ancestors and founders. Our analysis showed an overall completeness of 80% up to the ninth generation, representing one of the highest levels of genealogical coverage among projects in

our laboratory. This degree of completeness provides a sufficiently reliable basis for tracing the founder of the main ARSACS variant in earlier generations through further genotyping analyses and pedigree alignment methods. Figure 6 illustrates the rate of genealogical data completeness among ARSACS patients and families originating from SLSJ.

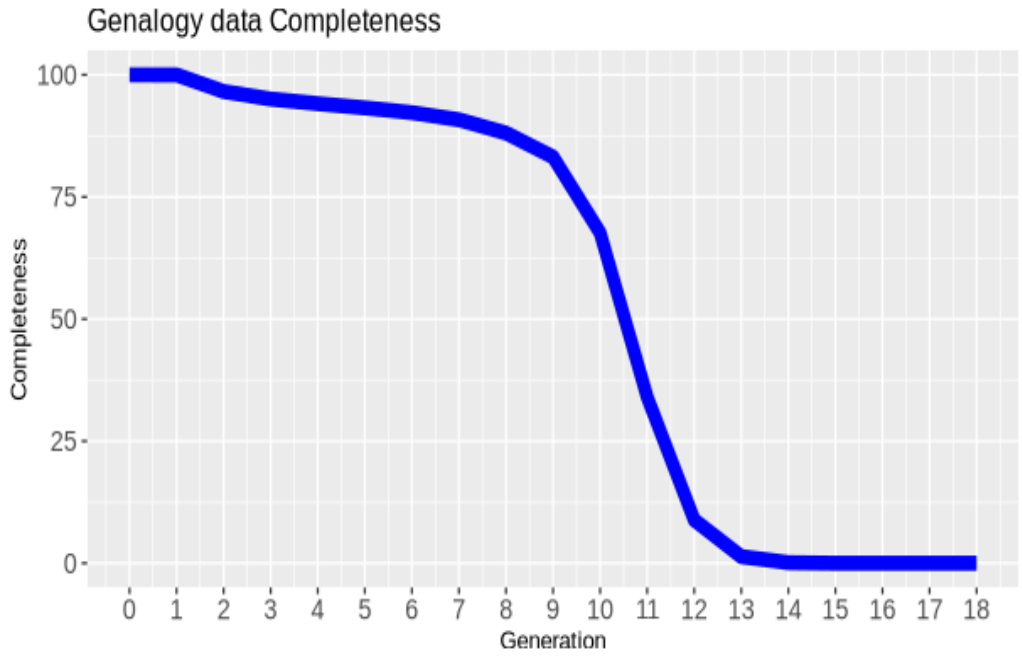

Figure 6. Genealogy data completeness of ARSACS patients and families up to 18 generations back.
We saw an overall completeness of 80% up to the ninth generation.

We also calculated the completeness for each patient for whom both genealogical and sequencing data were available, since in the final analysis these are the samples considered for aligning genotyping information with genealogical data to identify the founder of the main ARSACS variant. We calculated the 67 specific patients' genealogy data completeness and we found that one patient had only 50% completeness up to the last generation, as

genealogical data from the paternal side were missing. Consequently, this patient was excluded from further alignment analyses, where the haplotype of each individual and the identification of carrier ancestors played a critical role.

4.3.2 Identify ARSACS families

The goal of this project is to understand how the main ARSACS variant (c.8844delT) has been distributed within the population. To achieve this, we examined the genealogical distribution of the 166 patients for whom genealogical data were available across SLSJ families. Given the recessive inheritance pattern of the disease, which requires the involvement of both parents, we first calculated the number of parents associated with these patients.

To characterize the familial structure of ARSACS families and to understand how the disease has been distributed in the contemporary generation, we computed the pairwise kinship coefficients of these parents. This produced a 207 × 207 matrix, where each cell represents the kinship coefficient between a pair of parents. Based on this matrix, we constructed a dendro-heatmap (Figure 7). In the heatmap, each tile indicates the kinship coefficient between parents positioned on the x- and y-axes, while the diagonal represents each parent compared to themselves, yielding the highest kinship value (0.5). Familial relationship ranges are displayed in different colors. To the left of the heatmap, a dendrogram illustrates the relative closeness of individuals, where shorter branch lengths indicate closer relationships.

On the diagonal of the heatmap, clusters of tiles formed perfect squares, each corresponding to a close family group. In total, we identified 56 such squares, representing 56 unique families; however, 50 individuals could not be assigned to any family group. To validate these findings, we reconstructed the pedigree of each family and examined the number of patients included within them. As shown, each family unit includes a maximum of five patients. A pedigree example of the encircled square in Figure 7 is represented in Figure 8. Finally, we assigned unique family ID to each parent genealogical ID. This step was essential for subsequent analyses, where genotyping data will be integrated to reconstruct the genealogical structure of ARSACS families and to better understand the distribution pattern of the c.8844delT variant in the contemporary population. For confidentiality reasons, the IDs of the individuals have been masked.

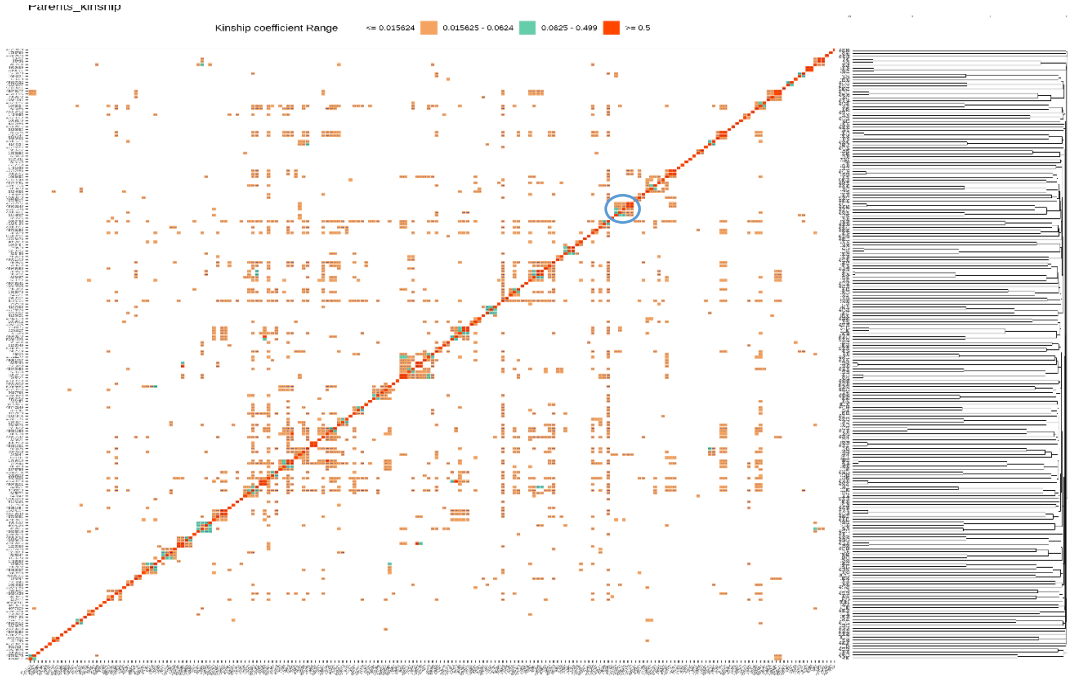

Figure 7. Dendro-heatmap representing kinship coefficient of carriers and familial relationship.

Based on the kinship coefficients of carrier parents, we redefined close family relationships. A heatmap was constructed to illustrate carriers' kinship coefficients exceeding 0.015625 (the coefficient for second cousins) across different ranges. Carrier IDs are displayed on both the x- and y-axes in a matrix format. The dendrogram on the right indicates the relative closeness of individuals, where shorter branch lengths represent closer relationships. In total, this analysis reconstructed 56 family units, one of which is highlighted and whose corresponding pedigree is presented in Figure 8.

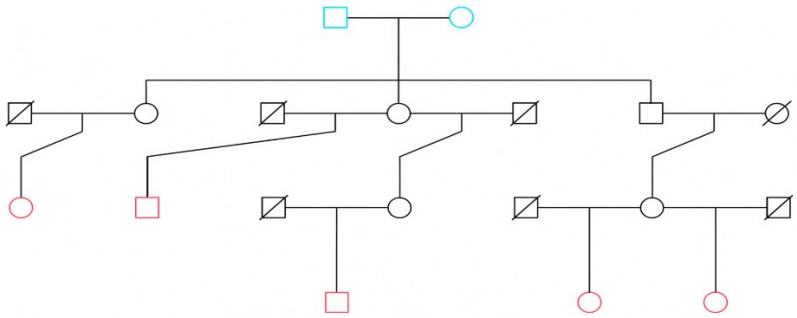

Figure 8. An example of ARSACS family pedigree.
The encircled family includes five probands (individuals in the first generation of the pedigree who do not have children), highlighted in red. All five can be in the corresponding square of the heatmap, each represented by different colors. As described in the previous figure, individuals shown in dark orange tile represent siblings, which are displayed as siblings on the right side of the pedigree. Those in blue tile correspond to cousins, illustrated on the left side of the pedigree. The remaining individual in the middle is a second cousin of the others, previously indicated by a light orange tile. The most recent common ancestor (MRCA) of this family has been identified and is shown in blue at the top of the pedigree. The crossed individuals represent those who have passed away.

4.3.3 Find the contribution of ARSACS founders based on genealogy data

Using the genealogical data and the GENLIB R library, we traced the founders of these patients to estimate the number of potential founders contributing to the genetic material of ARSACS patients. Overall, 62 founders were identified. We then calculated the genetic contribution of these founders to the carriers' gene pool to determine which individuals had the greatest influence on transmitting genetic material to their descendants. Four individuals were found to have the highest contributions; however, further analyses incorporating genotyping data and haplotype information are required to confirm which of these 62 founders introduced the main ARSACS variant into the population. Figure 9 illustrates the

contribution of each founder to the carriers' gene pool. For confidentiality reasons, the IDs of the individuals have been masked. For confidentiality reasons the IDs have not shown in the figure.

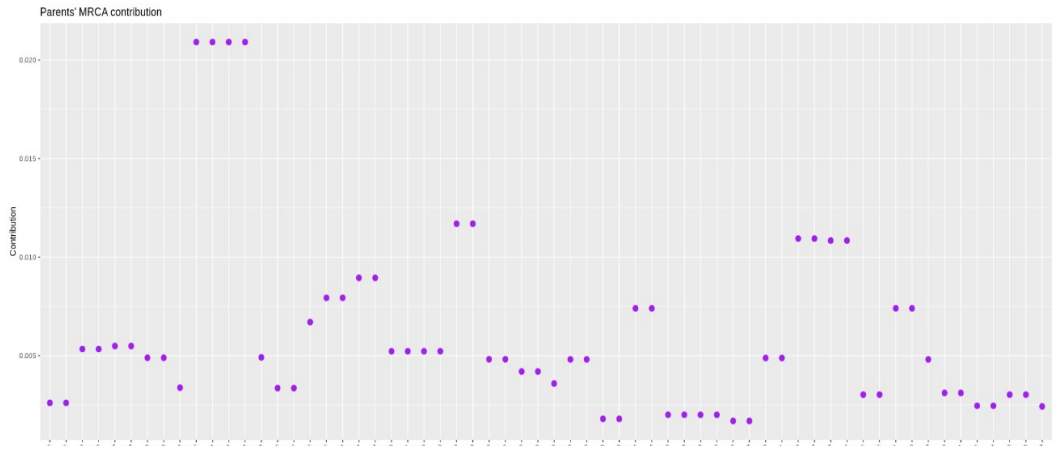

Figure 9. Contribution of founders to the carrier's gene pool.
Each dot on the y-axis represents the genetic contribution of a founder to the carriers' gene pool.

5 Chapter 2 (Genotyping Analysis)

5.1 Genetic data

The whole-genome genotyping data were generated for a total of 71 patients through the Centre d'expertise et de services Génome Québec. Genotyping refers to the process of identifying genetic variations across an individual's genome by examining specific positions, or markers, within the DNA sequence. This approach does not read the entire DNA sequence; instead, it detects known variants, usually single nucleotide polymorphisms (SNPs), that can provide valuable information about genetic diversity, disease risk, and population structure. The genotyping process typically uses microarray-based technologies, which allow researchers to analyze hundreds of thousands of markers simultaneously and efficiently at a relatively low cost.

The genetic data of the control group for this study come from CARTaGENE (CaG) cohort [100].CaG is a public research platform that consists of biological samples, phenotypic data and information on the lifestyle habits of individuals from Quebec. The cohort includes genetic data for 29,337 individuals who were recruited in six major centers in the province, regardless of their place of birth. Individuals were recruited in Gatineau, Montreal, Quebec City, Trois-Rivières, Sherbrooke and Saguenay [101] . Since the participants were not recruited on the basis of specific medical conditions, they constitute a random sample and as representative as possible of the population. CARTaGENE provided a reliable source of controls, allowing us to ensure that the observed genetic patterns in patients were not

influenced by hidden population stratification or the inclusion of individuals from outside SLSJ. Moreover, the comparison with CARTaGEN

controls coming from the SLSJ offered a broader perspective on the genetic structure of the population. Standard quality control procedure was done to ensure that our data was clean.

First, SNPs missing in at least 5% of individuals were eliminated, then individuals for which at least 5% of SNPs were missing were removed [102] . Next, SNPs that deviated from Hardy-Weinberg equilibrium with a p-value of 10^{-6} were eliminated [102] . Hardy-Weinberg equilibrium is a principle that describes the genetic variation expected in a population in the absence of evolutionary forces[103] . Failure to meet this criterion is indicative of genotyping errors. The cleaned datasets were then combined based on common genotyped positions. The cleaned dataset includes 3076 controls from CARTaGENE cohort and 71 ARSACS patients genotyped.

5.2 Methods

5.2.1 Visualization of population structure

Since genetic data files are complex and large, dimensionality reduction methods must be used to visualize similarities and differences between individuals. We first performed a principal component analysis (PCA) on the genotyping data of the 3076 individuals with the plink/1.9b_5.2-x86_64 software [104], then a uniform manifold approximation and projection (UMAP) with the R umap library [105]on the first two resulting components. To address this, we will compare the genetic structure of patients with that of CARTaGENE controls who will be confirmed to be from SLSJ.

5.2.1.1 Principal component analysis

In population genetics, PCA is a popular method to infer the ancestry of individuals or to control

for population structure in genome-wide association analyses [106]. It is an analytical method that aims to reduce the complexity of a data set while preserving important features as much as possible [106] . PCA transforms the initial data into a new set of principal components. These components are independent of each other and aim to preserve the majority of the variance [107] .Components explaining the greatest proportion of the variance can then be chosen to position the samples in a two-dimensional space to test for specific trends in the data [113]. This technique has the advantage of simplifying the interpretation of the data while allowing the visualization of the population structure within the sample [106],[108],[109] . Indeed, the coordinates of two individuals presenting a large genetic similarity will position them closer to each other than individuals presenting strong differences.

By applying PCA to genotyping data, it is possible to capture the population structure by separating individuals according to their ancestry. The first two components, which explain most of the variance in the data, allow to distinguish genetically very different groups, such as those from different continental origins [110] . Since no clear structure was observed in the initial PCA plot, we subsequently applied an additional dimensionality-reduction method, such as UMAP.

5.2.1.2 Uniform approximation and projection of manifolds

UMAP is a nonlinear dimensionality reduction algorithm that preserves the structure of the data [111] . The algorithm first constructs a graph that represents the relationships between points in the original data, and then projects them into a low-dimensional space. The points are moved in space until the similarities between the points in the low-dimensional space match the relationships that are observed in the high-dimensional data set. The points are moved gradually and randomly until the positioning reflects the structure of the original data [112] . The interpretation of UMAP results is different from that of PCA. In UMAP, the distance between points does not necessarily represent the structure of the original data. The points are distributed uniformly and the space between the points is artificially distorted [112]. Visualization of UMAPs allows to detect clusters formed by similar individuals, but the distance between different clusters is not informative [110].

5.2.2 Identity by descent segments analysis

IBD segments represent the common inheritance of an identical part of the genome between two individuals and provide information on the demographic history of populations [85],[86], [113]. In particular, they allow us to infer family ties, quantify the intensity of relatedness in a population and identify mutation transmission patterns[88].

IBD segments were inferred on the phased genotypes with refined IBD version 17Jan20 [114] by Beagle version 18May20 [115] . A VCF (Variant Call Format) file is a standard format used to record variation data from sequencing or genotyping. A phased VCF adds the extra information of which chromosome (maternal or paternal) each allele resides on, for example “0|1” instead of “0/1”. This phasing is critical for reconstructing haplotypes: when alleles are

phased, you can trace which variants are inherited together on the same parental chromosome. The IBD segments were detected using a sliding window of 40 SNPs, a log of odds threshold of 3, and a minimal length of 2 cM with a sensitivity scale of 10 to retain segments in familial data. Segments were then merged using the merge-ibd-segments 17Jan20.102 tool[116] .

We investigated how carriers of the disease transmitted identity-by-descent (IBD) segments. Specifically, we compared the rate of IBD sharing between patients, carriers, and controls to determine in which group the sharing was most elevated at each genomic position. By doing so, we sought to uncover the genetic signature of the disease variant and to evaluate whether the observed IBD sharing patterns were consistent with a strong founder effect. This analysis helped us to provide independent confirmation of c.8844delT origin as a founder mutation in the SLSJ population.

5.2.3 Haplotypes Analysis

5.2.3.1 Haplotypes

Haplotypes are defined as groups of alleles or genetic variants that are inherited together on the same chromosome. Because recombination events are relatively rare within short genomic intervals, variants in close physical proximity are often transmitted as a block, creating stable haplotype structures over multiple generations. Haplotypes can therefore serve as valuable markers for studying the inheritance of disease-associated variants and for tracing founder events in populations. Longer haplotypes sharing generally indicate more recent shared ancestry, whereas shorter haplotypes suggest older origins. This property makes haplotype

analysis particularly useful in understanding the history and transmission dynamics of rare genetic disorders in isolated populations[117].

Thus, in ARSACS study, defining haplotypes around ARSACS main variant allows us to identify shared ancestral segments and potential multiple founders (if divergent haplotypes exist), and also to estimate relative timing of variant entry based on haplotype breakdown by recombination.

We used the geneHap R package[118] to define haplotypes across the genomic regions of interest. This tool allows the construction and reconstruction of haplotypes from phased genotype data, enabling the identification of shared haplotypes among individuals. By applying different haplotype lengths, we were able to capture both short- and long patterns of genetic inheritance.

5.2.3.2 Linkage Disequilibrium blocks

Linkage disequilibrium is the nonrandom association of alleles at different loci and a sensitive indicator of the population genetic forces that structure a genome. Because of the explosive growth of methods for assessing genetic variation at a fine scale, evolutionary biologists and human geneticists are increasingly exploiting linkage disequilibrium in order to understand past evolutionary and demographic events, to map genes that are associated with quantitative characters and inherited diseases, and to understand the joint evolution of linked sets of genes[119].

A haplotype block is a chromosome region in which there are few haplotypes, for which the LD analysis provides evidence of a low rate of recombination. The discovery of haplotype blocks showed that LD usually extended over much longer chromosomal distances and suggested that testing one SNP within each block for significant association with a disease might be sufficient to indicate association with every SNP in that block, thus reducing the number of SNPs that need to be tested in case–control studies of disease association[120].

Using Haploview software[121],we identified the start and end coordinates of each linkage disequilibrium (LD) block within the haplotype regions of interest. Haploview computes pairwise LD between SNPs and groups them into discrete LD blocks based on statistical thresholds (e.g., D′ and r^2). To improve the visual interpretation of haplotype architecture, we assigned distinct colors to each LD block and displayed them adjacent to one another alongside a haplotype block bar. This visualization facilitated a clearer comparison of the internal structure of haplotypes across different genomic regions, enabling us to distinguish conserved ancestral segments from regions with higher recombination.

5.2.4 Phylogeny tree of haplotypes

Haplotype frequencies within a population tend to vary substantially. However, haplotypes often differ only due to few mutations, and leveraging similarities can improve the estimation of effects. The phylogenetic relationships can be either in a form of a tree or a network, and we refer to the model as the haplotype network model, which models haplotype effects by leveraging phylogenetic relationships described with a directed acyclic graph[122]. An evolutionary tree of haplotypes can be reconstructed through the analysis of recombination patterns. Haplotype trees can be used to reconstruct past human gene-flow patterns and

historical events, but any single tree captures only a small portion of evolutionary history, and is subject to error. A fuller view of human evolution requires multiple DNA regions, and errors can be minimized by cross-validating inferences across loci[123]. There are several methods to define the phylogeny tree of haplotypes. In this sturdy we used RAxML method[124] .RAxML (Randomized Axelerated Maximum Likelihood) is a widely used software tool for inferring phylogenetic trees based on the maximum likelihood (ML) principle. It is designed to handle large datasets with thousands of taxa and characters efficiently. RAxML implements rapid hill-climbing search algorithms to explore tree topologies and uses the general time-reversible (GTR) substitution model combined with models for rate heterogeneity across sites, such as the Gamma distribution or the CAT approximation. These approaches allow it to balance computational speed with accuracy when estimating evolutionary relationships. Additionally, RAxML supports both nucleotide and amino acid sequence data and includes bootstrap resampling to assess tree reliability [125],[126].

A key feature of RAxML is its optimization for parallel computing, enabling analyses on multicore CPUs and high-performance computing clusters. The program employs heuristics to reduce the search space and improve speed while maintaining high-quality likelihood scores. The introduction of RAxML-NG (Next Generation) further enhanced computational performance and scalability for very large phylogenomic datasets[127]. Because of its efficiency, flexibility, and robust implementation of statistical models, RAxML remains one of the standard tools in phylogenetics for reconstructing evolutionary histories based on genomic and molecular data.

5.2.5 Pedigree alignment method

To identify the founder of the disease-associated haplotype, we used a pedigree-alignment approach that combines genetic information from haplotypes with genealogical data from an extended multigenerational pedigree. This approach builds on the principle that haplotypes shared identical-by-descent (IBD) among affected individuals or carriers must originate from one or more common ancestors in the pedigree[128],[129].We first reconstructed a phylogenetic tree of disease-linked haplotypes using phased genotyping data from sampled individuals. This tree allowed us to visualize the relationships among haplotypes and to estimate the minimum number of ancestral copies and mutational events that explain the current diversity of haplotypes[130],[131].The phylogenetic tree provided the genetic framework for aligning observed haplotypes with their most likely ancestral nodes.

We then aligned the phylogenetic tree with the large pedigree assembled from genealogical records. Each sampled individual's haplotype was mapped to their known position in the pedigree, and we traced shared haplotypes upward through ancestral lines to identify the smallest set of common ancestors capable of transmitting the disease-associated haplotype to all current carriers [132],[133]. By examining how haplotypes co-segregated within family branches, we assigned ancestral origins to internal nodes of the phylogeny using a dedicated pedigree-alignment method[134].This alignment step ensured that the branching pattern of the haplotype tree was consistent with Mendelian inheritance and with the autosomal-recessive transmission pattern characteristic of ARSACS-associated haplotypes across generations.

Finally, we refined the founder inference by comparing the founder candidates identified at the root of the phylogenetic–pedigree alignment with those inferred solely from genealogical data using the GENLIB R package[97]. This comparison allowed us to integrate two complementary sources of evidence: the genetic signal from haplotype–phylogeny tree and the genealogical tracing of ancestral lines. By cross-referencing these two founder candidate sets, we identified a consensus group of ancestral individuals most likely to have introduced the ARSACS founder mutation (c.8844delT) into the Saguenay population. This integrative step strengthened the inference by ensuring that the final founder assignment was supported both by genetic inheritance patterns and by historical genealogical records.

5.3 Results

5.3.1 Visualization of population structure

To confirm that the 67 patient samples with available genotyping data originated from the Saguenay–Lac-Saint-Jean (SLSJ) region, we applied two complementary dimensional reduction methods to the large-scale SNP genotype data from both the patient group and the CARTaGENE control cohort. The control group consisted of individuals previously verified to have ancestry from the SLSJ region. By projecting both patients and controls into the same reduced-dimensional space, we were able to assess whether the patients clustered with the established SLSJ population, thus supporting their regional origin.

To increase the robustness and reliability of this verification, we used two approaches: PCA and UMAP. PCA was first applied to the genome-wide SNP dataset to summarize the major axes of genetic variation.

Next, we projected the individuals into a two-dimensional UMAP space using the first two principal components as input. In the resulting UMAP plot, figure 10, the orange dots (patients) clustered closely with the blue dots (controls from the SLSJ region), indicating that the genetic profiles of the patients align with those of the local reference population.

This concordance provided strong evidence that all of our samples indeed originate from the SLSJ region. Establishing this population consistency was an essential first step before subsequent analyses aimed at identifying founder haplotypes underlying the ARSACS mutation.

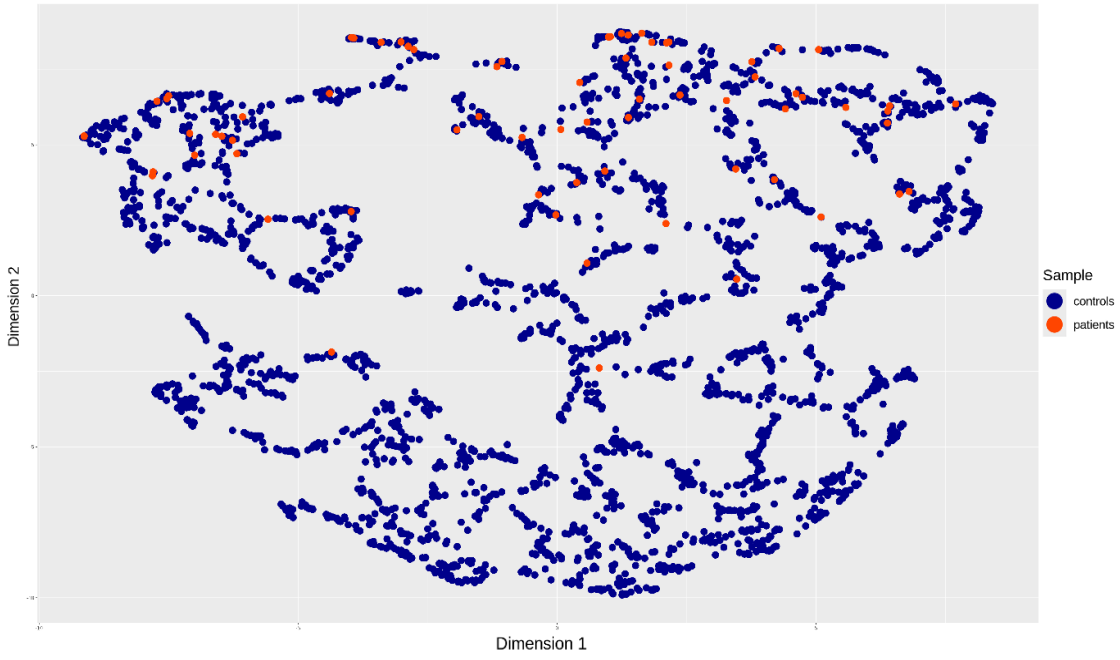

Figure 10. UMAP of the first 2 principal components of the genotype data of the patients and controls

The close clustering of the orange dots (patients) with the blue dots (SLSJ controls) indicates that the patients' genetic profiles closely match those of the local reference population.

5.3.2 Identity by descent segments analysis

Identity-by-descent (IBD) segments longer than 2 centimorgans (cM) were inferred from the phased genotypes using the Refined IBD algorithm (Browning & Browning, 2013), which is well-suited for detecting both recent and historical shared ancestry. We then quantified the rate of IBD sharing across genomic positions among three pairwise categories of individuals: patients–patients, patients–controls, and controls–controls.

As expected for a founder disease such as ARSACS, the analysis revealed a pronounced signal of excess of IBD sharing among patients in the region of the SACS gene on chromosome 13. This enrichment was notably higher than that observed in either the patients–controls or controls–controls comparisons. Such a pattern is characteristic of a founder effect, where a disease-causing mutation has been inherited from a limited number of common ancestors and propagated within a relatively isolated population over multiple generations.

Figure 11 illustrates the genome-wide distribution of IBD sharing across the three comparison groups. The distinct peak in the SACS locus highlights the strong founder signal and supports the hypothesis that the ARSACS mutation originated from a common ancestor in the Saguenay–Lac-Saint-Jean (SLSJ) population.

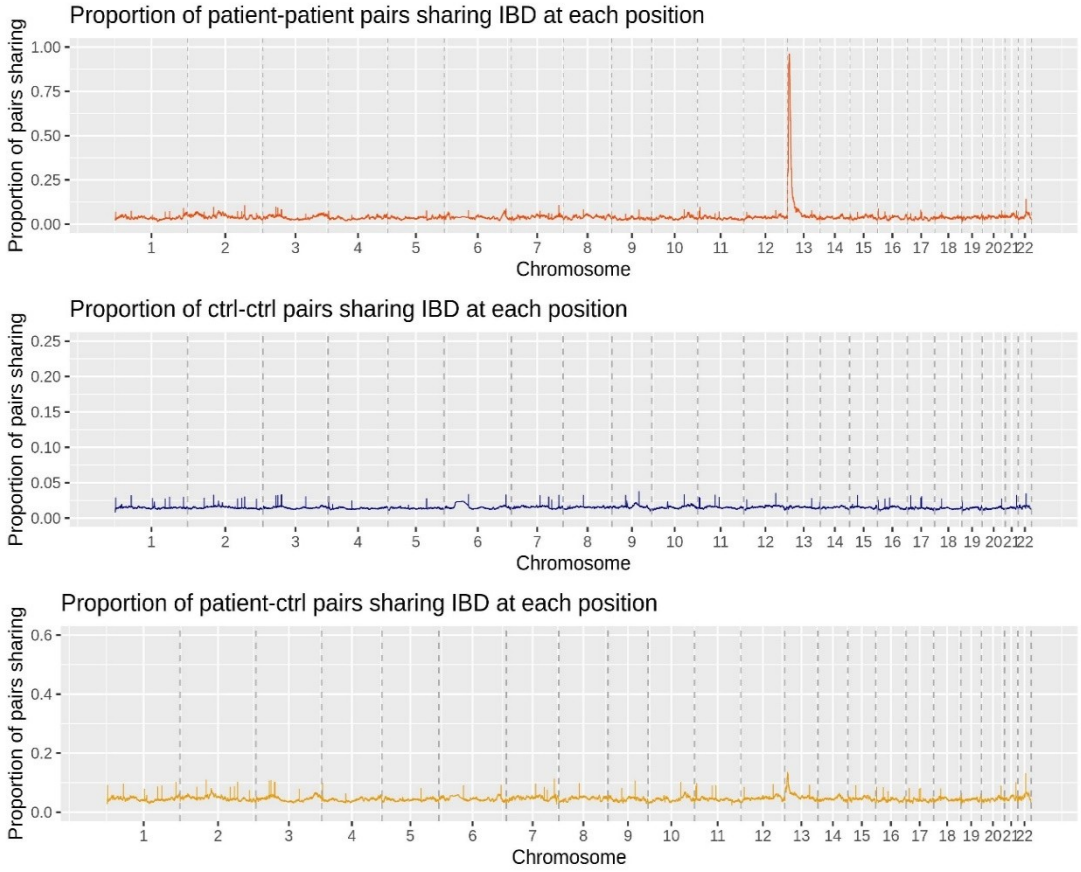

Figure 11. Genome-wide distribution of IBD sharing across three comparison groups: patients–patients, patients–controls, and controls–controls.
The first panel shows the IBD sharing among the patient's group, where nearly 94% of the patients carry the same IBD segment at the SACS gene locus on chromosome 13. In contrast, the patients–controls and controls–controls groups do not display a comparable peak in this region, highlighting the founder effect associated with the ARSACS mutation.

5.3.3 Haplotypes Analysis

Based on the number and distribution of IBD segments shared among patients upstream and downstream of the c.8844delT variant on chromosome 11, we defined haplotypes of varying lengths (ranging from 0.8 Mb to 8 Mb). This approach allowed us to explore the genomic structure surrounding the causal variant at different scales and to determine how many

haploid chromosomes, out of a total of 142 haploid chromosomes from 71 patients, carried identical haplotypes.

To characterize the internal organization of these haplotypes, we used Haploview software to identify the linkage disequilibrium (LD) block structure, including the start and end positions of each block, and to assign unique colors to each LD block for improved visualization. Subsequently, we employed the geneHapR package (Liu et al., 2022) to reconstruct and display each haplotype segment, enabling a clearer interpretation of haplotype diversity among patients.

As illustrated in Figure 12, we identified the structure of 16 unique haploblocks that were shared among ARSACS patients. To streamline further analyses—such as tracking the inheritance and distribution of these blocks across pedigrees—we assigned distinct names to each haploblock segment.

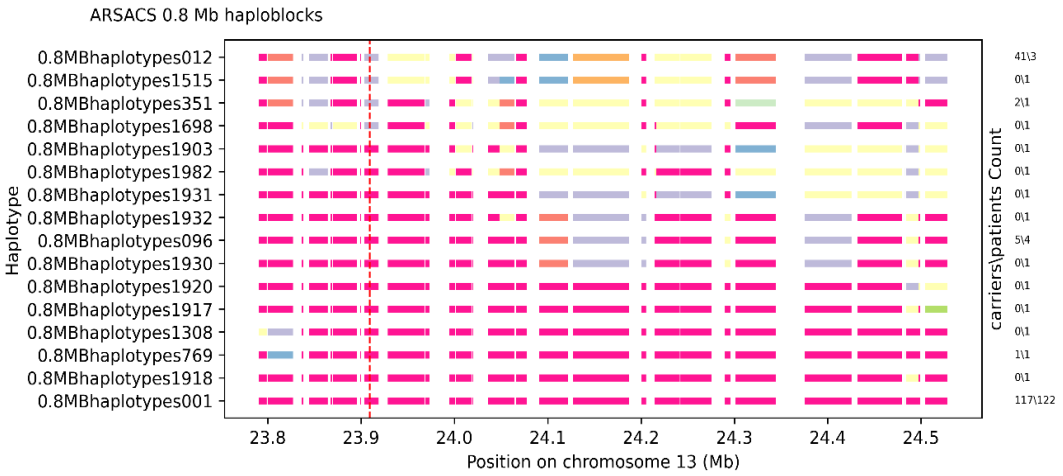

Figure 12 presents the visualization of the 16 unique 0.8 Mb haplotypes identified among ARSACS patients.

The x-axis indicates the approximate genomic positions of the LD (linkage disequilibrium) blocks that make up each haplotype (haploblock) structure. Each LD block is represented by a distinct color, allowing us to clearly differentiate the internal components of each haploblock. Notably, LD blocks that share the same color across different haploblocks indicate shared genomic segments, highlighting regions that are conserved among different haplotype units and may represent ancestral fragments. The y-axis lists the names assigned to each haploblock, which were defined during the haplotype reconstruction process to facilitate further analyses and tracking. On the right side of the figure, the distribution of individuals, patients versus carriers, sharing each haploblock segment is presented, providing insight into how these haplotypes are spread within the population. The exact genomic position of the ARSACS causal variant (c.8844delT), located at position 23,909,171 on chromosome 13 (based on the GRCh38 human reference genome), is marked by a red dashed vertical line. This annotation allows for a direct comparison of haplotype structures relative to the disease-causing variant and highlights the segments most closely associated with the founder mutation.

To investigate the broader patterns of recombination and gain a clearer understanding of the structural relationships among patient haplotypes, we extended our analysis—guided by the number of IBD segments shared among patients—to include longer haplotypes of 2.2, 4, and 8 Mb surrounding the *SACS* locus. The structures of these extended haplotypes, presented in the Supplementary Figures, offer a comparative view of how historical recombination events have shaped the genomic landscape in this region over generations.

Importantly, having access to the variant data for each patient enabled us to directly confirm which specific pathogenic mutation was present on each haploid chromosome. While the majority of patients carried two copies of the canonical c.8844delT frameshift mutation, we identified a subset of individuals who instead carried alternative mutations, including the 7504C>T nonsense mutation and the 814C>T missense mutation. These alternative haplotypes are represented as the first four haplotypes in Figure 12. Notably, the red dashed vertical line in Figure 12, marking the position of the canonical c.8844delT variant, passes through the pink-shaded region in most haplotypes but not in the first four, reflecting the presence of these alternative variants at that position.

Our analysis revealed that 122 out of 142 haploid chromosomes derived from the 71 genotyped patients carried the most common 0.8 Mb haplotype (haplotype 001), which harbors the founder mutation. In contrast, the remaining 20 haploid chromosomes, which either carried alternative mutations (as described above) or were associated with rare haplotypes shared by only a few individuals, were excluded from further analysis.

For subsequent analyses aimed at reconstructing the phylogenetic history of the haplotypes and identifying the most probable founder, we focused on a more genetically homogeneous subset of patients. Specifically, we selected 49 patients who were homozygous for the most common 0.8 Mb haplotype (haplotype 001) on both chromosomes. This step ensured that the downstream phylogeny reconstruction and founder inference were based on a consistent genetic background, thereby reducing noise from recombination events or alternative variant introductions.

5.3.4 Pedigree alignment

After identifying haplotypes of varying lengths and establishing the high frequency of the 0.8 Mb haplotype 001 among patients, we next sought to explore the evolutionary distribution and lineage of these haplotypes within the SLSJ population. To achieve this, we focused on constructing a phylogenetic tree based on the 2.2 Mb haplotypes using a carefully selected subset of 49 patients who were homozygous for the common 0.8 Mb haplotype 001. Focusing on this subset enabled us to analyze a genetically homogeneous group that was most likely to share a common ancestral origin for their disease-associated haplotypes, thereby reducing confounding heterogeneity.

A critical factor in reconstructing haplotype phylogenies is the impact of historical recombination events. Recombination disrupts ancestral haplotype structures over generations, which can obscure signals of common descent. Shorter haplotypes generally undergo fewer recombination events, thereby better preserving the original ancestral configuration. On the other hand, extremely short haplotypes may lack sufficient variation to resolve evolutionary relationships. Based on this balance, we selected the 2.2 Mb haplotype length, which was long enough to capture informative variation yet short enough to minimize the confounding effects of recombination. This choice allowed for a more reliable reconstruction of the evolutionary distribution of the founder haplotype in the contemporary SLSJ population.

Figure 13 presents the phylogenetic tree of the 2.2 Mb haplotypes, constructed using the RAxML maximum-likelihood method. To connect the genetic evidence with historical demographic records, we integrated the pedigree-alignment method, which enabled us to link

specific ancestral nodes of the phylogenetic tree with individuals identified in the BALSAC genealogical database. Using this combined approach, we identified eight candidate individuals at node 28 for whom genealogical records were available. These individuals are the most recent known carriers of the founder haplotype in the SLSJ region and therefore likely represent the key founders who introduced the ARSACS-associated haplotype into this population.

Importantly, genealogical data were not available for individuals ancestral upper than node 28 because these ancestors lived outside the Quebec province and were not documented in the BALSAC registry. To further confirm the central role of node 28 in the dissemination of the founder haplotype, we examined other branches of the phylogenetic tree—specifically nodes 40, 23, and 24—using the GenLib R package. These branches did not cluster within the main subtree rooted at node 28. However, genealogical tracing revealed that their ancestral lineages eventually converged in earlier generations, sharing the same set of eight candidate founders associated with node 28.

Taken together, these findings strongly suggest that the ARSACS-associated haplotype was introduced into the SLSJ population by eight ancestral individuals—organized as four couples—identified at node 28 of the phylogenetic tree. These individuals represent the most plausible sources of the original founder haplotype in this regional population, marking the point at which the pathogenic variant likely first entered the local gene pool.

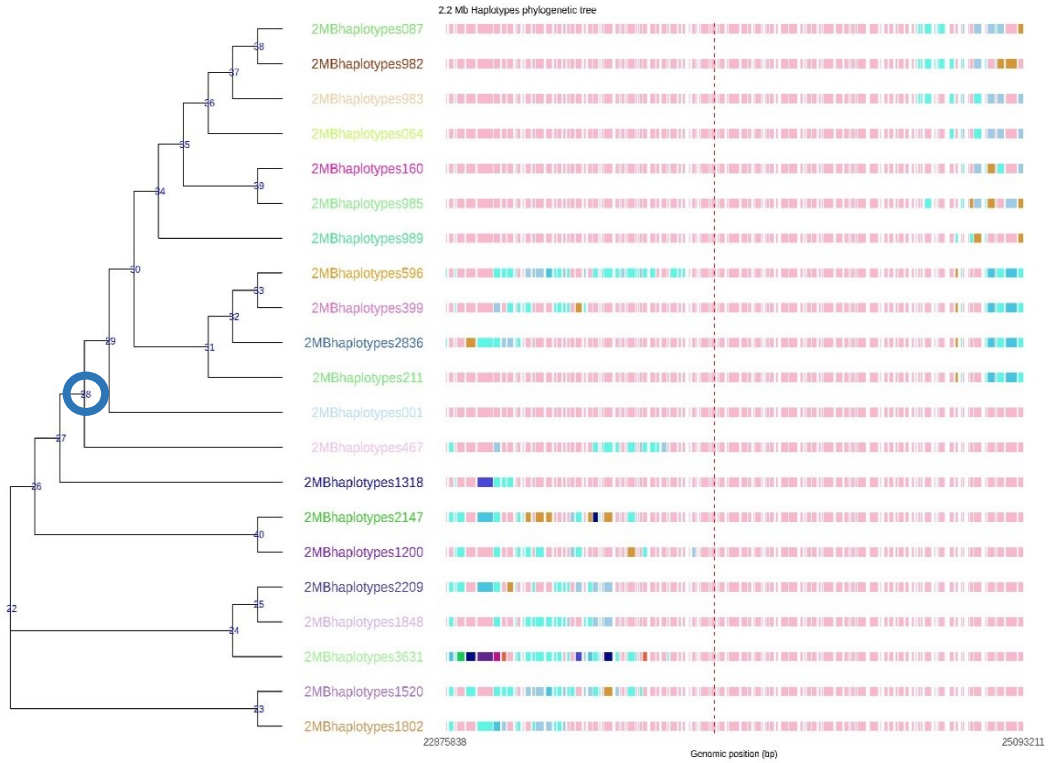

Figure 13 illustrates the phylogenetic tree of the 2.2 Mb haplotypes, along with the structure of the unique haploblocks.

On the left side of the figure, the phylogenetic tree displays the evolutionary relationships among the haplotypes. Using the pedigree-alignment method in combination with genealogical data, we were able to trace the lineages of these haplotypes back to node 28, which represents the most recent common ancestral branch accessible through available records. Within this node, we identified eight candidate individuals whose genealogical histories could be retrieved. These individuals are considered the likely founders of the first 2.2 Mb ARSACS-related haplotype, and they are presumed to have introduced the disease-causing ARSACS variant (c.8844delT) into the SLSJ population. In the middle section of the figure, the names assigned to each haplotype are shown. These haplotype identifiers were established during the reconstruction process to facilitate downstream comparative analyses and interpretation. On the right side of the figure, the linkage disequilibrium (LD) block structures of each haploblock are depicted. These structural representations reveal how segments of the haplotypes are organized. By comparing the LD block structures of each haplotype with those of their neighboring branches in the phylogenetic tree, we can observe patterns of recombination and structural shifts that have occurred across generations. This comparison provides insight into the evolutionary dynamics of haplotypes and highlights how recombination events have shaped the distribution and diversification of the ARSACS-associated haplotypes over time.

6 Chapter 3

6.1 Discussion

Populations that have experienced a founder effect have a gene pool that differs considerably from that of large, diverse populations. Their demographic history influences their genetic portrait, which is generally characterized by lower diversity[135]. This specificity offers a valuable perspective for identifying rare variants, as some can be observed at higher frequencies than what is observed globally[136]. Several rare genetic disorders have been shown to occur at unusually high frequencies in the Saguenay–Lac-Saint-Jean (SLSJ) region as a result of strong founder effects. Examples include myotonic dystrophy type 1 (DM1), Leigh syndrome, hereditary tyrosinemia type 1, and autosomal recessive spastic ataxia of Charlevoix–Saguenay (ARSACS)[137]. These conditions illustrate how the genetic contribution of a small number of early settlers, combined with subsequent population isolation and expansion, can lead to the enrichment of specific pathogenic variants.

6.1.1 Back to the objectives

In this study, we aimed to investigate the evolutionary history of ARSACS-related haplotypes in the SLSJ population by leveraging two unique resources: (1) the extensive genealogical records available through the BALSAC database, which document several centuries of regional family histories, and (2) the genotyping data from ARSACS patients collected in the neuromuscular clinic. By integrating these complementary datasets, we sought to reconstruct the origin and transmission of the ARSACS-associated haplotypes across generations. This approach provided an opportunity not only to trace the likely introduction of the primary pathogenic variant (c.8844delT) but also to better understand how recombination

and demographic history shaped the distribution of these haplotypes within the contemporary SLSJ population. The results presented here provide converging lines of evidence that support this hypothesis and offer new insights into the demographic and genetic processes that shaped the current distribution of ARSACS in the region.

6.1.2 Genealogy result interpretation

In first chapter, we investigated the genealogical structure of ARSACS in the Saguenay–Lac-Saint-Jean (SLSJ) population, aiming to understand how the disease has been distributed among affected families. Utilizing the BALSAC database, we were able to reconstruct family units, identify founders, and estimate their contributions to the carriers' gene pool.

Our genealogical data completeness allowed us to reliably trace familial relationships and reconstruct pedigrees. Kinship-based clustering revealed 56 unique family units, suggesting that the higher prevalence of ARSACS in the SLSJ region is likely due to a small number of founders who migrated from Charlevoix to Saguenay in the 1820s, rather than a high rate of consanguinity or cousin marriages.

The identification of 62 founders contributing to the gene pool of 67 patients illustrates the complex ancestral structure underlying in this population. Although four founders exhibited particularly high genetic contributions, further integration of haplotype and genotyping data was required to identify the individuals who actually introduced the main ARSACS variant, as this inference is based solely on genealogical data. Incorporating genotyping data could enable us to pinpoint the founder of the specific ARSACS-related allele, rather than the founder of the patients' entire genetic background. This highlights the utility of combining

genealogical reconstruction with genotyping data to distinguish between potential and true founders of a disease variant.

6.1.3 Genotyping results interpretation

A central aspect of our approach was the use of haplotypes of varying lengths to capture both recent and ancient recombination events. We first identified a common 0.8 Mb haplotype (haplotype 001) at high frequency among ARSACS patients, which served as a hallmark of the ancestral genetic background carrying the pathogenic variant. We then extended the analysis to longer haplotypes—2.2 Mb, 4 Mb, and 8 Mb—to better understand how recombination had fragmented the ancestral haplotype over time. This multi-scale approach revealed that the 2.2 Mb haplotypes offered an optimal balance between retaining informative ancestral segments and limiting the confounding effects of historical recombination. Such intermediate-length haplotypes preserved sufficient phylogenetic signal to reconstruct lineages, while still capturing variation informative about the divergence among patient lineages.

By focusing on 49 patients homozygous for the common 0.8 Mb haplotype 001, we constructed a robust phylogenetic tree based on their 2.2 Mb haplotypes. This strategy allowed us to concentrate on individuals most likely to share a common ancestral origin for the disease-associated haplotype. The phylogenetic reconstruction, performed using the RAxML maximum-likelihood method, revealed several well-defined clades and a principal lineage whose root was designated as node 28. Integration of genealogical information from the BALSAC database through the pedigree-alignment method provided a key bridge between genetic and demographic perspectives. We identified eight candidate individuals

corresponding to node 28 who were recorded in the regional genealogies and thus likely represent the most recent carriers of the founder haplotype within the SLSJ population.

We also examined the phylogenetic tree constructed using longer haplotypes from the same group of 49 patients who were homozygous for the common 0.8 Mb haplotype. However, increasing the haplotype length introduced a significant challenge: longer haplotypes are more fragmented by historical recombination events, which erode the continuity of the original ancestral segments. As a result, the phylogenetic tree of these longer haplotypes was more complex and broke into numerous smaller sub-trees rather than forming a single cohesive lineage.

Although we were able to identify a large number of intermediate or mediator ancestors for the roots of these sub-trees by integrating genealogical data, we were unable to trace all of them back to a single common root. This suggests that the original founder haplotype has been extensively reshaped over generations, obscuring the signal needed to pinpoint the ultimate founder who first introduced the ARSACS variant into the population. In other words, the longer haplotypes capture a more recent recombination history rather than the deeper shared ancestry that links all affected individuals to the original founder.

The identification of these eight candidates offers strong support for the historical founder hypothesis. Notably, genealogical data could not be traced for lineages ancestral to node 28 because these individuals originated outside the Quebec and were therefore not included in the BALSAC records. This finding suggests that the founder event predates the settlement of the region itself, consistent with the historical migration of early French settlers to Quebec in

the 17th century. Furthermore, the pedigree-alignment analysis of branches descending from nodes 40, 23, and 24 revealed that the ancestral lineages for these sub-branches also converge on the same eight individuals at node 28. This convergence strengthens the evidence that a single founder event—or at most a few closely related founders—introduced the pathogenic haplotype into the regional gene pool.

Our findings also highlight the value of large-scale genealogical resources such as BALSAC. By integrating these records with high-resolution genetic data, we could link contemporary patient haplotypes to historical individuals, thereby connecting molecular variation with demographic history. Such integrative approaches not only illuminate the origin of disease-associated variants but also provide a framework for similar studies in other founder populations.

From a clinical and public-health perspective, the identification of specific founder lineages has practical implications. By pinpointing the historical origin of the pathogenic haplotype and understanding its transmission dynamics, it becomes possible to refine genetic screening strategies in the region, anticipate recurrence risks in particular families, and design targeted carrier-testing programs. Furthermore, the methodological framework established here—particularly the integration of haplotype phylogenetics with pedigree alignment—can be adapted to investigate other rare disorders with suspected founder effects in the SLSJ region or in comparable populations worldwide.

6.2 Limitations

There are several important limitations to consider in this study. The most significant is the sample size. We had access to the genotyping data of only 71 ARSACS patients who had

been diagnosed and followed at the GRIMN neuromuscular clinic. Consequently, the founders we identified may not be fully representative of the entire ARSACS-affected population in the Saguenay–Lac-Saint-Jean (SLSJ) region. A larger sample would have increased the resolution of our analyses and potentially revealed additional founder lineages.

Another important limitation arises from the recessive inheritance pattern of ARSACS and the structure of our genotyping data. Although we phased the genotype data to facilitate the analysis of haplotypes directly inherited from each parent, we were still limited in our ability to determine parental origin. In other words, for each patient we could identify two distinct haplotypes inherited from their parents, but we could not confidently assign each haplotype to the maternal or paternal lineage. This lack of parental origin information introduced a degree of ambiguity when applying the pedigree-alignment method, which integrates genetic and genealogical records to trace haplotype lineages back to founders. For each haplotype represented in the phylogenetic tree derived from the genetic data, there were always two possible genealogical candidates—the father and the mother. This ambiguity complicated the tracing of haplotypes through historical generations, as each step in the analysis had to account for multiple inheritance pathways rather than a single, clearly defined lineage. While the pedigree-alignment method is capable of handling this complexity by evaluating all plausible transmission pathways, the process becomes substantially more intricate in the context of a recessive disease like ARSACS. If the disease had a dominant inheritance pattern, the situation would have been simpler: the pathogenic haplotype could be attributed to one parent only, eliminating the need to consider dual genealogical paths at every step. This would have streamlined the alignment process, reduced the uncertainty associated with

tracing lineages through intermediary ancestors, and potentially facilitated a more precise identification of the original founder(s).

Despite this limitation, our approach was still able to integrate genetic and genealogical data effectively, allowing us to infer plausible founder candidates. However, this challenge underscores the importance of developing future analytical frameworks or leveraging parent-of-origin information, such as from trio-based sequencing data, to refine lineage tracing in recessive disorders.

6.3 Perspective and Future Direction

While this study has provided valuable insights into the origin and transmission of the ARSACS founder haplotype in the SLSJ population, it also highlights several open questions and opportunities for deeper exploration. Future work will build upon the current framework by integrating advanced computational methods and additional layers of biological data to refine our understanding of the evolutionary history and clinical heterogeneity of ARSACS.

One of the next planned steps is to incorporate Ancestral Recombination Graph (ARG) [138]–based approaches to reconstruct the phylogenetic history of the *SACS* haplotypes. Unlike traditional phylogenetic methods such as RAxML, which infer a single bifurcating tree from sequence or haplotype data, ARG-based methods explicitly account for recombination events that have occurred over generations. This is particularly relevant for the SLSJ population, where the accumulation of recombination over centuries has fragmented the original founder haplotype, as seen in our analysis of longer haplotypes. By integrating recombination into the phylogenetic framework, ARG methods can generate more accurate

and realistic genealogical histories of chromosomal segments, allowing us to resolve complex evolutionary relationships that cannot be captured by standard tree-based models.

A key objective will be to cross-validate the founder lineages inferred from the ARG analysis with those previously identified using the RAxML-based approach. This comparison will allow us to evaluate the concordance and discrepancies between these methods and provide a more robust consensus on the historical founders who introduced the pathogenic *SACS* variant into the SLSJ population. The integration of ARG and RAxML findings will deepen our understanding of how the founder effect, recombination, and demographic history have collectively shaped the current distribution of ARSACS in the region.

Another crucial direction will be to expand our analyses beyond the genetic and genealogical data to include phenotypic and clinical information from ARSACS patients. Although all the patients in this study carry the same primary pathogenic variant (*c.8844delT*), there is substantial variability in the severity and nature of their clinical manifestations, including differences in age of onset, rate of progression, and the presence or absence of additional neurological features. This heterogeneity suggests the influence of genetic modifier factors—other genetic elements, possibly located in close proximity to the *SACS* locus, that may modulate the expression and impact of the primary pathogenic mutation.

By systematically correlating the genotypic data with clinical phenotypes, we aim to identify patterns that may point to such modifier genes. For example, variants in genes located near *SACS*, or in genes involved in related molecular pathways such as mitochondrial function or neuronal maintenance, could contribute to the observed diversity of clinical outcomes.

Integrating phenotypic data with haplotype analyses will also help disentangle the roles of environmental and genetic factors in shaping disease manifestation. This knowledge could have direct implications for precision medicine, potentially enabling risk stratification, targeted interventions, and better prognostic counseling for families affected by ARSACS.

Methodologically, the incorporation of ARG-based tools alongside traditional phylogenetic analyses represents an important advancement in the study of founder-effect disorders. It demonstrates how cutting-edge computational methods can be combined with historical genealogical resources and clinical data to create a more comprehensive picture of disease evolution. Moreover, the multi-layered approach proposed here—linking demographic history, recombination-aware haplotype evolution, and patient phenotypes—can serve as a model for similar studies in other founder populations and for other rare genetic diseases.

7 Conclusion

This study provides new insights into the demographic and genetic history of autosomal recessive spastic ataxia of Charlevoix–Saguenay (ARSACS) in the Saguenay–Lac-Saint-Jean (SLSJ) population. By integrating extensive genealogical records from the BALSAC database with high-resolution genotyping data from affected individuals, we were able to reconstruct the likely evolutionary trajectory of ARSACS-related haplotypes and identify key historical figures who contributed to the introduction and spread of the disease-associated variant.

Our results strongly support the historical founder hypothesis. We found that the majority of patients carry a shared ancestral 0.8 Mb haplotype (haplotype 001) encompassing the primary pathogenic variant c.8844delT in the SACS gene. Phylogenetic analysis of the 2.2 Mb haplotypes—selected to balance informativeness with minimal interference from historical recombination—revealed a major lineage converging on eight candidate individuals at node 28 in the genealogical records. These individuals appear to represent the most recent carriers of the founder haplotype within the SLSJ population. Importantly, the lineages ancestral to node 28 could not be traced in the regional genealogical data, consistent with the hypothesis that the disease-associated haplotype was introduced into the region by settlers of French origin prior to or during the early 17th-century migration into Saguenay.

The integration of genetic and genealogical evidence also revealed how recombination has reshaped the original founder haplotype over generations. While the 2.2 Mb haplotypes retained sufficient ancestral signal to identify a common lineage, analyses of longer haplotypes demonstrated a breakdown of this signal due to accumulated recombination events.

This highlights the importance of choosing an appropriate haplotype scale when reconstructing the evolutionary history of disease variants in founder populations.

Beyond the historical insights, our findings have practical implications for clinical genetics and public health. By pinpointing specific historical lineages associated with the disease, it becomes feasible to develop more efficient carrier screening and early diagnostic strategies tailored to the SLSJ region. This can improve risk assessment for families and inform genetic counseling, ultimately contributing to better disease management in affected communities.

Methodologically, this work demonstrates the power of combining haplotype-based phylogenetic approaches with genealogical data in founder populations. The use of the pedigree-alignment method provided a robust framework for bridging molecular data with demographic history, enabling the identification of plausible founders and clarifying the transmission dynamics of the pathogenic variant. Such integrative approaches can serve as a model for future investigations into other rare disorders that show strong founder effects, both in SLSJ and in similarly isolated populations worldwide.

In summary, this study advances our understanding of how historical migration, demographic isolation, and genetic inheritance together shaped the current distribution of ARSACS in the SLSJ region. It also establishes a methodological foundation for tracing the origins of rare pathogenic variants by leveraging the synergy between modern genomic tools and historical genealogical resources. Continued application and refinement of these methods hold promise for uncovering the roots of other rare diseases and improving targeted health interventions in founder-effect populations.

8 Bibliography:

[1] S. Vermeer, B. P. van de Warrenburg, E.-J. Kamsteeg, B. Brais, and M. Synofzik, "ARSACS," in *GeneReviews®*, M. P. Adam, J. Feldman, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, and A. Amemiya, Eds., Seattle (WA): University of Washington, Seattle, 1993. Accessed: Jan. 21, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK1255/

[2] M. Synofzik *et al.*, "Autosomal recessive spastic ataxia of Charlevoix Saguenay (ARSACS): expanding the genetic, clinical and imaging spectrum," *Orphanet J. Rare Dis.*, vol. 8, no. 1, p. 41, 2013, doi: 10.1186/1750-1172-8-41.

[3] J. C. Engert *et al.*, "ARSACS, a spastic ataxia common in northeastern Québec, is caused by mutations in a new gene encoding an 11.5-kb ORF," *Nat. Genet.*, vol. 24, no. 2, pp. 120–125, Feb. 2000, doi: 10.1038/72769.

[4] M. De Braekeleer, F. Giasson, J. Mathieu, M. Roy, J. P. Bouchard, and K. Morgan, "Genetic epidemiology of autosomal recessive spastic ataxia of Charlevoix-Saguenay in northeastern Quebec," *Genet. Epidemiol.*, vol. 10, no. 1, pp. 17–25, Jan. 1993, doi: 10.1002/gepi.1370100103.

[5] Sascha Vermeer, MD, PhD, Bart P van de Warrenburg, MD, PhD, Erik-Jan Kamsteeg, PhD, Bernard Brais, MD, PhD, and Matthis Synofzik, MD., *GeneReviews*.

[6] Griffiths, AJF, Wessler, SR, Carroll, SB & Doebley, J., *Introduction to genetic analysis*. New York: WH Freeman, 2012.

[7] Templeton, AR, *Human population genetics and genomics*. Academic press, 2019.

[8] Natasha Ramroop Singh, *Introduction to Genetics*. Thompson Rivers University, 2023.

[9] A. Gulani and T. Weiler, "Genetics, Autosomal Recessive," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Jan. 23, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK546620/

[10] Y. Takiyama, "Autosomal recessive spastic ataxia of Charlevoix-Saguenay," *Neuropathology*, vol. 26, no. 4, pp. 368–375, Aug. 2006, doi: 10.1111/j.1440-1789.2006.00664.x.

[11] D. A. Parfitt *et al.*, “The ataxia protein sacsin is a functional co-chaperone that protects against polyglutamine-expanded ataxin-1,” *Hum. Mol. Genet.*, vol. 18, no. 9, pp. 1556–1565, May 2009, doi: 10.1093/hmg/ddp067.

[12] B. J. Gentil *et al.*, “Sacsin, mutated in the ataxia ARSACS, regulates intermediate filament assembly and dynamics,” *FASEB J.*, vol. 33, no. 2, pp. 2982–2994, Feb. 2019, doi: 10.1096/fj.201801556R.

[13] C. M. Arellano *et al.*, “Generation of a human iPSC line from a patient with autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) caused by mutation in SACSIN gene,” *Stem Cell Res.*, vol. 31, pp. 249–252, Aug. 2018, doi: 10.1016/j.scr.2018.07.012.

[14] X. Li and K. Gehring, “Structural studies of parkin and sacsin: Mitochondrial dynamics in neurodegenerative diseases: Structural Studies of Parkin and Sacsin,” *Mov. Disord.*, vol. 30, no. 12, pp. 1610–1619, Oct. 2015, doi: 10.1002/mds.26357.

[15] A. Romano *et al.*, “Comparative Analysis and Functional Mapping of *SACS* Mutations Reveal Novel Insights into Sacsin Repeated Architecture,” *Hum. Mutat.*, vol. 34, no. 3, pp. 525–537, Mar. 2013, doi: 10.1002/humu.22269.

[16] G. Kozlov *et al.*, “Structural Basis of Defects in the Sacsin HEPN Domain Responsible for Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS),” *J. Biol. Chem.*, vol. 286, no. 23, pp. 20407–20412, June 2011, doi: 10.1074/jbc.M111.232884.

[17] J. F. Anderson, E. Siller, and J. M. Barral, “The Sacsin Repeating Region (SRR): A Novel Hsp90-Related Supra-Domain Associated with Neurodegeneration,” *J. Mol. Biol.*, vol. 400, no. 4, pp. 665–674, July 2010, doi: 10.1016/j.jmb.2010.05.023.

[18] M. Ménade *et al.*, “Structures of ubiquitin-like (Ubl) and Hsp90-like domains of sacsin provide insight into pathological mutations,” *J. Biol. Chem.*, vol. 293, no. 33, pp. 12832–12842, Aug. 2018, doi: 10.1074/jbc.RA118.003939.

[19] P. L. Greer *et al.*, “The Angelman Syndrome Protein Ube3A Regulates Synapse Development by Ubiquitinating Arc,” *Cell*, vol. 140, no. 5, pp. 704–716, Mar. 2010, doi: 10.1016/j.cell.2010.01.026.

[20] “Angelaman syndrome.” [Online]. Available: https://www.ninds.nih.gov/health-information/disorders/angelman-syndrome?utm_source=chatgpt.com

[21] L. E. L. Romano *et al.*, “Multi-omic profiling reveals the ataxia protein sacsin is required for integrin trafficking and synaptic organization,” *Cell Rep.*, vol. 41, no. 5, p. 111580, Nov. 2022, doi: 10.1016/j.celrep.2022.111580.

[22] E. J. Duncan *et al.*, “Altered organization of the intermediate filament cytoskeleton and relocalization of proteostasis modulators in cells lacking the ataxia protein sacsin,” *Hum. Mol. Genet.*, May 2017, doi: 10.1093/hmg/ddx197.

[23] A. Dabbaghizadeh *et al.*, “The J domain of sacsin disrupts intermediate filament assembly,” July 14, 2021. doi: 10.1101/2021.07.14.452331.

[24] M. Girard *et al.*, “Mitochondrial dysfunction and Purkinje cell loss in autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS),” *Proc. Natl. Acad. Sci.*, vol. 109, no. 5, pp. 1661–1666, Jan. 2012, doi: 10.1073/pnas.1113166109.

[25] P. Mesdom *et al.*, “Human Dermal Fibroblast: A Promising Cellular Model to Study Biological Mechanisms of Major Depression and Antidepressant Drug Response,” *Curr. Neuropharmacol.*, vol. 18, no. 4, pp. 301–318, Mar. 2020, doi: 10.2174/1570159X17666191021141057.

[26] S. Kálmán, K. A. Garbett, Z. Janka, and K. Mirnics, “Human dermal fibroblasts in psychiatry research,” *Neuroscience*, vol. 320, pp. 105–121, Apr. 2016, doi: 10.1016/j.neuroscience.2016.01.067.

[27] X. Li *et al.*, “High-Throughput Screening for Ligands of the HEPN Domain of Sacsin,” *PLOS ONE*, vol. 10, no. 9, p. e0137298, Sept. 2015, doi: 10.1371/journal.pone.0137298.

[28] M. Grynberg, H. Erlandsen, and A. Godzik, “HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins,” *Trends Biochem. Sci.*, vol. 28, no. 5, pp. 224–226, May 2003, doi: 10.1016/S0968-0004(03)00060-4.

[29] L. E. L. Romano *et al.*, “The ataxia protein sacsin is required for integrin trafficking and synaptic organization,” Aug. 20, 2021. doi: 10.1101/2021.08.20.456807.

[30] A. Del Bondio *et al.*, “Restoring calcium homeostasis in Purkinje cells arrests neurodegeneration and neuroinflammation in the ARSACS mouse model,” *JCI Insight*, vol. 8, no. 12, p. e163576, June 2023, doi: 10.1172/jci.insight.163576.

[31] J. Bagaria, E. Bagyinszky, and S. S. A. An, “Genetics of Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay (ARSACS) and Role of Sacsin in Neurodegeneration,” *Int. J. Mol. Sci.*, vol. 23, no. 1, p. 552, Jan. 2022, doi: 10.3390/ijms23010552.

[32] M. Bchetnia *et al.*, “Genetic burden linked to founder effects in Saguenay–Lac-Saint-Jean illustrates the importance of genetic screening test availability,” *J. Med. Genet.*, vol. 58, no. 10, pp. 653–665, Oct. 2021, doi: 10.1136/jmedgenet-2021-107809.

[33] S. C. Zschocke J, Potocki L, *Human genetics: from molecules to medicine*. Philadelphia, Pennsylvania, United States: Lippincott Williams & Wilkins, 2011.

[34] F. Longo *et al.*, “Assessment of Sacsin Turnover in Patients With ARSACS: Implications for Molecular Diagnosis and Pathogenesis,” *Neurology*, vol. 97, no. 23, Dec. 2021, doi: 10.1212/WNL.0000000000012962.

[35] “Clinvar c.8844delT.” [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/variation/5512/

[36] I. Thiffault *et al.*, “Diversity of ARSACS Mutations in French-Canadians,” *Can. J. Neurol. Sci. J. Can. Sci. Neurol.*, vol. 40, no. 1, pp. 61–66, Jan. 2013, doi: 10.1017/S0317167100012968.

[37] “clinvar c.7504C>T.” [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/RCV000005848/?utm_source=chatgpt.com

[38] J. P. Bouchard, A. Barbeau, R. Bouchard, and R. W. Bouchard, “Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay,” *Can. J. Neurol. Sci. J. Can. Sci. Neurol.*, vol. 5, no. 1, pp. 61–69, Feb. 1978, doi: 10.1017/S0317167100024793.

[39] E. D. McKenzie, P. N. Sharma, J. S. Parboosingh, Forge Canada Consortium, and O. Suchowersky, “Novel *SACS* Mutation Deviates from the French Canadian ARSACS Phenotype,” *Can. J. Neurol. Sci. J. Can. Sci. Neurol.*, vol. 41, no. 1, pp. 88–89, Jan. 2014, doi: 10.1017/S0317167100016334.

[40] A. W. Crosby, *Ecological imperialism: the biological expansion of Europe, 900-1900*, 2nd ed., new Ed. in Studies in environment and history. Cambridge ; New York: Cambridge University Press, 2004.

[41] A. W. Crosby, “Virgin Soil Epidemics as a Factor in the Aboriginal Depopulation in America,” *William Mary Q.*, vol. 33, no. 2, p. 289, Apr. 1976, doi: 10.2307/1922166.

[42] I. de la statistique du Québec, “Les indicateurs de l’occupation et de la vitalité des territoires - Taux de croissance annuel de la population et population par groupe d’âge,” Apr. 2023, Accessed: Feb. 04, 2025. [Online]. Available: https://policycommons.net/artifacts/3714106/les-indicateurs-de-loccupation-et-de-la-vitalite-des-territoires/4519952/

[43] “demographie_03.pdf.” Accessed: Feb. 04, 2025. [Online]. Available: https://statistique.quebec.ca/docs-ken/vitrine/occupation-vitalite-territoire/documents/demographie_03.pdf

[44] H. Charbonneau, B. Desjardins, J. Légaré, and H. Denis, “The population of the St-Lawrence Valley, 1608-1760,” *Popul. Hist. N. Am.*, pp. 99–142, Jan. 2000.

[45] C. Laprise, “The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population,” *Genes Immun.*, vol. 15, no. 4, pp. 247–255, June 2014, doi: 10.1038/gene.2014.12.

[46] D. Fougères and R. MacLeod, *Montreal: the history of a North American city*. Montreal (Québec): McGill-Queen’s University Press, 2017.

[47] N. Kelley and M. J. Trebilcock, *The making of the mosaic: a history of Canadian immigration policy*, Second edition. Toronto: University of Toronto Press, 2010.

[48] G. Bouchard, C. Laberge, and C. R. Scriver, “Repŕoduction demographique et transmission genetique dans le Nord-Est de la province de Quebec (18e–20e siecles),” *Eur. J. Popul.*, vol. 4, no. 1, pp. 39–67, Sept. 1988, doi: 10.1007/BF01797106.

[49] H. Vézina, M. Tremblay, B. Desjardins, and L. Houde, “Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise,” *Cah. Qué. Démographie*, vol. 34, no. 2, pp. 235–258, Nov. 2006, doi: 10.7202/014011ar.

[50] G. Bouchard, Ed., *Histoire d’un génôme: population et génétique dans l’est du Québec*. Québec: Presses de l’Univ. du Québec, 1991.

[51] H. Charbonneau, Ed., *Naissance d'une population: les Français établis au Canada au XVIIe siècle*, Nouvelle édition revue et corrigée. Aubervilliers] : [Montréal, Québec: Ined Éditions ; Les Presses de l'Université de Montréal, 2020.

[52] Y. Landry, "Les filles du roi émigrées au Canada au XVIIe siècle, ou un exemple de choix du conjoint en situation de déséquilibre des sexes," *Hist. Économie Société*, vol. 11, no. 2, pp. 197–216, 1992, doi: 10.3406/hes.1992.1632.

[53] A. Filser and K. P. Willführ, "Sex ratios and union formation in the historical population of the St. Lawrence Valley," *PLOS ONE*, vol. 17, no. 6, p. e0268039, June 2022, doi: 10.1371/journal.pone.0268039.

[54] A. Gagnon, K. R. Smith, M. Tremblay, H. Vézina, P. Paré, and B. Desjardins, "Is there a trade-off between fertility and longevity? A comparative study of women from three large historical databases accounting for mortality selection," *Am. J. Hum. Biol.*, vol. 21, no. 4, pp. 533–540, July 2009, doi: 10.1002/ajhb.20893.

[55] M. R. Haines and R. H. Steckel, *A population history of North America*. Cambridge: Cambridge university press, 2000.

[56] M. Tremblay, M. Letendre, L. Houde, and H. Vézina, "The Contribution of Irish Immigrants to the Quebec (Canada) Gene Pool: An Estimation Using Data from Deep-Rooted Genealogies: La contribution des immigrants irlandais au pool génique du Québec (Canada) : une estimation à partir de données généalogiques," *Eur. J. Popul. Rev. Eur. Démographie*, vol. 25, no. 2, pp. 215–233, May 2009, doi: 10.1007/s10680-008-9170-x.

[57] A. Laberge *et al.*, "Population history and its impact on medical genetics in Quebec," *Clin. Genet.*, vol. 68, no. 4, pp. 287–301, Oct. 2005, doi: 10.1111/j.1399-0004.2005.00497.x.

[58] C. R. Scriver, "Human Genetics: Lessons from Quebec Populations," *Annu. Rev. Genomics Hum. Genet.*, vol. 2, no. 1, pp. 69–101, Sept. 2001, doi: 10.1146/annurev.genom.2.1.69.

[59] Bouchard G, Tremblay M, *Français de France et Français du Canada Les parlers de l'Ouest de la France, du Québec et de l'Acadie*. in Série «Dialectologie». Université Lyon III, Centre d'Etudes Linguistiques Jacques Goudet, 1995.

[60] C. Bherer, D. Labuda, M. Roy-Gagnon, L. Houde, M. Tremblay, and H. Vézina, “Admixed ancestry and stratification of Quebec regional populations,” *Am. J. Phys. Anthropol.*, vol. 144, no. 3, pp. 432–441, Mar. 2011, doi: 10.1002/ajpa.21424.

[61] L. Gagnon, C. Moreau, C. Laprise, H. Vézina, and S. L. Girard, “Deciphering the genetic structure of the Quebec founder population using genealogies,” *Eur. J. Hum. Genet.*, vol. 32, no. 1, pp. 91–97, Jan. 2024, doi: 10.1038/s41431-023-01356-2.

[62] M.-H. Roy-Gagnon *et al.*, “Genomic and genealogical investigation of the French Canadian founder population structure,” *Hum. Genet.*, vol. 129, no. 5, pp. 521–531, May 2011, doi: 10.1007/s00439-010-0945-x.

[63] C. Moreau *et al.*, “Genetic heterogeneity in regional populations of Quebec—Parental lineages in the Gaspe Peninsula,” *Am. J. Phys. Anthropol.*, vol. 139, no. 4, pp. 512–522, Aug. 2009, doi: 10.1002/ajpa.21012.

[64] G. Rouleau, “Portrait démogénétique de la RMR de Saguenay”.

[65] L. Anderson-Trocmé *et al.*, “On the genes, genealogies, and geographies of Quebec,” *Science*, vol. 380, no. 6647, pp. 849–855, May 2023, doi: 10.1126/science.add5300.

[66] G. Bouchard, “Sur la reproduction familiale en milieu rural : systèmes ouverts et systèmes clos,” *Rech. Sociographiques*, vol. 28, no. 2–3, pp. 229–251, Apr. 2005, doi: 10.7202/056289ar.

[67] H. Vézina, M. Tremblay, and L. Houde, “Mesures de l’apparentement biologique au Saguenay-Lac-St-Jean (Québec, Canada) à partir de reconstitutions généalogiques,” *Ann. Démographie Hist.*, vol. 108, no. 2, p. 67, 2004, doi: 10.3917/adh.108.0067.

[68] C. Pouyez, Y. Lavoie, and G. Bouchard, *Les Saguenayens: introduction à l’histoire des populations du Saguenay, XVIe-XXe siècles*. Sillery, Québec: Presses de l’Université du Québec, 1983.

[69] E.-M. Lavoie, M. Tremblay, L. Houde, and H. Vézina, “Demogenetic Study of Three Populations within a Region with Strong Founder Effects,” *Public Health Genomics*, vol. 8, no. 3, pp. 152–160, 2005, doi: 10.1159/000086758.

[70] E. Heyer and M. Tremblay, “Variability of the genetic contribution of Quebec population founders associated to some deleterious genes,” *Am. J. Hum. Genet.*, vol. 56, no. 4, pp. 970–978, Apr. 1995.

[71] E. Heyer, M. Tremblay, and B. Desjardins, “Seventeenth-century European origins of hereditary diseases in the Saguenay population (Quebec, Canada),” *Hum. Biol.*, vol. 69, no. 2, pp. 209–225, Apr. 1997.

[72] T. Cruz Marino *et al.*, “Portrait of autosomal recessive diseases in the FRENCH-CANADIAN founder population of SAGUENAY-LAC-SAINT-JEAN,” *Am. J. Med. Genet. A.*, vol. 191, no. 5, pp. 1145–1163, May 2023, doi: 10.1002/ajmg.a.63147.

[73] “Offre de tests de porteur pour quatre maladies héréditaires récessives chez les personnes originaires des régions du Saguenay–Lac-Saint-Jean, de Charlevoix et de la Haute-Côte-Nord,” Gouvernement du Québec. Accessed: Feb. 04, 2025. [Online]. Available: https://www.quebec.ca/sante/conseils-et-prevention/depistage-et-offre-de-tests-de-porteur/tests-de-porteur-maladies-hereditaires-recessive

[74] M. Bchetnia *et al.*, “Genetic burden linked to founder effects in Saguenay–Lac-Saint-Jean illustrates the importance of genetic screening test availability,” *J. Med. Genet.*, vol. 58, no. 10, pp. 653–665, Oct. 2021, doi: 10.1136/jmedgenet-2021-107809.

[75] É. Michel *et al.*, “Rare diseases load through the study of a regional population,” Oct. 30, 2024. doi: 10.1101/2024.10.29.24316346.

[76] T. Kivisild, “Founder Effect,” in *Brenner’s Encyclopedia of Genetics*, Elsevier, 2013, pp. 100–101. doi: 10.1016/B978-0-12-374984-0.00552-0.

[77] C. T. Bergstrom and L. A. Dugatkin, *Evolution*, Second edition. New York London: W.W. Norton & Company, 2016.

[78] F. B. Livingstone, “The founder effect and deleterious genes,” *Am. J. Phys. Anthropol.*, vol. 30, no. 1, pp. 55–59, Jan. 1969, doi: 10.1002/ajpa.1330300107.

[79] M. Kirk, “Genetic drift and founder effects: Implications for population genetics, conservation, and human health,” *Genet. Mol. Res.*.

[80] W. H. Lowe, R. P. Kovach, and F. W. Allendorf, “Population Genetics and Demography Unite Ecology and Evolution,” *Trends Ecol. Evol.*, vol. 32, no. 2, pp. 141–152, Feb. 2017, doi: 10.1016/j.tree.2016.12.002.

[81] S. Kc, “Population Stabilization, Human,” in *Encyclopedia of Biodiversity*, Elsevier, 2013, pp. 199–209. doi: 10.1016/B978-0-12-384719-5.00413-5.

[82] D. S. Woodruff, “Populations, Species, and Conservation Genetics,” in *Encyclopedia of Biodiversity*, Elsevier, 2001, pp. 811–829. doi: 10.1016/B0-12-226865-2/00355-2.

[83] S. Wright, “EVOLUTION IN MENDELIAN POPULATIONS,” *Genetics*, vol. 16, no. 2, pp. 97–159, Mar. 1931, doi: 10.1093/genetics/16.2.97.

[84] D. Drayna, “Founder Mutations,” *Sci. Am.*, vol. 293, no. 4, pp. 78–85, Oct. 2005, doi: 10.1038/scientificamerican1005-78.

[85] S. R. Browning, “Estimation of Pairwise Identity by Descent From Dense Genetic Marker Data in a Population Sample of Haplotypes,” *Genetics*, vol. 178, no. 4, pp. 2123–2132, Apr. 2008, doi: 10.1534/genetics.107.084624.

[86] S. R. Browning and B. L. Browning, “High-Resolution Detection of Identity by Descent in Unrelated Individuals,” *Am. J. Hum. Genet.*, vol. 86, no. 4, pp. 526–539, Apr. 2010, doi: 10.1016/j.ajhg.2010.02.021.

[87] “meiosis.” [Online]. Available: https://www.genome.gov/genetics-glossary/Meiosis

[88] E. L. Sticca, G. M. Belbin, and C. R. Gignoux, “Current Developments in Detection of Identity-by-Descent Methods and Applications,” *Front. Genet.*, vol. 12, p. 722602, Sept. 2021, doi: 10.3389/fgene.2021.722602.

[89] R. Ferla *et al.*, “Founder mutations in BRCA1 and BRCA2 genes,” *Ann. Oncol.*, vol. 18, pp. vi93–vi98, June 2007, doi: 10.1093/annonc/mdm234.

[90] M. De Braekeleer, “Hereditary Disorders in Saguenay-Lac-St-Jean (Quebec, Canada),” *Hum. Hered.*, vol. 41, no. 3, pp. 141–146, 1991, doi: 10.1159/000153992.

[91] D. Gauvreau and M. Bourque, “Mouvements migratoires et familles : le peuplement du Saguenay avant 1911,” *Rev. Hist. Am. Fr.*, vol. 42, no. 2, pp. 167–192, Sept. 2008, doi: 10.7202/304677ar.

[92] “Ongoing Research for Autosomal Recessive Spastic Ataxia of Charlevoix-Saguenay.” [Online]. Available: https://checkrare.com/ongoing-research-for-autosomal-recessive-spastic-ataxia-of-charlevoix-saguenay/

[93] “BALSAC [Internet]. BALSAC. Available from: https://balsac.uqac.ca/.”

[94] M. H. D. Larmuseau, A. Van Geystelen, M. Van Oven, and R. Decorte, “Genetic genealogy comes of age: Perspectives on the use of deep-rooted pedigrees in human population genetics,” *Am. J. Phys. Anthropol.*, vol. 150, no. 4, pp. 505–511, Apr. 2013, doi: 10.1002/ajpa.22233.

[95] H. Vézina and J.-S. Bournival, “An Overview of the BALSAC Population Database. Past Developments, Current State and Future Prospects,” *Hist. Life Course Stud.*, vol. 9, pp. 114–129, Aug. 2020, doi: 10.51964/hlcs9299.

[96] H. Gauvin *et al.*, “GENLIB: an R package for the analysis of genealogical data,” *BMC Bioinformatics*, vol. 16, no. 1, p. 160, Dec. 2015, doi: 10.1186/s12859-015-0581-5.

[97] L. Houde, J.-F. Lefebvre, V. Roy-Lagace, and S. Lemieux, “GENLIB: Genealogical Data Analysis.” p. 1.1.10, Sept. 22, 2014. doi: 10.32614/CRAN.package.GENLIB.

[98] A. Ochoa and J. D. Storey, “Estimating FST and kinship for arbitrary population structures,” *PLoS Genet.*, vol. 17, no. 1, p. e1009241, Jan. 2021, doi: 10.1371/journal.pgen.1009241.

[99] C. Moreau, C. Bhérer, H. Vézina, M. Jomphe, D. Labuda, and L. Excoffier, “Deep Human Genealogies Reveal a Selective Advantage to Be on an Expanding Wave Front,” *Science*, vol. 334, no. 6059, pp. 1148–1150, Nov. 2011, doi: 10.1126/science.1212880.

[100] “https://cartagene.qc.ca/en/.”

[101] P. Awadalla *et al.*, “Cohort profile of the CARTaGENE study: Quebec’s population-based biobank for public health and personalized genomics,” *Int. J. Epidemiol.*, vol. 42, no. 5, pp. 1285–1299, Oct. 2013, doi: 10.1093/ije/dys160.

[102] S. H. Slifer, “PLINK: Key Functions for Data Analysis,” *Curr. Protoc. Hum. Genet.*, vol. 97, no. 1, p. e59, Apr. 2018, doi: 10.1002/cphg.59.

[103] “Hardy-Weinberg law.” [Online]. Available: https://www.britannica.com/science/Hardy-Weinberg-law

[104] S. Purcell *et al.*, “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sept. 2007, doi: 10.1086/519795.

[105] “UMAP.” [Online]. Available: https://github.com/tkonopka/umap

[106] O. Francois, M. Currat, N. Ray, E. Han, L. Excoffier, and J. Novembre, “Principal Component Analysis under Population Genetic Models of Range Expansion and Admixture,” *Mol. Biol. Evol.*, vol. 27, no. 6, pp. 1257–1268, June 2010, doi: 10.1093/molbev/msq010.

[107] M. Ringnér, “What is principal component analysis?,” *Nat. Biotechnol.*, vol. 26, no. 3, pp. 303–304, Mar. 2008, doi: 10.1038/nbt0308-303.

[108] J. Novembre and M. Stephens, “Interpreting principal component analyses of spatial population genetic variation,” *Nat. Genet.*, vol. 40, no. 5, pp. 646–649, May 2008, doi: 10.1038/ng.139.

[109] D. Reich, A. L. Price, and N. Patterson, “Principal component analysis of genetic data,” *Nat. Genet.*, vol. 40, no. 5, pp. 491–492, May 2008, doi: 10.1038/ng0508-491.

[110] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel, “UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts,” *PLOS Genet.*, vol. 15, no. 11, p. e1008432, Nov. 2019, doi: 10.1371/journal.pgen.1008432.

[111] J. Healy and L. McInnes, “Uniform manifold approximation and projection,” *Nat. Rev. Methods Primer*, vol. 4, no. 1, p. 82, Nov. 2024, doi: 10.1038/s43586-024-00363-x.

[112] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” 2018, *arXiv*. doi: 10.48550/ARXIV.1802.03426.

[113] S. R. Browning and B. L. Browning, “Identity by Descent Between Distant Relatives: Detection and Applications,” *Annu. Rev. Genet.*, vol. 46, no. 1, pp. 617–633, Dec. 2012, doi: 10.1146/annurev-genet-110711-155534.

[114] B. L. Browning and S. R. Browning, “Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data,” *Genetics*, vol. 194, no. 2, pp. 459–471, June 2013, doi: 10.1534/genetics.113.150029.

[115] S. R. Browning and B. L. Browning, “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering,” *Am. J. Hum. Genet.*, vol. 81, no. 5, pp. 1084–1097, Nov. 2007, doi: 10.1086/521987.

[116] B. L. Browning and S. R. Browning, “Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data,” *Genetics*, vol. 194, no. 2, pp. 459–471, June 2013, doi: 10.1534/genetics.113.150029.

[117] G. Glusman, H. C. Cox, and J. C. Roach, “Whole-genome haplotyping approaches and genomic medicine,” *Genome Med.*, vol. 6, no. 9, p. 73, Sept. 2014, doi: 10.1186/s13073-014-0073-7.

[118] “geneHapR Library.” [Online]. Available: https://cran.r-project.org/web/packages/geneHapR/index.html

[119] M. Slatkin, “Linkage disequilibrium — understanding the evolutionary past and mapping the medical future,” *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–485, June 2008, doi: 10.1038/nrg2361.

[120] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, “Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium,” *Am. J. Hum. Genet.*, vol. 74, no. 1, pp. 106–120, Jan. 2004, doi: 10.1086/381000.

[121] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, “Haploview: analysis and visualization of LD and haplotype maps,” *Bioinformatics*, vol. 21, no. 2, pp. 263–265, Jan. 2005, doi: 10.1093/bioinformatics/bth457.

[122] M. L. Selle, I. Steinsland, F. Lindgren, V. Brajkovic, V. Cubric-Curik, and G. Gorjanc, "Hierarchical Modelling of Haplotype Effects on a Phylogeny," *Front. Genet.*, vol. 11, p. 531218, Jan. 2021, doi: 10.3389/fgene.2020.531218.

[123] A. R. Templeton, "Haplotype Trees and Modern Human Origins," *Am. J. Phys. Anthropol.*, vol. 128, no. S41, pp. 33–59, 2005, doi: 10.1002/ajpa.20351.

[124] *RAxML*. [Online]. Available: https://github.com/stamatak/standard-RAxML

[125] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, Nov. 2006, doi: 10.1093/bioinformatics/btl446.

[126] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, May 2014, doi: 10.1093/bioinformatics/btu033.

[127] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinformatics*, vol. 35, no. 21, pp. 4453–4455, Nov. 2019, doi: 10.1093/bioinformatics/btz305.

[128] S. R. Browning and B. L. Browning, "Identity by Descent Between Distant Relatives: Detection and Applications," *Annu. Rev. Genet.*, vol. 46, no. 1, pp. 617–633, Dec. 2012, doi: 10.1146/annurev-genet-110711-155534.

[129] L. Almasy and J. Blangero, "Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees," *Am. J. Hum. Genet.*, vol. 62, no. 5, pp. 1198–1211, May 1998, doi: 10.1086/301844.

[130] D. H. Huson and D. Bryant, "Application of Phylogenetic Networks in Evolutionary Studies," *Mol. Biol. Evol.*, vol. 23, no. 2, pp. 254–267, Feb. 2006, doi: 10.1093/molbev/msj030.

[131] H. J. Bandelt, P. Forster, and A. Rohl, "Median-joining networks for inferring intraspecific phylogenies," *Mol. Biol. Evol.*, vol. 16, no. 1, pp. 37–48, Jan. 1999, doi: 10.1093/oxfordjournals.molbev.a026036.

[132] E. A. Thompson, “Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations,” *Genetics*, vol. 194, no. 2, pp. 301–326, June 2013, doi: 10.1534/genetics.112.148825.

[133] A.-L. Leutenegger *et al.*, “Estimation of the inbreeding coefficient through use of genomic data,” *Am. J. Hum. Genet.*, vol. 73, no. 3, pp. 516–523, Sept. 2003, doi: 10.1086/378207.

[134] *Pedigree-alignment-method*. [Online]. Available: https://gravellab.github.io/pedigree-ARG-matching/

[135] K. Hatzikotoulas, A. Gilly, and E. Zeggini, “Using population isolates in genetic association studies,” *Brief. Funct. Genomics*, vol. 13, no. 5, pp. 371–377, Sept. 2014, doi: 10.1093/bfgp/elu022.

[136] M. T. Lynch *et al.*, “The burden of pathogenic variants in clinically actionable genes in a founder population,” *Am. J. Med. Genet. A.*, vol. 185, no. 11, pp. 3476–3484, Nov. 2021, doi: 10.1002/ajmg.a.62472.

[137] M. Bchetnia *et al.*, “Genetic burden linked to founder effects in Saguenay–Lac-Saint-Jean illustrates the importance of genetic screening test availability,” *J. Med. Genet.*, vol. 58, no. 10, pp. 653–665, Oct. 2021, doi: 10.1136/jmedgenet-2021-107809.

[138] A. L. Lewanski, M. C. Grundler, and G. S. Bradburd, “The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics,” *PLOS Genet.*, vol. 20, no. 1, p. e1011110, Jan. 2024, doi: 10.1371/journal.pgen.1011110.

9 Ethical considerations

The study was approved by the Ethics Review Board of the CIUSSS of the SLSJ. A written informed consent was obtained from all participants.

This study was approved by the ethics board of the Université du Québec à Chicoutimi (UQAC).

10 Appendix

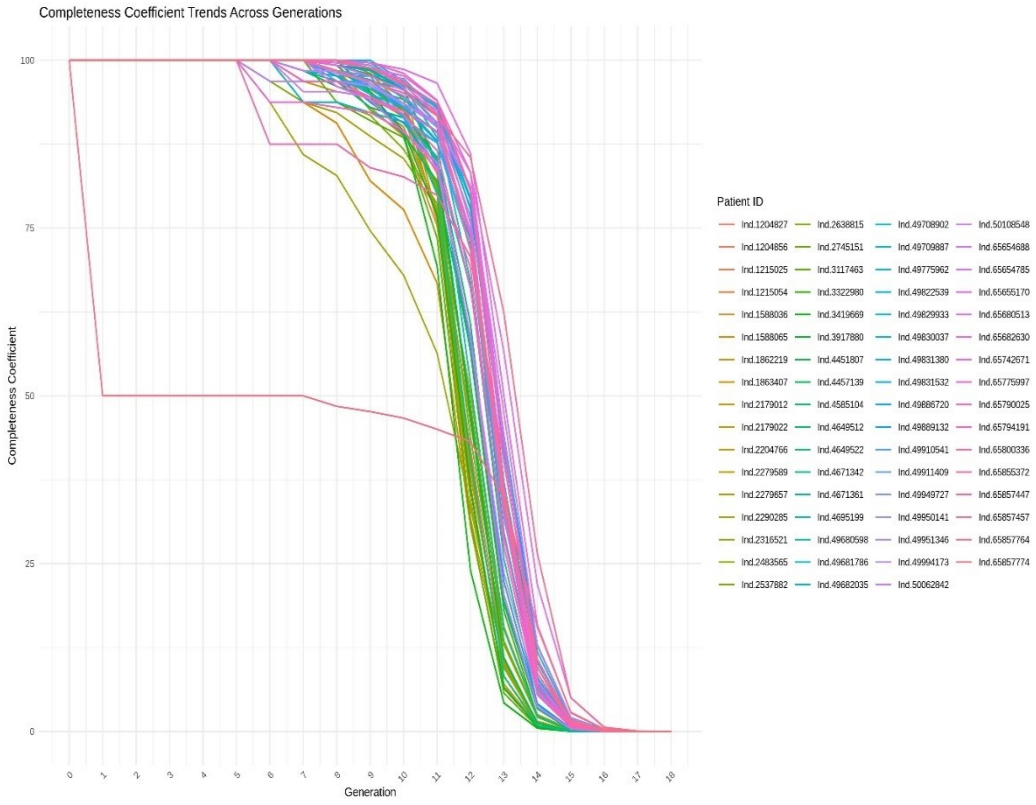

Figure A1. The completeness for 67 patients for whom both genealogical and sequencing data were available.

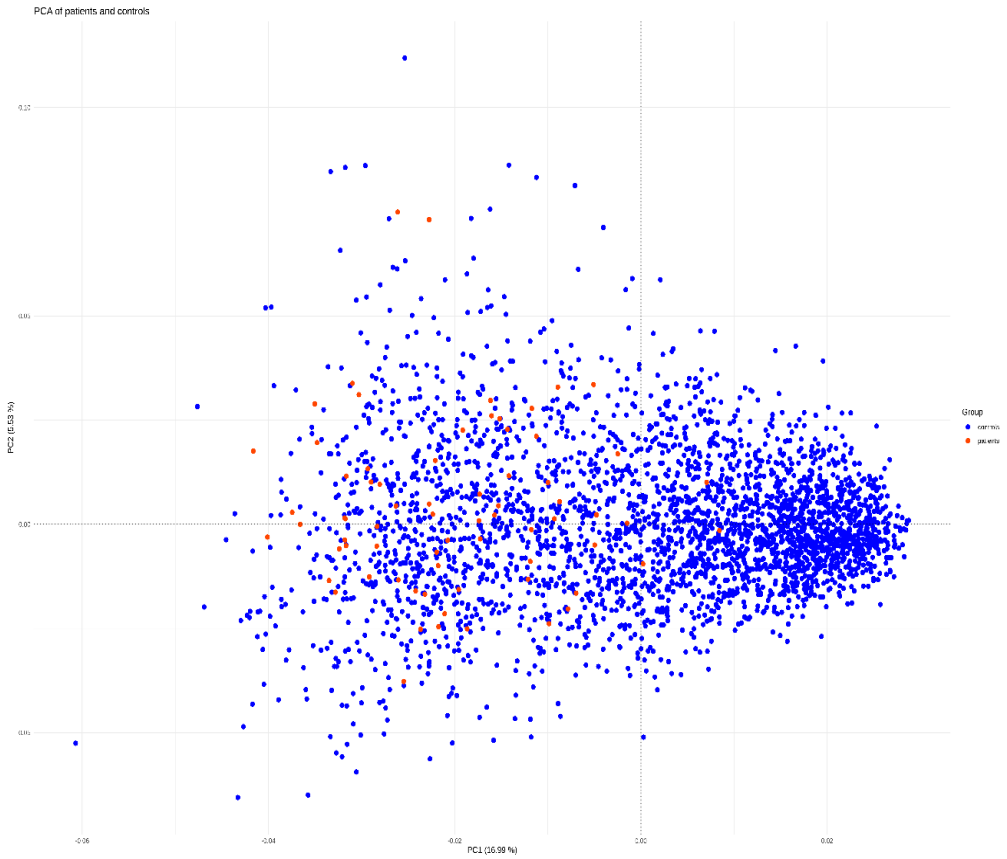

Figure A2. PCA of the first 2 principal components of the genotype data of the patients and controls.

This captures the largest proportion of the total variance in the data. Individuals with similar genetic backgrounds tend to cluster closely together.

Figure A3. Visualization of the 2 Mb haplotypes surrounding the SACS locus.
The x-axis represents the approximate genomic positions of the linkage disequilibrium (LD) blocks that constitute each haploblock structure, as identified in our haplotype reconstruction. The y-axis lists the unique haploblock identifiers, defined during the haplotype analysis. On the right side of the figure, the distribution of individuals—patients versus carriers—sharing each haploblock segment is shown, illustrating how these haplotypes are distributed within the study population and offering insights into the inheritance patterns of the founder haplotypes. The ARSACS causal variant (c.8844delT), located at position 23,909,171 on chromosome 13 (GRCh38 reference genome), is marked by a red dashed vertical line. This annotation provides a clear point of reference for comparing the haplotype structures relative to the disease-causing variant and highlights the segments most strongly associated with the founder mutation.

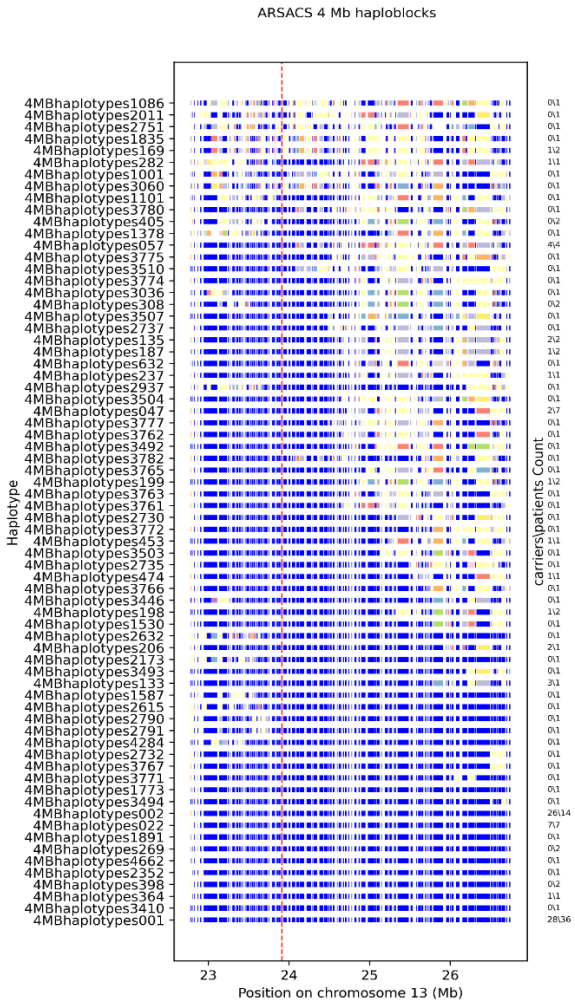

Figure A4. Visualization of the 4 Mb haplotypes surrounding the SACS locus. The x-axis represents the approximate genomic positions of the linkage disequilibrium (LD) blocks that constitute each haploblock structure, as identified in our haplotype reconstruction. The y-axis lists the unique haploblock identifiers, defined during the haplotype analysis. On the right side of the figure, the distribution of individuals—patients versus carriers—sharing each haploblock segment is shown, illustrating how these haplotypes are distributed within the study population and offering insights into the inheritance patterns of the founder haplotypes. The ARSACS causal variant (c.8844delT), located at position 23,909,171 on chromosome 13 (GRCh38 reference genome), is marked by a red dashed vertical line. This annotation provides a clear point of reference for comparing the haplotype structures relative to the disease-causing variant and highlights the segments most strongly associated with the founder mutation.

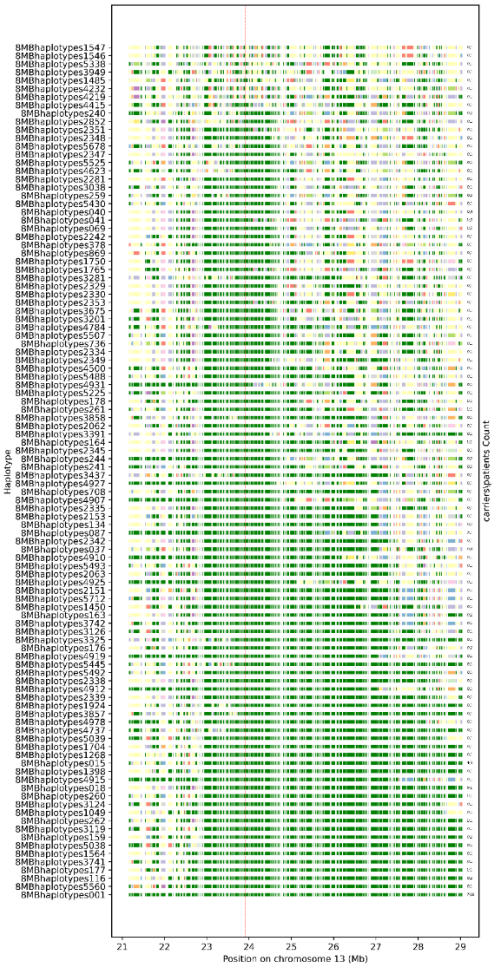

Figure A5. Visualization of the 8 Mb haplotypes surrounding the SACS locus. The x-axis represents the approximate genomic positions of the linkage disequilibrium (LD) blocks that constitute each haploblock structure, as identified in our haplotype reconstruction. The y-axis lists the unique haploblock identifiers, defined during the haplotype analysis. On the right side of the figure, the distribution of individuals—patients versus carriers—sharing each haploblock segment is shown, illustrating how these haplotypes are distributed within the study population and offering insights into the inheritance patterns of the founder haplotypes. The ARSACS causal variant (c.8844delT), located at position 23,909,171 on chromosome 13 (GRCh38 reference genome), is marked by a red dashed vertical line. This annotation provides a clear point of reference for comparing the haplotype structures relative to the disease-causing variant and highlights the segments most strongly associated with the founder mutation.