Genetic Investigation of Mendelian Disorders in the Founder Populations of Newfoundland & Labrador.

by

© Daniel R. Evans

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirement for the degree of Doctor of Philosophy

Discipline of Genetics, Faculty of Medicine

Memorial University of Newfoundland

May 2019

St. John's Newfoundland

Abstract

Exploring the genetic basis of monogenic disorders imparts fundamental insights into human molecular biology. Moreover, cataloguing and describing pathogenic mutations offers new genetic tests to diagnose and identify mutation carriers worldwide. Sometimes, reaching a genetic diagnosis provides long sought-after answers for patients and their families. Likewise, comprehending the genetic and clinical spectrum of rare disorders aids clinicians in managing their patients. Further, clinical knowledge obtained from genetic studies allows clinicians to anticipate and screen for potential complications.

Founder populations experience higher burdens of certain Mendelian disorders and thereby provide unique circumstances to conduct genetic studies. Several founder populations exist in the Canadian province of Newfoundland & Labrador. These have proven instrumental for genetic discoveries. My thesis investigated the genetic basis of three Mendelian disorders in families from Newfoundland. These disorders include retinitis pigmentosa, Weill-Marchesani Syndrome and hereditary colorectal cancer.

First, whole exome sequencing and linkage analyses were employed to investigate retinitis pigmentosa in a large kindred living on the Great Northern Peninsula of Newfoundland. This identified a linked region on chromosome 2 (logarithm of the odds 4.89 [$\theta = 0$]) encompassing a novel pathogenic 25 kb deletion (c.845-1450del; p.Ala282_His483del) in *MERTK*, which segregated in the family. The

molecular features and clinical manifestations of the deletion were characterized and study findings were explained to family members.

Next, Weill-Marchesani Syndrome was investigated in a Newfoundland family using whole exome sequencing and homozygosity mapping. This identified a novel homozygous pathogenic missense variant (c.3068G>A, p.C1023Y) in *ADAMTS17*. Transfection of *ADAMTS17* p.C1023Y expression plasmids into HEK293T cells revealed significantly reduced secretion of ADAMTS17 into the extracellular matrix. The clinical and molecular consequences of the variant were characterized and the family members were counselled.

Finally, candidate gene screening of *GALNT12* in a cohort of 479 colorectal cancer cases from Newfoundland identified eight rare variants (p.Asp303Asp; p.Arg297Trp; p.His101Gln; p.Ile142Thr; p.Glu239Gln; p.Thr286Met; p.Val290Phe; c.732-8G>T), which were overrepresented in cases compared to controls (*N*=400) (P=0.0381). Six of eight variants showed reduced GALNT12 enzyme activity, providing supportive evidence of a role for this gene in colorectal cancer. In summary, each study above imparted novel insights into genetic disorders in Newfoundland.

Foreword

"Nature is nowhere accustomed more openly to display her secret mysteries than in cases where she shows traces of her workings apart from the beaten path; nor is there any better way to advance the proper practice of medicine than to give our minds to the discovery of the usual law of Nature by careful investigation of cases of rarer forms of disease. For it has been found, in almost all things, that what they contain of useful or applicable is hardly perceived unless we are deprived of them, or they become deranged in some way."

Sir William Harvey as translated by Sir Archibald Garrod (1928)

“There are many of us to whom a scientific research appeals in proportion to its practical utility, and others for whom a discovery loses much of its charm when applied to practical uses. So, too, I take it, there have always been, and always will be, many members of our profession to whom the study of those common diseases which are of so great practical importance, and to cope with which is the daily task of the medical man, make the strongest appeal, whereas for others, who form but a small minority, the investigation of rare maladies has a special attraction.”

Sir Archibald Garrod (1928)

Acknowledgments

I would like to thank all the individuals who lent their confidence and support along this journey. Beginning in the summer of 2011, when I was a 23-year-old undergraduate on a summer studentship, Dr. Jane Green first inspired me to pursue a PhD in genetics. She taught me the value of the patient's perspective, critical analysis of data, and the application of science in the context of Mendelian disorders. My dream of pursuing a combined M.D-Ph.D was first culminated somewhere amongst a stack of 100 charts and this inspiring mentor.

Absolutely none of my ambitions could ever been realized without the many opportunities granted to me by Dr. Michael Woods. He took a chance on a young student who was interested in graduate studies, but who also had aspirations for medical school -- in a university culture that does not cultivate such aspirations. Never will I be able to repay Mike for his wisdom, mentorship, and countless life lessons over the years. I consider him my supervisor and mentor, but more importantly, a valued lifelong friend.

I would especially like to thank Dr. Ann Dorward, whom I have known always known to be a strong student advocate, particularly with respect to the M.D-Ph.D program. I would like to thank Dr. Donald McKay, Associate Dean of Undergraduate Medical Education (UGME), who facilitated my transitions between M.D and Ph.D programs. Importantly, I would like to thank the School of Graduate Studies for their vision and leadership.

To my parents Wendy Evans and Rob Hayes -- thank you so much for your unwavering support and belief in me. Thanks also to my siblings Rhyder Evans, Robyn Evans and Cody Evans for always being there. To all the friends and colleagues who I have met through the Discipline of Genetics over the years, notably my labmates Julia Pennell, Amy Powell and Robyn Byrne, I would like to express my sincere gratitude.

This work embodies an exercise in dedication and perseverance, and I hope it will someday allow me to become a well-rounded clinician-scientist. As I look forward to everything the future holds, I will remember to question things and always to challenge my own personal beliefs.

Sincerely,

Daniel Evans

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3D: 3 dimensional

^{32}P: radioactive phosphate isotope

AC-1: Amsterdam Criteria 1

ACMG: American College of Medical Genetics and Genomics

adRP: autosomal dominant Retinitis Pigmentosa

AP: Andrew Patterson

arRP: autosomal recessive RP

ARVC5: Arrhythmogenic Right Ventricular Cardiomyopathy Type 5

ARVD5: Arrhythmogenic Right Ventricular Dysplasia Type 5

BAC: Bacterial Artificial Chromosome

BAF: Bridget A. Fernandez

BBS: Bardet-Biedl Syndrome

BLAST: Basic Local Alignment Search Tool

cal YBP: Calibrated Years Before Present

CLB: Chandree L. Beaulieu

CMG: Center for Mendelian Genomics

CNIB: The Canadian National Institute for the Blind

CRC: Colorectal cancer

CRM: Christian R. Marshall

DRE: Daniel Rory Evans

DH: Dirk Hubmacher

ECM: Extracellular Matrix

EQ: Erina Quinn

FAP: Familial Adenomatous Polyposis

FBN1: Fibrillin-1

FCCTX: Familial Colorectal Cancer Type X

FORGE: Finding of Rare Disease Genes (Consortium)

FSS: Freeman–Sheldon syndrome

GJJ: Gordon J. Johnson

GS20: Genome Sequencer 20

HGMD: The Human Gene Mutation Database

HHT: Hereditary Hemorrhagic Telangactasia

HMG-CoA-reductase: 3-hydroxy-3-methylglutaryl coenzyme A reductase

HNPCC: Hereditary Nonpolyposis Colorectal Cancer

JF: Justin French

JHW: James G. Whelan

JM: Jacek Majewski

JPS: Juvenile Polyposis Syndrome

JS: Jeremy Schwartzentruber

JSG: Jane S. Green

KG: Kishore Guda

KMB: Kym M. Boycott

LaR: Lakshmeswari Ravi

LDL: Low-density lipoprotein

LeR: Leslie Reveredo

LOD: Logarithm of Odds

LS: Lynch Syndrome

MAD: Matthew A. Deardorff

MAP: *MUTYH*-associated polyposis

MEN: Multiple Endocrine Neoplasia

MIM: Mendelian Inheritance in Man

MMR: Mismatch Repair

MOW: Michael O. Woods

MPS: Massively Parallel Sequencing

MSI: Microsatellite Instability

MSI-H: High Microsatellite Instability

MSS: Microsatellite Stable

NCL: Neuronal Ceroid Lipofuscinosis

NIH: National Institutes of Health

NL: Newfoundland & Labrador

NMR: Nicole M. Roslin

OMIM: Online Mendelian Inheritance in Man

PCR: Polymerase Chain Reaction

PEDF: Pigment Epithelium Derived Factor

PJS: Peutz-Jeghers syndrome

PKD: Polycystic Kidney Disease

RCS: Royal College of Surgeons

RFLP: Restriction Fragment Length Polymorphisms

RNase: Ribonuclease

RP: Retinitis pigmentosa

RPE: Retinal Pigment Epithelium

SA: Suneel Apte

SF: Somayyeh Fahiminiya

SNP: Single Nucleotide Polymorphism

SoLiD: Sequencing by Oligonucleotide Ligation and Detection

SV: Srividya Venkitachalam

TAP: Tara A. Paton

TAG: Thomas A. Gerkin

TGF-β: Transforming Growth Factor Beta

tRNA: Transfer RNA

WMS: Weill-Marchesani Syndrome

WQ: Wen Qin

List of Appendices

Chapter 1. Overview: Mendelian Genetics and Unique Aspects of the Newfoundland & Labrador Founder Populations

1.1 Genetic Discovery

For me, nothing is more exciting than pursuing a potential genetic discovery. These moments of enthusiasm sustained me during my M.D. and Ph.D. programs (Memorial University of Newfoundland). This thesis outlines genetic discoveries garnered from the investigation of three Mendelian disorders: retinitis pigmentosa (RP), Weill-Marchesani Syndrome (WMS) and hereditary predisposition for colorectal cancer (CRC). The aim of each research project was to apply modern genetic strategies to uncover novel genes and characterize variants implicated in unsolved Mendelian disorders in Newfoundland & Labrador (NL). The purpose of the endeavor was to reach a genetic diagnosis for NL families (and potentially others worldwide), while gaining insight into pathologic mechanisms underlying each disorder.

In this general introductory section, I review the importance and purpose of Mendelian disease gene exploration and describe how technological advances have shifted the paradigm of genetic discovery. I then discuss the importance of genetic isolates for Mendelian discovery, before acquainting the reader with unique aspects of the NL population. Thereafter, each proceeding thesis chapter examines an independent research project investigating one of the three Mendelian disorders listed previously. Such studies expand our knowledge of the pathogenesis of these

disorders, further our understanding of the unique founder populations of NL and provide answers to family members.

Such studies would not be possible without the generosity and kindness of all research study participants, who contributed their personal and family histories, as well as their biological heritage. Certainly, these gifts are what drive genetic discoveries. Given such substantial contributions, an emergent question becomes – what are Mendelian disorders, and why do we study them?

1.2 Study of Mendelian Disorders

Mendelian disorders are phenotypes caused by genetic aberration and germline transmission of a single altered gene product or locus (i.e. monogenic inheritance). These alterations cause observable or measurable phenotypes. Studying Mendelian disorders provides fundamental clues into the pathophysiology of human diseases. Indeed, renowned clinician-scientist Sir Archibald Garrod (1857-1936) once said "…we find in rare diseases the keys to not a few dark places of physiology and pathology" (Garrod, 1928). Understanding basic science can unlock keys to many clinically impactful questions.

Genetic discovery allows clinicans to diagnose Mendelian disorders based on genetic tests. Early genetic diagnoses can provide individuals with life-saving foreknowledge about potentially fatal conditions. For example, in Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 (ARVC5), identifying mutation carriers is valuable; as individuals predisposed to sudden cardiac death can receive implantable cardioverter defibrillators (Hodgkinson *et al.,* 2005).

Carrier identification is also important in the medical management of Mendelian disorders such as phenylketonuria, which is treatable by dietary restriction of phenylalanine (MacLeod & Ney, 2010). Likewise, identifying affected carriers for Mendelian conditions such as WMS can help recommend patients for screening of common complications such as thoracic anuerysms (see Section 3.1). Similarly, identifying individuals who carry pathogenic mutations for hereditary cancer predisposition can greatly reduce mortality of the disease by enrollment in tailored surveillence programs.

Identifying non-carriers by genetic testing may also re-assure individuals that they do not possess a particular pathogenic mutation. Moreover, these individuals can avoid the tailored surveillance regimens of their affected family members. This practice not only reduces patient anxiety associated with these screening tests, it also reduces healthcare costs associated with unnecessary testing. Lastly, an important role for distinguishing between both carrier and non-carrier status exists for genetic counselling and family planning purposes.

The ability to find new causal genes and then develop genetic testing for individuals suspected of harboring a Mendelian disorder is largely dependent on sequencing technology. Historically, rapid advances in these DNA sequencing technologies have greatly facilitated genetic discoveries for many Mendelian disorders.

1.3 Technology Drives Genetic Discoveries

Historically, Sanger DNA sequencing and genetic mapping were critical tools for genetic discovery. For example, by 1983, genetic linkage analysis identified a critical disease interval for Huntington disease on chromosome 4 (Gusella *et al.,* 1983), after which the trinucleotide repeat expansion of the gene *HTT* was discovered (MacDonald *et al*., 1993). By 1986, a positional cloning study identified the first gene (*CYBB*) for a Mendelian disorder (i.e. x-linked chronic granulomatous disease), without any prior isolation of the affected protein (Royer-Pokora *et al.,* 1986). Later, the same method uncovered the role of *CFTR* in cystic fibrosis, which was the first gene for an autosomal recessive disorder (Riordan *et al.,* 1989). The increasing sophistication of genetic mapping and sequencing strategies led to a dramatic rise in the number of genes mapped to online databases, which grew from approximately 500 genes in early 1980 to nearly 3000 in 1993 (McKusick, 2007).

By 1990, approval of the Human Genome Project provided three billion dollars over a 15 year period to assemble a human reference genome (International Human Genome Sequencing Consortium, 2001). This project eventually led to the development of massively parallel sequencing (MPS) or 'next-generation' DNA sequencing (Heather & Chain, 2016). Next-generation sequencing is a catchall term referring to technologies that are successors to Sanger DNA sequencing. These methods capture much larger volumes of sequence data compared with Sanger sequencing. For example, coverage of entire genomes at substantial read depth is

achieved with next-generation sequencing, as compared with Sanger, which captures only short 200-1000 base pair fragements of a single locus of the genome.

Next-generation sequencing begins with capture of the target region by binding (i.e. hybridization) of probes to target regions of the genome. Target regions can include specific loci, the entire protein coding region of the genome (i.e. exome), or the entire genome. DNA in captured regions is then amplified by clonal amplification of DNA fragments, after which DNA sequencing commonly using reversible dye terminators allows for generation of vast quantities of data (Voelkerding *et al*., 2009) (Figure 1.1). Raw sequence data is then managed by variant annotation and filtering pipelines to identify novel disease-associated variants (Goh & Choi, 2012) (Figure 1.2).

Ultimately, the first personal human genome was sequenced (i.e. genome of Craig Venter), at a cost of approximately 100 million dollars (Levy *et al.,* 2007). Early on, next-generation DNA sequencing (i.e. whole exome and whole genome sequencing) was too expensive for use in research studies. Initial cost estimates for clinical genome sequencing were between 25 and 50 million dollars (Schwarze *et al.,* 2018). By 2004, the US National Human Genome Research Institute funded the 'Advanced Sequencing Technology Awards', which were a series of grants for technology companies to innovate cheaper sequencing methods, with the hope of reaching a '$1000 genome' price range (Hayden, 2014). The cost of next-generation sequencing eventually dropped to 10 million dollars per genome, and eventually to several thousand dollars (Dewey *et al*., 2012).

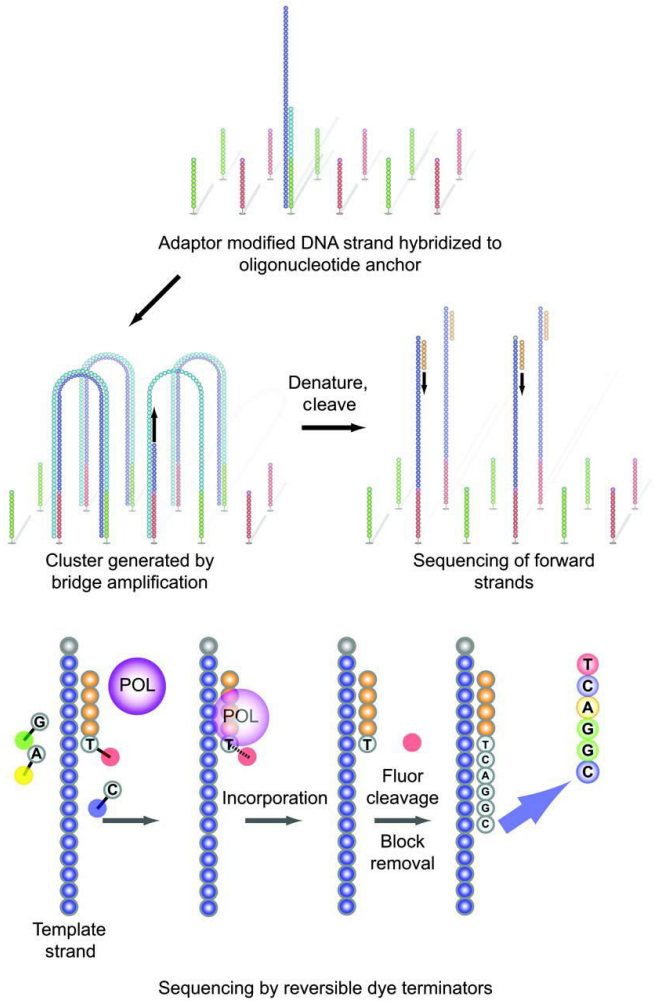


Figure 1.1 Whole Exome Sequencing Using Reversible Dye Terminators. Reprinted from Voelkerding *et al.,* (2009) with copyright permission.

This figure shows the general process of whole exome sequencing using reversible chain terminators. Adaptor molecules anchor DNA fragments to a flow cell. Each DNA fragment multiplies by bridge amplification. Reversible dye terminators are sequentially added and exposed to a fluorophore as the DNA polymerase (POL) reaction proceeds in this sequencing-by-synthesis method.

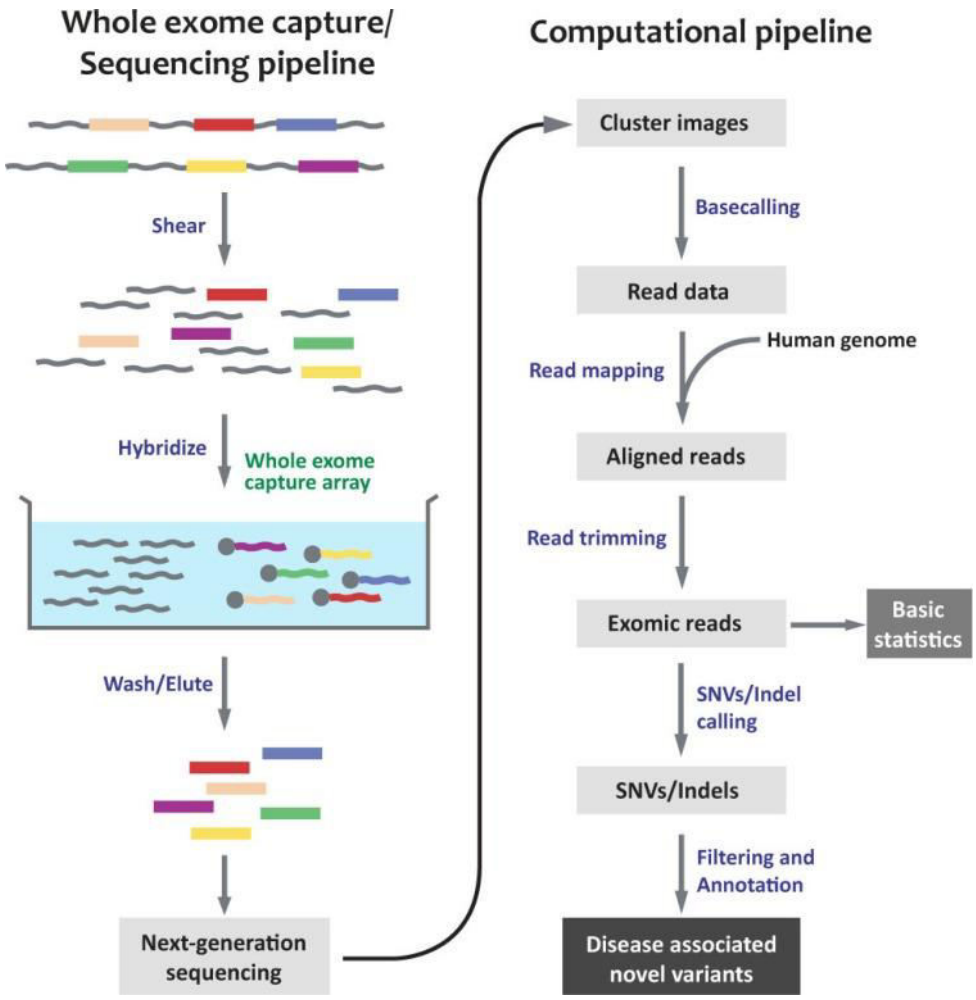


Figure 1.2 DNA Sequencing and Data Management Pipeline for Whole Exome Sequencing. Reprinted from Goh & Choi (2012) with Copyright Permission under Creative Commons Attribution Non-Commercial License.

This figure shows whole exome capture and DNA sequencing on a next-generation DNA sequencing platform, the raw data is processed through a computational pipeline to first call base pairs, then map reads to the human reference genome, and finally trim reads based on quality thresholds. Next, variants and indels are identified and annotated based on known gene function and reported allele frequencies among other metrics, before filtering variants to identify a disease-associated variant.

Given the precipitous decrease in next-generation DNA sequencing costs, Mendelian discovery studies could begin pursing whole exome sequencing by 2008. By 2010, the first study to use whole exome sequencing to identify a novel disease-causing gene (*DHODH*) for a Mendelian disorder (Miller Syndrome) was published (Ng *et al.,* 2010). In fact, another 13 novel gene discoveries for Mendelian discoveries were made in 2010 using next-generation sequencing (Rabbani *et al.,* 2012). By 2011, a further 54 new candidate genes were identified for various Mendelian disorders, and by May 2012 -- over 100 new gene discoveries were made.

It soon was clear that the dogma of genetic mapping and Sanger sequencing studies was giving way to a paradigm shift toward next-generation sequencing technology. The work of this thesis began in 2012, shortly after the first commercial (i.e. mass-produced) whole exome sequencers became accessible in Canada. An early national initiative called 'The Finding of Rare Disease Genes' (FORGE) Canada Consortium was funded to facilitate research studies into unsolved Mendelian disorders by providing the infrastructure and bioinformatic expertise necessary, which was not widely available in Canada at the time (Beaulieu *et al.,* 2014).

This broad initiative encompassed 21 genetic centers across 10 provinces of Canada, involving approximately 95 clinical geneticists and 170 other FORGE members (i.e. pediatric subspecialists, molecular biologists, genetic researchers). With the increasing availability of whole genome sequencing data, new revelations

about the frequency of rare genetic variants soon emerged. The next section discusses how next-generation sequencing has influenced our ability to determine pathogenic mutations with confidence.

1.4 Refinement of 'Disease-causing Mutations' in the Genomic Era

The Human Gene Mutation Database (HGMD) is an online catalog of all genetic variants responsible for inherited diseases as well as disease-associated or functional polymorphisms (Stenson *et al.,* 2017). At present, HMGD lists more than 224,000 germline mutation entries, including missense, nonsense, splice variants, as well as structural changes (http://www.hgmd.cf.ac.uk/ac/index.php). There are 127,200 HMGD missense and nonsense variants described, with splicing variants accounting for 20,132 entries. Meanwhile, small deletions and small insertions account for 33,090 and 13,864 entries respectively (October 2018). Following widespread next-generation DNA sequencing initiatives, it soon became clear that higher standards were needed for variant interpretation, as genomes and exomes of healthy individuals showed an excessive number of previously reportedly disease-causing mutations (Abouelhoda *et al.,* 2016; Bell *et al.,* 2011; MacArthur *et al.,* 2014; Wang and Shen, 2014).

For example, Xue and co-authors examined preliminary whole genome data from 179 individuals of the 1000 genomes pilot project (Xue *et al.,* 2012). Among the 179 healthy control genomes examined, there were 577 variants classified as 'disease-causing mutations' according to HGMD – indicating many of the healthy individuals carried alleles thought to be highly penetrant for Mendelian disorders.

There were also 191 apparently disease mutations in the homozygous state in these healthy controls. On average, between 40 and 110 disease causing mutations were carried by each individual, and 3 to 24 of these variants (per patient) were homozygous. These findings strongly suggested that numerous variants in HGMD were likely falsely reported as pathogenic mutations. This was in part due to HGMD reporting criteria, whereby a variant was determined to be disease causing based solely on what was published in the literature. It soon became clear that peer-reviewed publication were not enough to define pathogenicity, and clinicians needed a system to describe pathogenic variants in more rigid terms.

This led the American College of Medical Genetics and Genomics (ACMG) to adopt a five-tier terminology system (i.e. pathogenic, likely pathogenic, variant of uncertain significance, likely benign and benign) to categorize our varied understanding of the effects of genetic variants, with the intention of avoiding false positives in clinical care (Richards *et al.,* 2015).

Using this five-tiered framework, the strength of evidence supporting or refuting pathogenicity can be weighed. For example, Table 1.1 shows a list of criteria for assessing the pathogenicity of a variant. One of the strongest levels of evidence for pathogenicity (i.e. very strong pathogenic) occurs when there is a null variant in a gene known to cause a Mendelian disorder, and a loss of function mechanism has already been proven. It is intuitive to see how this would be very strong evidence to support pathogenicity. For the next level, that is, 'strong' support of pathogenicity, a number of pieces of evidence can be considered, including well-established *in vitro*

or animal studies that indicate damaging consequences to the gene or protein product. An alternative strong supporting criterion (PS4) occurs when the prevalence of a variant is significantly higher in cases compared to population controls.

Conversely, more 'moderate' evidence to suggest pathogenicity (i.e. PM5) would be an instance of a novel missense variant being found in a gene that has previously discovered disease causing missense variants; whereas weaker 'supporting' evidence (i.e. PP3) would be an instance where multiple *in silico* tools suggest a deleterious effect. There are separate criteria for defining variants as 'benign' also. Multiple pieces of evidence can then be used to classify a variant overall as either clinically pathogenic, likely pathogenic, uncertain significance, likely benign or benign (Table 1.2).

Although the ACMG scoring system is still rather primitive, it provides an initial point of agreement for what should be called pathogenic verses benign. Unfortunately, the ACMG guidelines do not take into account much of context and disease specific data that is present in the literature – particularly for well-studied disorders such as CRC. This led to the creation of several disease-specific expert panels which implement a more detailed ACMG-like scoring system that is more effective at determining variant pathogenicity (Thompson *et al.,* 2014).

The ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) represents the culmination of an effort to score all potential disease causing variants using ACMG

Table 1.1 ACMG guidelines for assessing pathogenicity of a candidate variant using a multi-tiered evidence-based assessment scheme. Reprinted from Richards *et al.,* (2015) with copyright permission.

Evidence of pathogenicity	Category
Very strong	PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease Caveats: • Beware of genes where LOF is not a known disease mechanism (e.g., *GFAP*, *MYH7*) • Use caution interpreting LOF variants at the extreme 3′ end of a gene • Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact • Use caution in the presence of multiple transcripts
Strong	PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change Example: Val→Leu caused by either G>C or G>T in the same codon Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level PS2 De novo (both maternity and paternity confirmed) in a patient with the disease and no family history Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, and so on, can contribute to nonmaternity. PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well established. PS4 The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls Note 1: Relative risk or OR, as obtained from case–control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance. Note 2: In instances of very rare variants where case–control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.
Moderate	PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation PM2 Absent from controls (or at extremely low frequency if recessive) (Table 6) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium Caveat: Population data for insertions/deletions may be poorly called by next-generation sequencing. PM3 For recessive disorders, detected in *trans* with a pathogenic variant Note: This requires testing of parents (or offspring) to determine phase. PM4 Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants PM5 Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before Example: Arg156His is pathogenic; now you observe Arg156Cys Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level. PM6 Assumed de novo, but without confirmation of paternity and maternity
Supporting	PP1 Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease Note: May be used as stronger evidence with increasing segregation data PP2 Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease PP3 Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.) Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of a variant. PP4 Patient's phenotype or family history is highly specific for a disease with a single genetic etiology PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation

LOF, loss of function; OR, odds ratio.

Table 1.2 ACMG guidelines for stratifying variants into categories based on evidence assessments. Reprinted with permission from Richards *et al.,* (2015).

Pathogenic	(i) 1 Very strong (PVS1) *AND* (a) ≥1 Strong (PS1–PS4) *OR* (b) ≥2 Moderate (PM1–PM6) *OR* (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) *OR* (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) *OR* (iii) 1 Strong (PS1–PS4) *AND* (a)≥3 Moderate (PM1–PM6) *OR* (b)2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR* (c)1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Likely pathogenic	(i) 1 Very strong (PVS1) *AND* 1 moderate (PM1–PM6) *OR* (ii) 1 Strong (PS1–PS4) *AND* 1–2 moderate (PM1–PM6) *OR* (iii) 1 Strong (PS1–PS4) *AND* ≥2 supporting (PP1–PP5) *OR* (iv) ≥3 Moderate (PM1–PM6) *OR* (v) 2 Moderate (PM1–PM6) *AND* ≥2 supporting (PP1–PP5) *OR* (vi) 1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Benign	(i) 1 Stand-alone (BA1) *OR* (ii) ≥2 Strong (BS1–BS4)
Likely benign	(i) 1 Strong (BS1–BS4) and 1 supporting (BP1–BP7) *OR* (ii) ≥2 Supporting (BP1–BP7)
Uncertain significance	(i) Other criteria shown above are not met OR (ii) the criteria for benign and pathogenic are contradictory

guidelines and focus on the clinically actionable using the ACMG guidelines (Landrum *et al.,* 2016). ClinVar restricts entries to assertions based on the ACMG five tiered confidence system and uses only a limited pool of recognized data submitters (i.e. clinical testing laboratories, genetics clinics and researchers). According to ClinVar, there are 74,824 high confidence 'pathogenic' Mendelian variants, and 5,504 'likely pathogenic variants'. The next section emphasizes the global burden of Mendelian disorders, and key considerations in applying special populations (i.e. founder populations) for discovery of new genetic variants.

1.5 The Burden of Mendelian Disorders

The Online Mendelian Inheritance of Man (OMIM) is a vast repository of collective knowledge about Mendelian and complex traits (Amberger *et al.,* 2015). OMIM periodically updates and synthesizes descriptions as new genes and phenotypes become available. As of October 2018, OMIM described an impressive 5,173 phenotypes and 3,548 genes causing Mendelian traits or disorders (https://www.omim.org/statistics/geneMap). Approximately 300 new Mendelian phenotypes are described annually (Chong *et al.,* 2015). Moreover, total OMIM entries have risen every month since the inception of the database in 1995. The dramatic rise in OMIM entries coincides with increasing sophistication of genetic mapping and gene discovery strategies.

Given the observation that more than 17% of an estimated 20,000 human genes are now linked to a Mendelian trait or disorder, Mendelian genetics clearly affects a wide swath of our biology, and consequently, the human condition. Many

organ systems and physiological processes are impacted by Mendelian disorders. Broadly, these include ocular disorders (Rattner *et al.,* 1999; Young, 2003), blood disorders (Sonati & Costa, 2008), connective tissue disorders (Tilstra and Byers, 1994), metabolic disorders (Dipple & McCabe, 2000), neurological disorders (Huang *et al.,* 2014), cardiovascular diseases (McBride and Garg, 2010), pulmonary disorders (Garcia and Raghu, 2004), renal diseases (Hildebrandt, 2010) and hereditary cancers (Nagy *et al.,* 2004) among others.

From both financial and emotional perspectives, there are high burdens associated with Mendelian disorders. In the United States for example, the lifetime healthcare cost for a person with a genetic disorder is approximately $5,000,000 (Chong *et al.,* 2015). Meanwhile, some families carry an unidentified gene mutation, and the uncertainty regarding if – or when – their genetic disease might strike another family member can be anxiety provoking. For other families, this problem can be exacerbated by failure to reach a clinical diagnosis in the first place. One retrospective study found that patients with mitochondrial disorders visited an average of eight clinicians before being diagnosed (Grier *et al.,* 2018). Indeed, delayed diagnosis can be a significant source of emotional trauma, known as the "diagnostic odyssey" (Basel & McCarrier, 2017).

Likewise, the demographic burden of Mendelian disorders is appreciable. Population studies demonstrate that genetic disorders occur in 5.32% of all live births by age 25 (Chong *et al.,* 2015). In a landmark study, Baird *et al*. reviewed over one million (*N*=1,169,873) consecutive live births of the British Columbia Health

Surveillance Registry. Data was collected from more than 60 provincial health registries and encompassed a 25-year period from 1952-1983 (Baird *et al.,* 1988). The authors reviewed medical records to assess burden of genetic disease in British Columbia, Canada. They found that Mendelian disorders occurred in 0.36% of live births: with autosomal dominant (0.014%), autosomal recessive (0.017%) and X-linked (0.05%) disorders each represented. Moreover, this series does not include Mendelian disorders that were lethal in the embryonic stage, nor does it include late-onset adult monogenic disorders (i.e. hereditary cancers) diagnosed after age 25 (Baird et al., 1988; Verma and Puri, 2015).

Thus, Mendelian disorders are individually rare, but with a substantial collective burden. Additionally, geographic factors considerably influence the demographic burden of Mendelian disorders. In special cases, isolated founder populations may experience higher frequencies of certain Mendelian disorders, providing ideal research conditions for genetic discoveries. Such discoveries will help to end the diagnostic odyssey for future patients, and provide answers to families.

1.6 Role of Founder Populations in Mendelian Discoveries

Genetic diversity acts as a buffer against autosomal recessive disorders. Gene flow in and out of populations are important mechanisms for maintaining genetic diversity (Amos & Harwood, 1998). Geographic and/or cultural isolation can cause limited flow of genes into a population and thereby reduce genetic diversity. The 'founder effect' is a special circumstance whereby a small group of individuals (i.e. founders)

become geographically and/or culturally isolated from their larger and well-established population – often owing to migration to a new region (Matute, 2013). The new founder population is vulnerable to genetic drift because of its relatively small size, and non-random sampling of alleles from the original population.

Colonization of island populations can lead to founder effects, as geographic barriers restrict gene flow in and out of these populations. With a reduction in genetic diversity, conditions exist for a higher prevalence of rare disorders -- especially autosomal recessive disorders. Moreover, in isolated populations, higher rates of consanguinity and inbreeding can then also contribute to genetic homogeneity and reduced genetic diversity. Higher disease frequency facilitates gene discovery studies as it potentially allows recruitment of more study participants with pathogenic mutations on a reduced genetic background due to consanguinity.

Well-recognized founder populations exist around the world and include populations of Finland (Norio *et al.,* 1973), Iceland (Kristiansson *et al.,* 2008), Quebec (Scriver, 2001), the Netherlands (Zeegers *et al.,* 2004), Sardinia (Chiò *et al.,* 2011) and NL (Rahman et al., 2003). The province of NL harbors several founder populations owing to a unique settlement history and population expansion that has resulted in a cluster of genetically distinct isolates residing in geographically restricted areas. It is important to understand the structure of this unique population in order to design genetic studies.

1.7 The Newfoundland & Labrador Founder Populations: Colonization & Population Expansion

The province of NL is part of Atlantic Canada, and consists of two geographically distinct regions: the large landmass of Labrador (294,330 km^2) on the easternmost portion of mainland Canada; and the island of Newfoundland (108,860 km^2) -- the most eastern point of North America.

Genetic evidence indicates that indigenous peoples arrived in Labrador from western Canada around 10,000 calibrated years before present (cal YBP), and subsequently migrated to Newfoundland across ice patches in the straight between Newfoundland and Labrador nearly 6,000 cal YBP (Duggan *et al.,* 2017). At least three indigenous cultural groups (Maritime Archaic, Palaeoeskimo and Beothuk peoples) inhabited NL from around 8000 to 200 cal YBP. In 1497, Giovanni Caboto (known as "John Cabot"), working by commission of King Henry VII, found and claimed Newfoundland for England (Allen, 1992). Thus began a period of prolonged European contact leading to the cultural extinction of the original indigenous groups. By the 17th century, only the Beothuk remained. Shanawdithit -- the last of the Beothuk peoples -- died in captivity in 1829 (Marshall, 1998).

While the discovery of Newfoudland occurred in 1497, colonization of the island did not begin in earnest until the 1600s. Commercial fishing in the waters surrounding NL provided lucrative economic opportunities and was a primary driving force in the settlement of the island. Early settlement was therefore seasonal, as English and Irish fishermen resided in Newfoundland during the summer and

Figure 1.3 The Canadian Province of Newfoundland & Labrador is the most eastern point in North America. Reprinted from https://commons.wikimedia.org/wiki/File:Newfoundland_and_Labrador_in_Canada_(modern).svg and licensed by Allice Hunter under the Creative Commons Attribution-Share Alike 4.0 International license.

In this figure, Canada is shaded light blue, while Newfoundland & Labrador is shaded dark blue.

returned home in the winter. The geographic distribution of most settlements around coastal parts of the island reflects this history. These small 'outport' communities were often geographically restricted from one another, with few roads, and many only accessible by boat.

During this early settlement period, the population ebbed and flowed in an unplanned manner, with very slow rates of population growth, punctuated by larger expansions driven largely by immigration primarily from Dorset and Devon, England. Subsequently, permanent settlement increased while seasonal settlement began to decline. Permanent migration to NL reached a pinnacle between 1780 and 1830 (Mannion, 1977). By 1803, the population of 19,034 would eventually expand to 220,249 in 1901 -- this time driven by immigration and natural population expansion. Immigants came from several founding populations including English, Irish, French and Spanish (Basque).

From 1901 onwards, Newfoundland consisted of numerous small outport communities separated by rugged coastlines. Within these small communities, cultural and religious factors (i.e. Protestant and Catholic denominations) played a role in mate selection, as individuals married within their denominations. Population expansion continued but with limited in-migration. The population increased to a modern day size of about 528,488. Remarkably, 95% of the present day population of NL trace their ancestry to either southeast England or southwest Ireland (Mannion, 1977) .

What we find in Newfoundland is a large number of small genetic isolates, with varying degrees of admixture and genetic diversity (Zhai *et al.,* 2016).

Additional factors provide optimal conditions for genetics research studies. First, vital statistics and marriage records from 1840 to 1949 are publically available. This, in addition to large family structure, greatly facilitates ancestry tracing and pedigree construction. The high incidence of particular Mendelian disorders and the recurrence of several founder mutations in nearby communities have played large roles in previous NL genetic discoveries.

1.8 Reviewing Current Understanding of Mendelian Genetics in the Newfoundland & Labrador Founder Populations

Extensive investigation of the genetic etiology of the NL population has demonstrated that outport communities of NL are genetically heterogeneous with numerous founder effects and higher inbreeding (Rahman et al., 2003; Zhai et al., 2016). The willingness of the population to participate in genetic research studies has provided unique insights into the clinical manifestation and natural history of numerous genetic diseases, and has played an integral role in gene discovery -- most notably for Bardet-Biedl Syndrome (BBS) (i.e. *BBS6* discovery), CRC (i.e. *MSH2* discovery) and ARVD5 (*TMEM43* discovery). The following is a review of Mendelian disorders in the province of NL.

1.8.1 Inherited Blood Disorders

One of the earliest case reports describing a genetic disorder in NL was the characterization of a large family with severe congenital deficiency of factor XIII,

which was only the second family in Canada to be reported at the time (Hanna, 1970). The prevalence of hemophilia A is also quite notable. Early on, clinicians recognized a very high incidence of X-linked factor VIII deficiency (i.e. hemophilia A), which clustered in communities on the northern shore of Newfoundland (i.e. Twillingate and New World Island). Early estimates suggested the incidence was 130 times higher than reported elsewhere in North America at the time (Bear *et al.,* 1987). The prevalence was lated identified as 44 in 3,330 males -- the highest population density of hemophilia A reported in the world (Xie *et al.,* 2002). Extensive genealogical and genetic studies of a multiplex pedigree consisting of 1,630 family members across ten generations demonstrated the founder effect, caused by the pathogenic variant pathogenic variant (p.Val2016Ala) in the *F8* gene.

Founder effects are also observed for other less common blood disorders, including for α^0-thalassaemia. There are 11 NL families carrying a large 26 kb deletion in both *HBA* genes designated α^0-thalassaemia ($--^{BRIT}$) common in the British population; as well as another founder mutation (90.7 kb deletion) designated α^0-thalassaemia ($--^{NFLD}$) (Eng *et al.,* 2009).

1.8.2 Hereditary Ocular Disorders

Studies by Green and colleagues have established a high prevalence of monogenic ocular disorders in NL based on data from ocular genetics clinics in St. John's and rural NL as well as by review of CNIB records and extensive pedigree analysis of famlilies (Green *et al.,* 1986; Green & Johnson, 1983). Strikingly, monogenic disorders accounted for 30% of all provincially registered CNIB cases in 1986. The

largest proportion included RP or other rod-cone dystrophies, accounting for 31% of all ocular genetic disorders. Additionally, cone, cone-rod or macular dystrophies accounted for 22% of cases. Ocular albinism and oculocutaneous albinism each contributed an additional 9% of genetically determined CNIB cases, while congenital cataract disorders accounted for a further 6% among other lesser causes.

While a comprehensive review of NL RP mutations has not been published, there are reports of a notable rod-cone dystrophy due to *RLBP1* mutations called Newfoundland Rod-Cone Dystrophy (OMIM #607476). While *RLBP1* mutations are typically a very rare cause of rod-cone dystrophies worldwide, there are six NL families carrying homozygous or compound heterozygous combinations of two pathogenic splice variants in *RLBP1* (Eichers *et al.,* 2002). Subsequently, four other *RLBP1* mutations have been identified in a further four NL families (Green J, unpublished data), indicating significant diversity of mutations for this rare rod-cone dystrophy.

Meanwhile, among inherited macular dystrophies, Stargardt disease is a juvenile macular degeneration prominently seen in NL -- with 31 families identified since the 1970s. A founder mutation in *ABCA4* (c.5714+5G>A) is the most common cause among 21 *ABCA4* mutations identified in NL (Green *et al.,* 2010). Likewise, since the 1970s, X-linked ocular albinism has been prevalent along the central and northern coast and in central Newfoundland. Genealogical investigation of a six-generation pedigree of 200 family members identified a founder mutation with a common ancestor living in Moreton's Harbor ~178 years ago (Johnson *et al.,* 1971).

Subsequent genetic studies of NL families were integral to early genetic mapping studies, which identified a critical interval for X-linked ocular albinism on chromosome Xp22.3 (Charles *et al.,* 1993; Pearce *et al.,* 1971), which facilitated a discovery implicating *GPR143* in this disorder (Bassi *et al*., 1995).

Finally, a number of cases of congenital iris atrophy (*N*=40), previously coined 'peninsula pupil' were identified on the Great Northern Peninsula of Newfoundland (Bosanquet and Johnson, 1981). Also, a large NL family with anterior segment disorders including Peter's Anomaly is described (Green & Johnson, 1986), with subsequent genetic studies identifying a pathogenic variant in *FOXE3* (Doucette *et al.,* 2011). Founder effects causing achromatopsia are also seen in families with pathogenic *CNGB3* (p.T383IfsX) and *CNGA3* (p.L527R) variants, and a family with Jalili syndrome (OMIM #217080) -- an extremely rare cone-rod dystrophy associated with amelogenesis imperfecta was also identified (Doucette *et al.,* 2013).

1.8.3 Bardet-Biedl Syndrome

There is a significantly high incidence of BBS in NL (1 in 17,500) (Moore *et al.,* 2005). Clinical characterization of this autosomal recessive syndromic retinopathy among NL BBS families has made significant contributions to our understanding of the cardinal manifestations of BBS (Green *et al.,* 1989). Furthermore, analysis of 46 BBS patients from 26 NL families provided strong evidence that BBS and Laurence-Moon Syndrome were distinct entities (Moore *et al.,* 2005). Significant genetic heterogeneity has been documented among NL BBS families (Figure 1.4). Consequently, NL founder families have played a significant role in early genetic

studies mapping the BBS loci. These include studies investigating the *BBS1* locus (Young *et al.,* 1999; Fan *et al.,* 2004); the *BBS3* locus with subsequent discovery of *ARL6* mutations (Young *et al.,* 1998; Fan *et al.,* 2004b); the *BBS5* locus and discovery of the *BBS5* gene (Woods *et al.,* 1999; Young et al., 1999; Li *et al*., 2004); and mapping the *BBS6* locus with identification of *MKKS* as the first gene identified for BBS (Katsanis *et al.,* 1999). Moore and colleagues summarize clinical and genetic findings of NL founder BBS mutations in six large families homozygous for p.M390R in *BBS1*, three families homozygous for p.F94fsX103 in *MKKS* and two families homozygous for p.D143fsX157 in *MKKS*, among several other loci (Moore *et al.,* 2005) (Green J, unpublished data).

1.8.4 Connective Tissue and Muscular Disorders

There is a high prevalence of certain muscular and connective tissue disorders in NL. For example, a population-based study from 2006 to 2011 identified 28 families (36% of total cases) with family history of pulmonary fibrosis (Fernandez *et al.,* 2012). This relatively high proportion of familial pulmonary fibrosis (compared to only 0.5-20% of familial cases reported in other populations) clustered in communities along the eastern region of Newfoundland (i.e. Avalon Peninsula) (Figure 1.5). Moreover, Fernandez *et al*. (2012) identified several pathogenic *TERT* mutations (p.Phe883Cys, p.Arg631Gln and p.Arg865His) in 3/28 families (10.7%).

In terms of muscular disorders, a very high prevalence of myotonic dystrophy has been seen in Labrador. For example, in 1978, the prevalence was 75 per 100,000 – which was approximately 13x higher than reported elsewhere

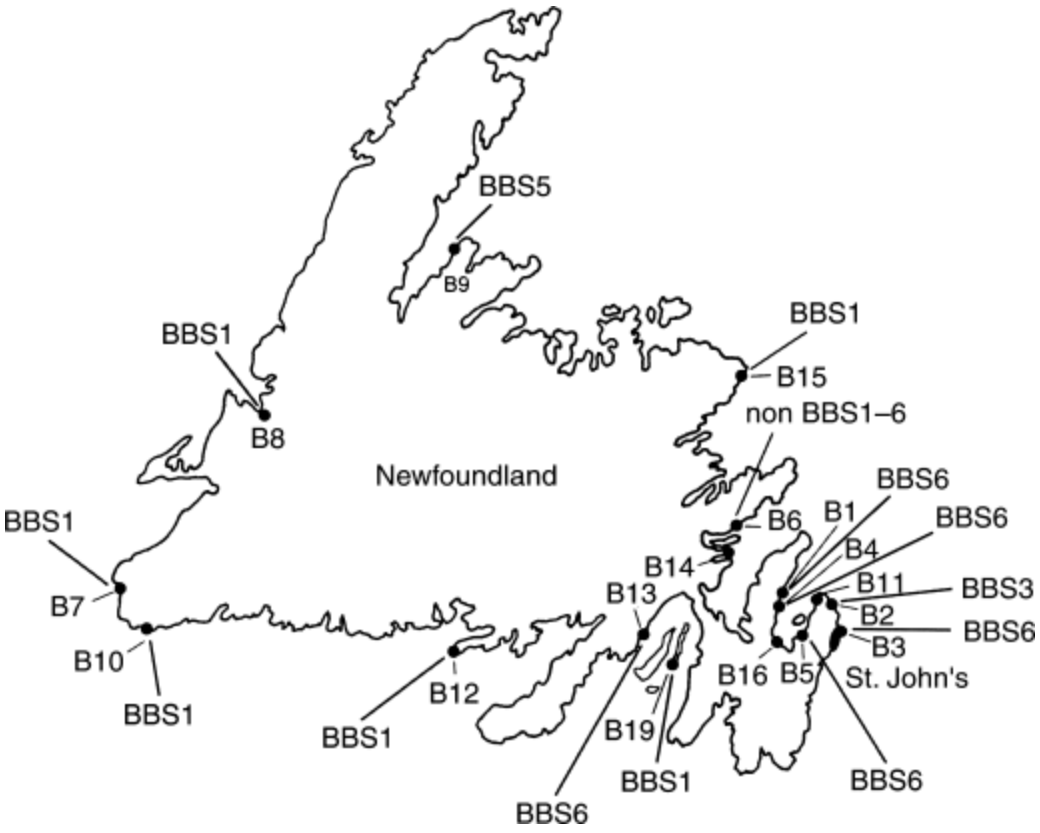


Figure 1.4 The geographic distribution of BBS families and their respective haplotypes, around several communities in Newfoundland. Reprinted from Parfrey *et al.,* (2002) with copyright permission.

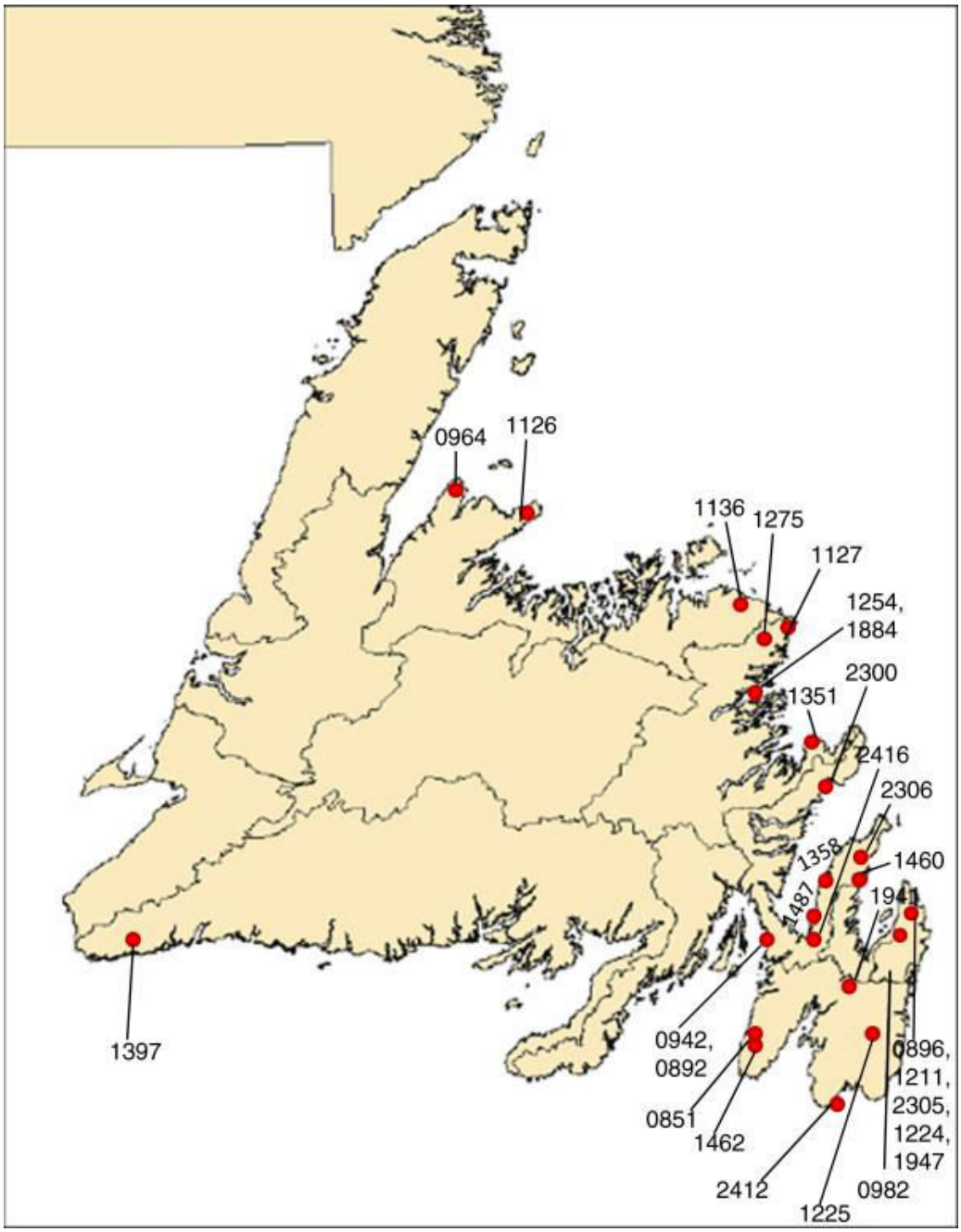


Figure 1.5 The geographic distribution of FPF study families around Newfoundland. Reprinted from Fernandez *et al.,* (2012) with copyright permission under creative commons license.

(Webb *et al.,* 1978). Meanwhile, on the west coast of Newfoundland, one family with an exceptionally rare skeletal and connective tissue disorder, known as WMS, was reported (Johnson & Bosanquet, 1983). This family is the subject of a genetic investigation of this thesis.

1.8.5 Cardiovascular disorders

Studies of NL families have yielded significant insights into arrhythmogenic right ventricular dysplasia type 5 (ARVD5). For example, one large NL ARVD5 family was first described in the late 1980s (Marshall *et al.,* 1988). Investigation of this extended pedigree of more than 200 individuals with ten alive and affected at this time identified an ARVD5 locus on chromosome 3p25 (Ahmad *et al.,* 1998). Subsequently, Merner and colleagues fine mapped this locus in 14 additional NL ARVD5 families, identifying a 2.36 Mb critical interval. Thereafter, they discovered the causal role of *TMEM43* in ARVD5 (Merner *et al.,* 2008). This *TMEM43* founder mutation (p.S358L) was found in all 15 NL ARVD5 families and has provided extensive clinical data, characterizing the natural history of p.S358L within the NL population (Hodgkinson *et al.,* 2016, Hodgkinson *et al.,* 2013). The *TMEM43* p.S358L mutation has been independently replicated other places nationally and globally, including in Canada (Baskin *et al.,* 2013) and Denmark (Christensen *et al.,* 2011). Moreover, a common haplotype has been identified between some Danish, German and NL ARVD5 families, suggesting ancestral origins from within Europe (Milting *et al.,* 2015).

Finally, in terms of other hereditary cardiovascular disorders, familial

hyperlipoproteinemia has been reported in two large families with at least 39 affected individuals clustering near Baie Verte, on the north coast of NL (Anjilvel, 1973).

1.8.6 Neurological Disorders

Studies indicate extensive genetic heterogeneity and founder effects are present among neurological disorders in NL. There is a notably high prevalence (13.4 per 100,000) of neuronal ceroid lipofuscinosis (NCL) in NL, compared with relatively low prevalence (0.15 to 0.78 per 100,000) seen in other populations (Moore *et al.,* 2008). Moreover, if considering the late infantile NCL subtype, NL reports the highest prevalence (9.0 per 100,000) in the world. Most of these families have been followed since the 1980s (Andermann *et al.,* 1988), and characterization of their clinical manifestations has provided substantial insights into NCL.

Most NCL families cluster around the northern Avalon Peninsula or the southern coast of Newfoundland (Figure 1.6). Substantial genetic heterogeneity and numerous NCL founder mutations have been identified in the NL population. For example, Moore *et al.,* (2008) reviewed all NCL cases ascertained by adult and pediatric neurology clinics between 1960 and 2005. Among 54 patients from 32 families, pathogenic mutations were found in *TTP1* (*CLN2*), *CLN3*, *DNAJC5* (*CLN4*), *CLN5* and *CLN6*. A number of founder mutations were identified in *TTP1*, as two families were homozygous for p.Ser475TrpfsX13, five families were homozygous for p.Gly284Val and four families were homozygous for a splice variant (c.509-1G>C). Moreover, as many as seven additional families harbored compound heterozygous

alleles of the above, or other nonsense variants; including p.Lys104X and p.Ala208X. Finally, in *CLN6*, there were five families homozygous for p.Val91GlufsX42, with *CLN3* and *DNAJC5* causing a smaller proportion of cases – and highlighting the extensive genetic heterogeneity in the province.

In terms of other neurological disorders within NL, the rate of pediatric idiopathic epilepsy (107 per 100,000 individuals) is approximately 3x higher than other developed countries (Mahoney *et al.,* 2012). Among 88 NL families with pedigree data available, family history was present in many (n=55), with eight families demonstrating clear autosomal dominant inheritance. To date, the genetic etiology of only one family has been solved, with identification of a pathogenic mutation (p.Y388H) in *SCN1A* (Mahoney *et al.,* 2009).

Meanwhile, among hereditary spastic ataxias, NL families have significantly contributed to our understanding of these disorders. The first locus identified for an autosomal dominant spastic ataxia was the 'SAX1' locus on chromosome 12p13, which was found in 3 large NL families (Meijer *et al.,* 2002). Later, these families were further clinically characterized and the locus fine mapped to a 1.9 Mb region (Grewal *et al.,* 2004). Subsequently, candidate gene sequencing discovered *VAMP1* mutations in autosomal dominant spastic ataxia (Bourassa *et al.,* 2012). The causal pathogenic mutation (c.340+2T>G) segregated in 50 affected individuals from four multiplex families living on the northern Avalon Peninsula of Newfoundland – demonstrating a founder effect.

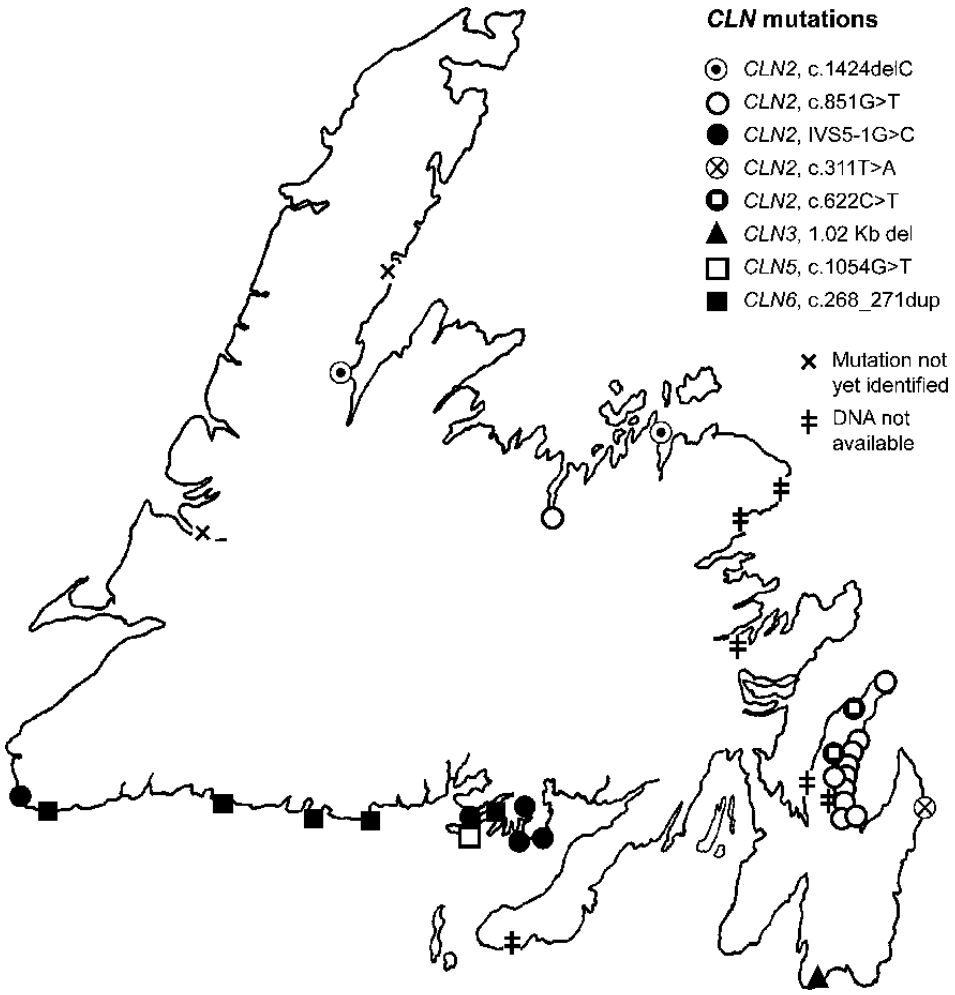


Figure 1.6 The geographic distribution of NCL families in Newfoundland stratified by mutation status. Reprinted from Moore *et al.,* (2008) with copyright permission.

Finally, investigation of NL families with hereditary sensory and autonomic neuropathy type II has yielded additional founder effects. This was demonstrated by mapping and identification of homozygous *WNK1* c.594delA mutations in NL families (Lafreniere *et al.,* 2004). Additionally, linkage analysis and homozygosity mapping in a family with congenital lack of pain sensation identified a nonsense mutation (p.Y328*) in *SCN9A* (Ahmad *et al.,* 2007).

1.8.7 Inherited Renal Disorders

The incidence of monogenic renal disorders has been extensively studied in the NL population (O'Dea *et al.,* 1998). Parfrey *et al*. (2002) comprehensively reviewed the clinical and epidemiological distribution of monogenic and complex kidney diseases across NL. Within a cohort of 669 patients who developed end stage renal disease between 1987 and 1993, they found that 8.5% of cases resulted from a monogenic disorder. Among the autosomal dominant disorders, the most common disorders were polycystic kidney disease (PKD) (4.5% of cases), Alport Syndrome (2.7% of cases) and Charcot-Marie Tooth syndrome (0.3% of cases). Meanwhile, among autosomal recessive disorders there were BBS (0.3% of cases); 2,8 Dihydroxyadenine stone disease (0.3% of cases); hyperoxaluria (0.3% of cases); and autosomal recessive PKD (0.2% of cases).

PKD is the most prominently studied inherited renal disorder in NL, with at least 17 families with autosomal dominant PKD (Figure 1.7) (Parfrey *et al.,* 1990). Many of these families have been studied since the 1980s (Churchill *et al.,* 1984). For example, a founder effect at the *PKD1* locus was identified in 91 affected individuals

of four NL families living on the south coast of Newfoundland (Parfrey *et al.,* 2002). Meanwhile, at least four other *PKD1* mutations account for a further 54 affected individuals, from another four families -- demonstrating genetic heterogeneity for *PKD1* in NL. Likewise, at least two families share a founder mutation in *PKD2* (c.1510del127), with another two families segregating different *PKD2* mutations. Finally, linkage and haplotype analyses have demonstrated at least one NL family that appears to have trans-heterozygotes for *PKD1* and *PKD2* loci, suggesting bi-lineal inheritance (Pei *et al.,* 2001).

1.8.8 Hereditary Cancers

Hereditary cancer syndromes have important implications in NL. The genetics of CRC is comprehensively reviewed in Section 4.1, while NL population-related CRC studies are briefly described here. NL has the highest incidence of CRC in Canada (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2015). Moreover, Green and colleagues observed the highest reported familial incidence of CRC in the world (Green *et al.,* 2007).

In terms of Familial Adenomatous Polyposis (FAP), a study of 90 individuals from 5 families identified a founder mutation in *APC* originating from Twillingate island, NL (Spirio *et al.,* 1999). Meanwhile, investigation of Hereditary Non-polyposis CRC (HNPCC) in NL families has yielded several important discoveries. Primarily, the investigation of an NL family dubbed 'Family C' was integral to the discovery of *MSH2*, the first gene identified for HNPCC (Fishel *et al.,* 1993; Leach *et al.,* 1993), and a common ancestor has been identified for many of these families. The founder

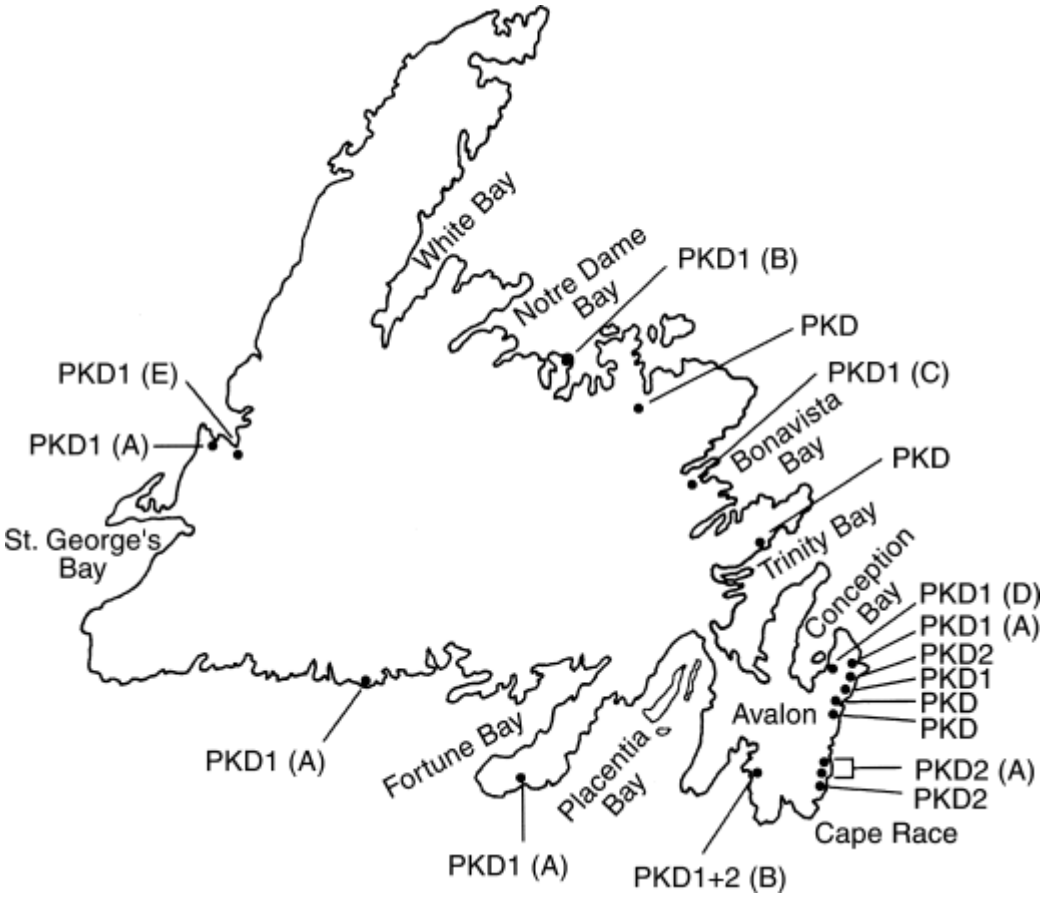


Figure 1.7 The geographic distribution of PKD families in Newfoundland with their respective haplotypes. Reprinted from Parfrey *et al.,* (2002) with copyright permission.

mutation in Family C (c.942+3A>T) is found in 21 NL families with eight families known to share a common haplotype (Froggatt *et al.,* 1999). The same mutation has also been found in the United Kingdom and the United States (Lynch *et al.,* 2004). Anadditional *MSH2* founder mutation, an exon 8 deletion, was identified in a number of NL families as well as another large family harboring a deletion of exons 4-16 (Green J, personal communication). The clinical sequalae of these three *MSH2* mutatations have been extensively characterized (Stuckless *et al*., 2007).

Population-based studies genetically screening NL patients have yielded important insights as well. For example, Woods *et al*. (2010) comprehensively screened all incident CRC cases (N=750 CRC patients, N=708 families) in the province who were ascertained over 5-year period from 1999 to 2003. This genetic screening identified 12 mutations in known CRC genes. An additional 0.9% of cases were ascribed to *APC* or biallelic *MUTYH* mutations. Critically, this study identified several Familial CRC Type X (FCCTX) families, or families with strong genetic predisposition of unknown etiology -- which remain a particularly important avenue for ongoing genetic research studies. The geographic distribution and common HNPCC with MMR mutations (now called Lynch Syndrome) in NL families are shown in Figure 1.8. Meanwhile, the geographic distribution of FCCTX families is shown in Figure 1.9.

Finally, in terms of other hereditary cancer syndromes, the incidence of gastric cancer is high in NL as well -- which has been appreciated since the 1970s (Pfeiffer *et al.,* 1973). The high incidence is partly explained by a recurrent *CDH1*

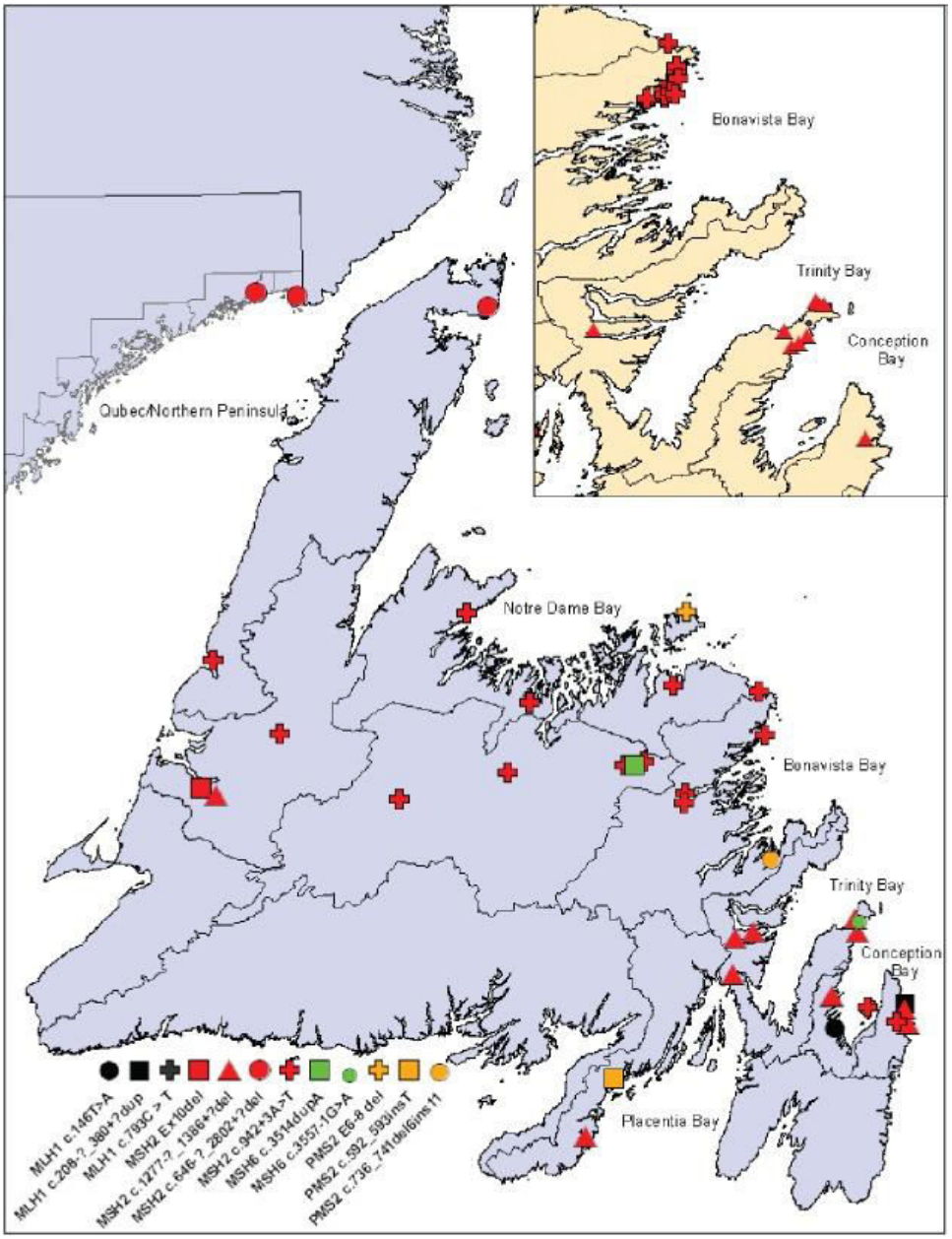


Figure 1.8 The geographic distribution of families with known Lynch Syndrome mutations in Newfoundland. Reprinted from Warden *et al.,* (2013) with copyright permission.

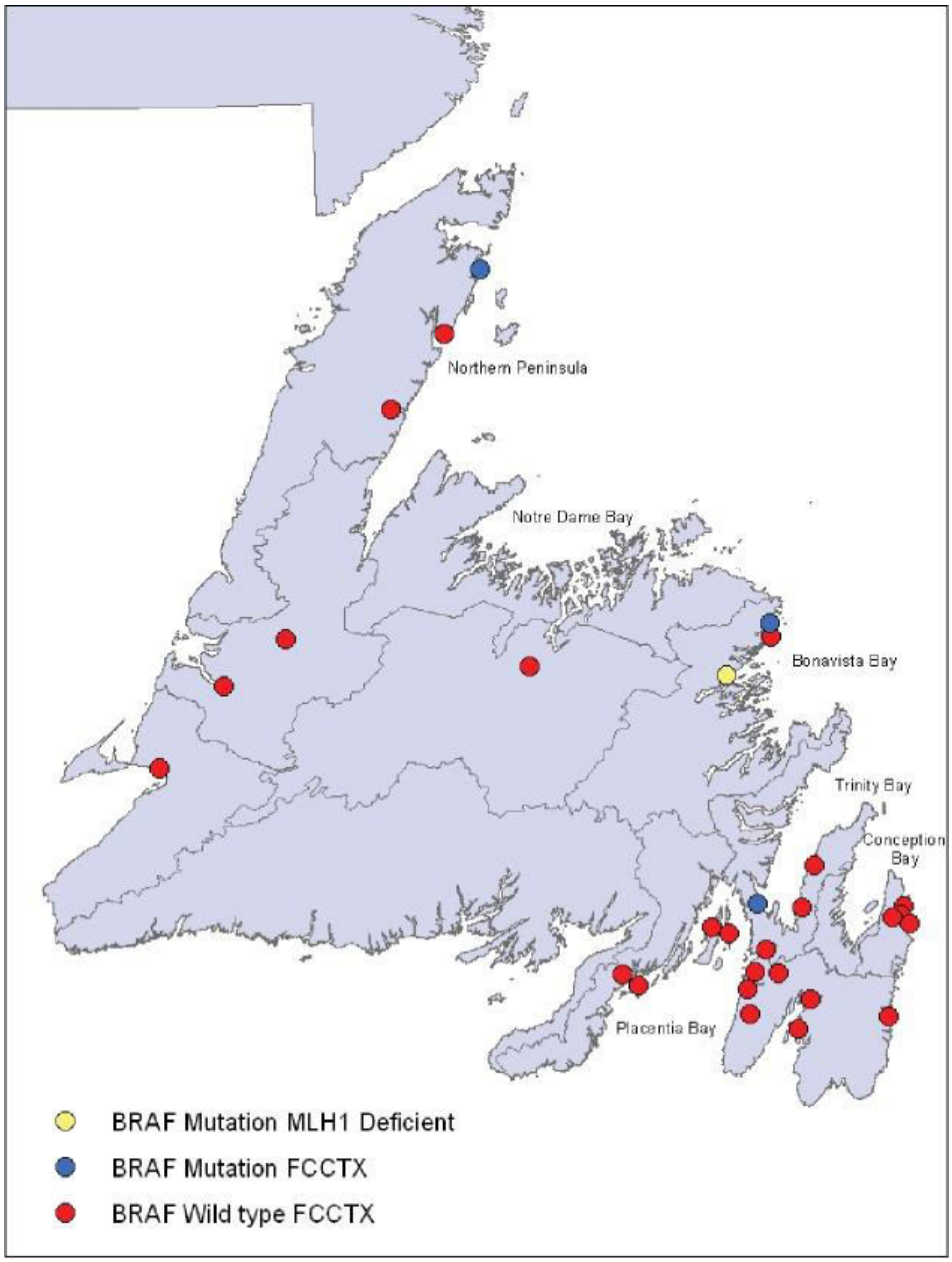


Figure 1.9 The geographic distribution of FCCTX families in Newfoundland. Reprinted from Warden *et al.,* (2013) with copyright permission.

founder mutation (c.2398delC) found in 23 gastric cancer and 16 breast cancer cases belonging to four families residing near the southeast coast and Avalon Peninsula (Kaurah *et al.,* 2007). Lastly, Von Hippel-Lindau disease has been identified in at least one large family from NL (Green *et al.,* 1986b).

1.8.9 Unsolved Mendelian Disorders in NL

Taken together, this broad overview of NL Mendelian genetics highlights the extensive genetic heterogeneity in the presence of multiple founder effects in NL. The willingness of this population to participate in genetic research studies has provided unique insights into the clinical manifestation and natural history of numerous rare diseases over the years, and has played an integral role in gene discovery; most notably for BBS (i.e. *BBS6*), CRC (i.e. *MSH2*) and ARVD5 (*TMEM43*). Studying NL families with unsolved genetic etiology is poised to yield new discoveries for disorders including RP, WMS, CRC/FCCTX and FPF; suggesting new insights are on the horizon.

1.9 Thesis Aims

The thesis aim was to explore three Mendelian disorders (arRP, WMS, CRC) in families of the NL founder populations by applying modern genetic strategies for the discovery of novel genes and/or genetic variants underying each disorder. Secondary aims were to characterize the clinical manifestation of these rare disorders, as well as the genetic variants uncovered. The purpose of this endeavor was to provide the families, and potentially others around the world, with a genetic

diagnosis, while gaining unique insights into the underlying molecular pathogenesis of these disorders.

Three main methods were employed including a.) family-based studies using whole exome DNA sequencing with genetic linkage analysis or homozygosity mapping (arRP and WMS studies, respectively), and b.) cohort-based candidate gene approach using Sanger sequencing (CRC project). Following identification of candidate variants, the next objective was to assess and interpret variants for pathogenicity using control DNA sequencing, segregation analysis, bioinformatic prediction and functional assays where appropriate. Finally, the last objective was to synthesize the clinical and molecular genetic data in order to provide meaningful insights into these Mendelian disorders.

1.10 Co-authorship Statement

The following three chapters are independent and published research articles, written by me as the primary author. As a major contributor, I participated in data generation, interpretation and critical appraisal of data, and I made significant intellectual contributions to each project. I am grateful to a number of contributors and collaborators who provided their expertise during each of these projects. Individual contributions made by each co-author are described in the author contributions section of Chapters 2-4.

Chapter 2. Discovery of a Novel Mutation for Retinitis Pigmentosa in a Large Kindred from Newfoundland Using Whole Exome Sequencing and Genetic Linkage Analysis.

2.1 Background: Retinitis Pigmentosa

During the visual process, light is detected within the retina, which is a complex tissue consisting of ten histological layers (Joselevitch, 2008) (Figure 2.1). Photons reach the inner limiting membrane and travel toward the outermost layer of the retina called the pigment epithelium (RPE). The photoreceptor layer lies above the RPE, and both play critical roles in the visual cycle. The photoreceptor cells (i.e. rods and cones) are highly specialized neuroepithelial cells within the photoreceptor layer that detect incoming photons, and there are more than 120 million rods and six million cones within the retina. Whereas rods are highly sensitive to photons (i.e. providing vision in low light conditions), the cones are less sensitive, but able to perceive color vision. The photo transduction cascade is the physiological process of vision that is responsible for detecting incoming photons (Arshavsky *et al.,* 2002).

Photoreceptor cells are structurally unique (Figure 2.2). These cells have outer segments, which contain layers of flattened disc-like membranes studded with a photopigment called rhodopsin. There are approximately 700-1000 of these discs per photoreceptor cell (Mayhew & Astle, 1997). Moreover, approximately 90% of the protein content in each disc is rhodopsin (Hargrave, 2001), which equates to approximately 150,000 rhodopsin molecules per disc, and 1.5 million molecules per

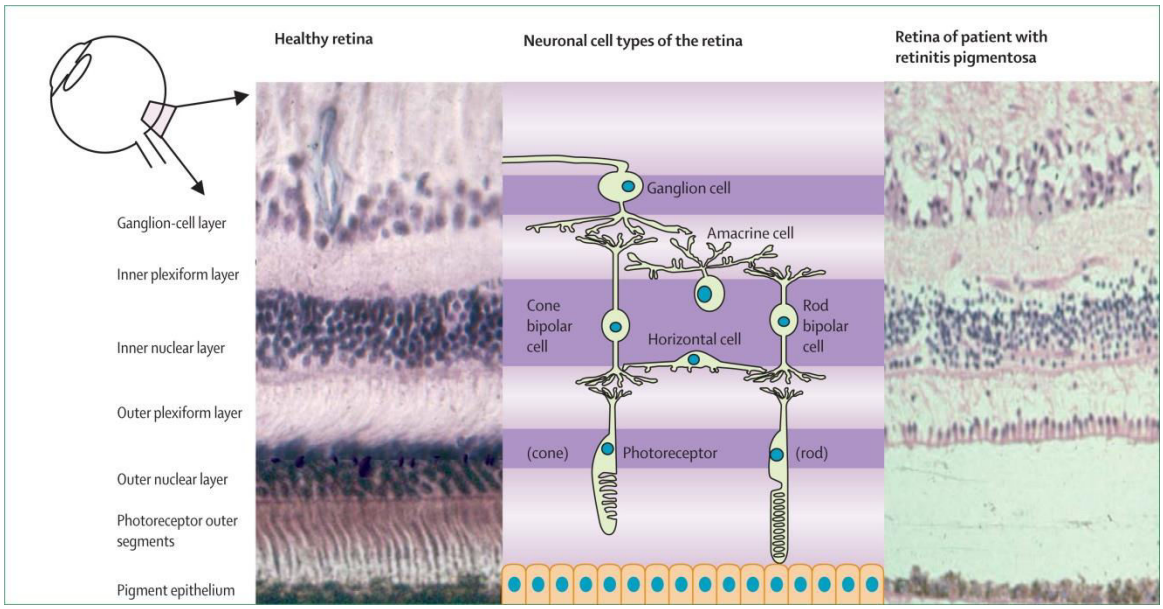


Figure 2.1 The histological and cellular structure of a healthy retina compared to changes seen in RP. Reprinted from Hartong *et al.,* (2006) with copyright permission.

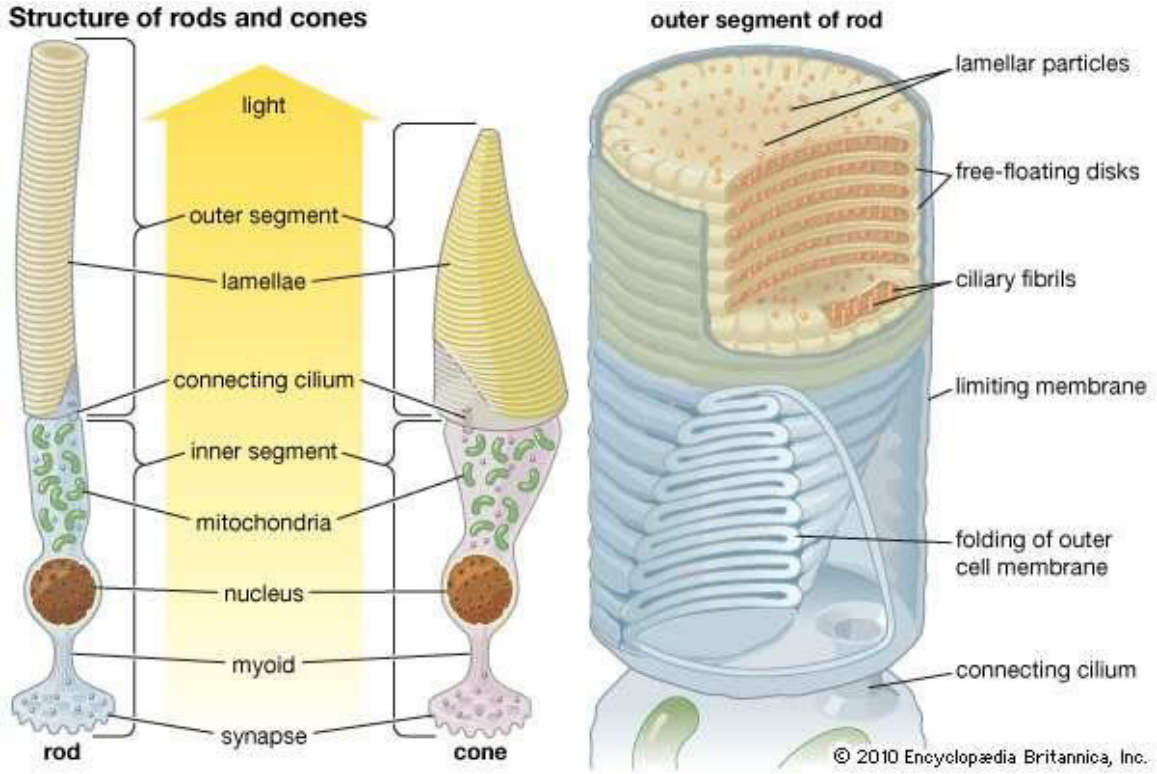


Figure 2.2 Structure of the retinal photoreceptors, demonstrating the membranous discs of the photoreceptor outer segments. Courtesy of Encyclopaedia Britannica, Inc., copyright 2012; used with permission.

retina. Rhodopsin consists of opsin and covalently bound 11-cis-retinal, which is a chromophore that is central to the photo transduction cascade.

During the photo transduction cascade, photons travel to the photoreceptor layer and are absorbed by rhodopsin. This results in a conformational change of the 11-cis-retinal chromophore, which isomerizes to all-trans retinal – creating conformational changes to the opsin molecule. This initiates a series of G-protein coupled signalling events leading to the photo transduction cascade. A conformational change of rhodopsin to metarhodopsin II activates a G-protein called transducin (Shichida & Morizumi, 2007). The exchange of guanosine triphosphate to the metarhodopsin-activated transducin α-subunit causes dissociation of the subunit from the transducin β and γ subunits. The activated transducin α-subunit then proceeds to activate cGMP phosphodiesterase (Fung *et al.,* 1981). Normally, under dark conditions, cGMP molecules are bound to Na^+ and Ca^{2+} channels on the disc membrane, which keep the channels open and the cells depolarized. Activation of cGMP phosphodiesterase during the photo transduction cascade hydrolyzes cGMG causing these channels to close and photoreceptors to hyperpolarize. The now-activated photoreceptor cell then transmits stimulatory signals to the inner retinal membrane via bipolar cells, which relay signals to the optic nerve via retinal ganglion cells for visual perception (Euler *et al.,* 2014).

Ultimately, the isomerized all-trans retinal must be recycled back to 11-cis retinal in order for a new photon to re-initiate the cascade, while also preventing accumulation of phototoxic products. To accomplish this, a 'bleach and recycling'

pathway acts to convert all-trans retinal back to 11-cis retinal by a series of enzymatic reactions that occur within photoreceptor outer segments and the RPE (Figure 2.3). This complex and multi-step pathway demonstrates the important roles of the photoreceptor outer segments and the RPE. Communication between these two layers is critical, and the complex interplay between them is essential for maintaining healthy and properly functioning photoreceptors.

The RPE plays a critical role in maintaining photoreceptor excitability and viability (Strauss, 2005). Several mechanisms contribute to this. First, secretion is an important process of the RPE that facilitates communication between photoreceptors and blood vessels. The RPE secretes a number of growth factors and other important molecules, including fibroblast growth factors, insulin-like growth factors, vascular endothelial growth factor and pigment epithelium derived factor (PEDF). PEDF is secreted from the apical surface of the RPE toward the photoreceptor cells. PEDF is a neurotrophic factor which prevents apoptosis and stabilizes photoreceptor cells (Cayouette *et al.,* 1999). Furthermore, as previously mentioned, the photoreceptors are constantly exposed to phototoxic products and free radicals, which accumulate and cause damage in the absense of a properly functioning retina. The RPE also plays a critical role in light absorption, which improves image quality by absorbing scattered light. Melanosomes within the RPE are dynamic molecules which can migrate to the apical surface to absorb more photons under light conditions (Strauss, 2005). Likewise, Na^+/K^+-ATPases are important mediators of K+ transport to photoreceptors and glial cells. As well, the

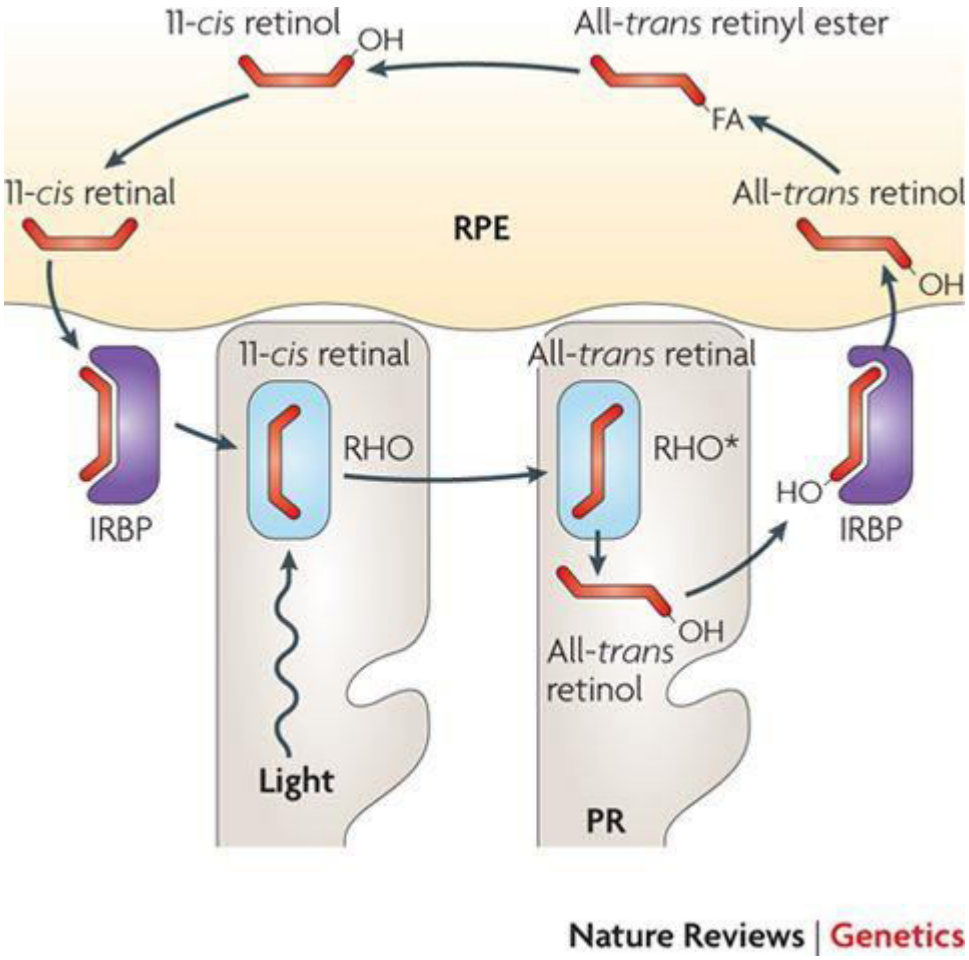


Figure 2.3 The bleach and recycling pathway restores 11-cis retinal for use in the photoreceptor transduction cascade. Reprinted from (Wright *et al.,* 2010) with Copyright Permission.

epithelial transport of water, ions, and retinal precursors such as vitamin A across the sub-retinal space is important in maintaining a healthy retina.

Lastly, phagocytosis is a key process performed by the RPE. Given the toxic accumulation of phototoxic products during the visual cycle, the photoreceptor outer segments must undergo continuous renewal – and are therefore shed into the sub-retinal space for further breakdown. Once shed, the outer segments traverse the subretinal space and must be phagocytosed by the RPE. Meanwhile, new discs are formed at the proximal end (near the cell body), as older discs move distally toward the microvilli of the RPE. Phagocytosis occurs at the apical membranes of the RPE and involves three key protein mediators: CD36, MERTK and $\alpha_V\beta_5$ integrin. Whereas CD36 mediates internalization of photoreceptor outer segments, MERTK plays a critical role in signal transduction to activate phagocytosis. Meanwhile, $\alpha_V\beta_5$ integrin mediates the binding process. The importance of the MERTK receptor in phagocytosis is notable from studies of the Royal College of Surgeons (RCS) rat, which was the first animal model for inherited retinal degeneration (D'Cruz *et al.,* 2000). By ensuring adequate activation and signal transduction during phagocytosis, the photoreceptor outer segments can be cleared from the sub retinal space. Failure to phagocytize the outer segments causes a build-up of debris in the sub-retinal space, which hinders the flow of nutrients and leads to photoreceptor degeneration.

Photoreceptor degeneration has been extensively studied in humans, given that blindness can substantially impair one's quality of life. During the normal aging process, retinal photoreceptor density decreases by about 0.2% to 0.4% per year

(Panda-Jonas *et al.,* 1995), which leads to a corresponding decrease in visual acuity. Inherited retinal dystrophies (IRDs) however, cause early vision loss and blindness. The IRDs constitute a large heterogeneous group of retinal disorders, united by their progressive degeneration of photoreceptors, variable clinical manifestations and monogenic patterns of inheritance. The more than 300 genes or loci known to cause IRDs, highlights the substantial complexity of the visual cycle; and the critical role that photoreceptors play in this process.

A broad array of molecular pathways has been implicated in hereditary retinal degeneration (Figure 2.4). While some mutations occur in genes in obvious pathways such as the visual cycle, a broad variety of processes impacted include more generalized cellular functions such as protein folding, extracellular matrix regulation and lipid metabolism (Wright *et al.,* 2010). In fact, the pathway accounting for the largest proportion of IRDs involves ciliary trafficking. Meanwhile, the clinical pattern of visual problems associated with IRDs, along with electrophysiological findings can broadly sub-classify IRDs into cone-rod dystrophies, rod-cone dystrophies and inherited macular dystrophies.

In rod-cone dystrophies, there is primarily a loss of rod photoreceptors followed by loss of cone photoreceptors. As rods are concentrated in the mid-periphery, visual field loss is often first noticed in the periphery, with sparing of the cone-rich macula, responsible for central vision, until the later stages. As rods are involved in low light responses, early onset of night-blindness occurs in rod-cone

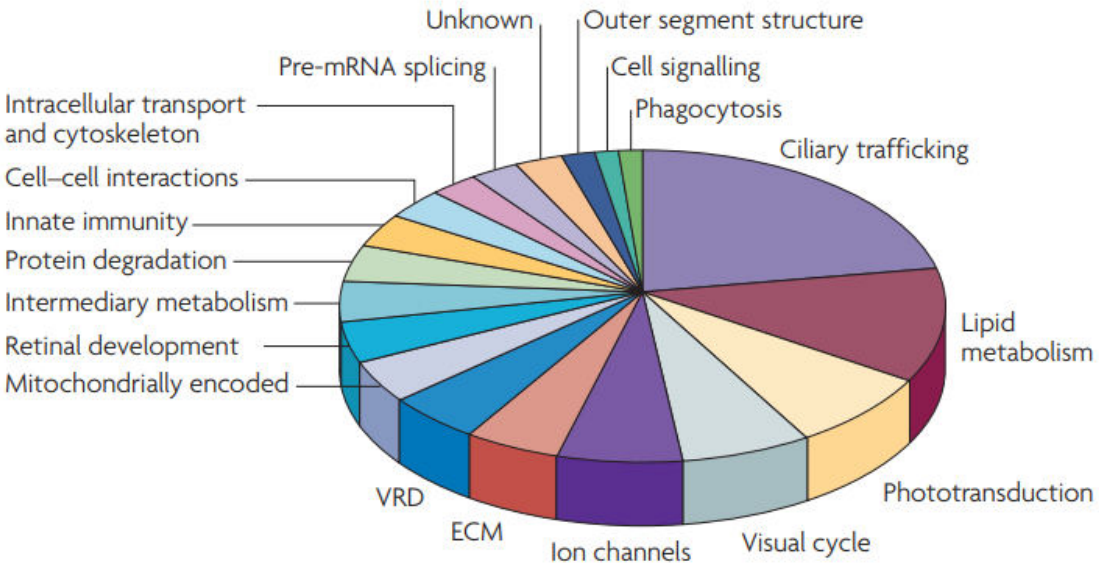


Figure 2.4 The relative proportion of known genes causing photoreceptor death and their respective molecular pathways. Reprinted from Wright *et al.,* (2010) with copyright permission.

dystrophies. Color vision and central visual fields and acuities may then be affected later on, as the disease progresses.

RP is the most common IRD, and is essentially a rod-cone dystrophy. There are more than 1.5 million people worldwide living with RP. The progressive decrease in visual fields and acuities that accompany it and eventual blindness are a significant and incurable cause of disability. The global prevalence of RP is approximately 1:4000 (Pagon, 1988). There is also geographic variability in RP, with one study in India estimating RP prevalence of 1:750 (Nangia *et al.,* 2012). RP is a broad and heterogeneous disorder, united by a triad of clinical features: waxy pallor of the optic disc, attenuation of retinal blood vessels, and pigmentary aggregates visualized in the fundus, called bone spicules (Figure 2.5).

Night blindness and loss of visual fields may occur early (typically by the second decade) with blindness probably as early as the 40th decade. Based on accompanying clinical features, RP is sub-divided into syndromic and non-syndromic forms. Syndromic RP includes conditions with extra-ocular manifestations. The syndromic forms of RP are well characterized, and include Usher Syndrome, Bassen-Kornzweig Syndrome, Refsum disease, Bardet-Biedl Syndrome, and Batten disease. Meanwhile non-syndromic RP involves only ocular features. The genetic and molecular pathogenesis of RP is broad and explained by a large number of genes resulting in autosomal dominant, autosomal recessive and X-linked inheritance patterns (Figure 2.6). X-linked RP is the least common form of RP (5-15% of cases) and demonstrates the least heterogeneity, with pathogenic

Figure 2.5 Clinical manifestation of RP revealed by fundus appearance using ophthalmoscopy. The panel on the left shows a normal healthy fundus while the panel on the right shows characteristic features of RP. Reprinted from Hartong *et al.,* (2006) with copyright permission.

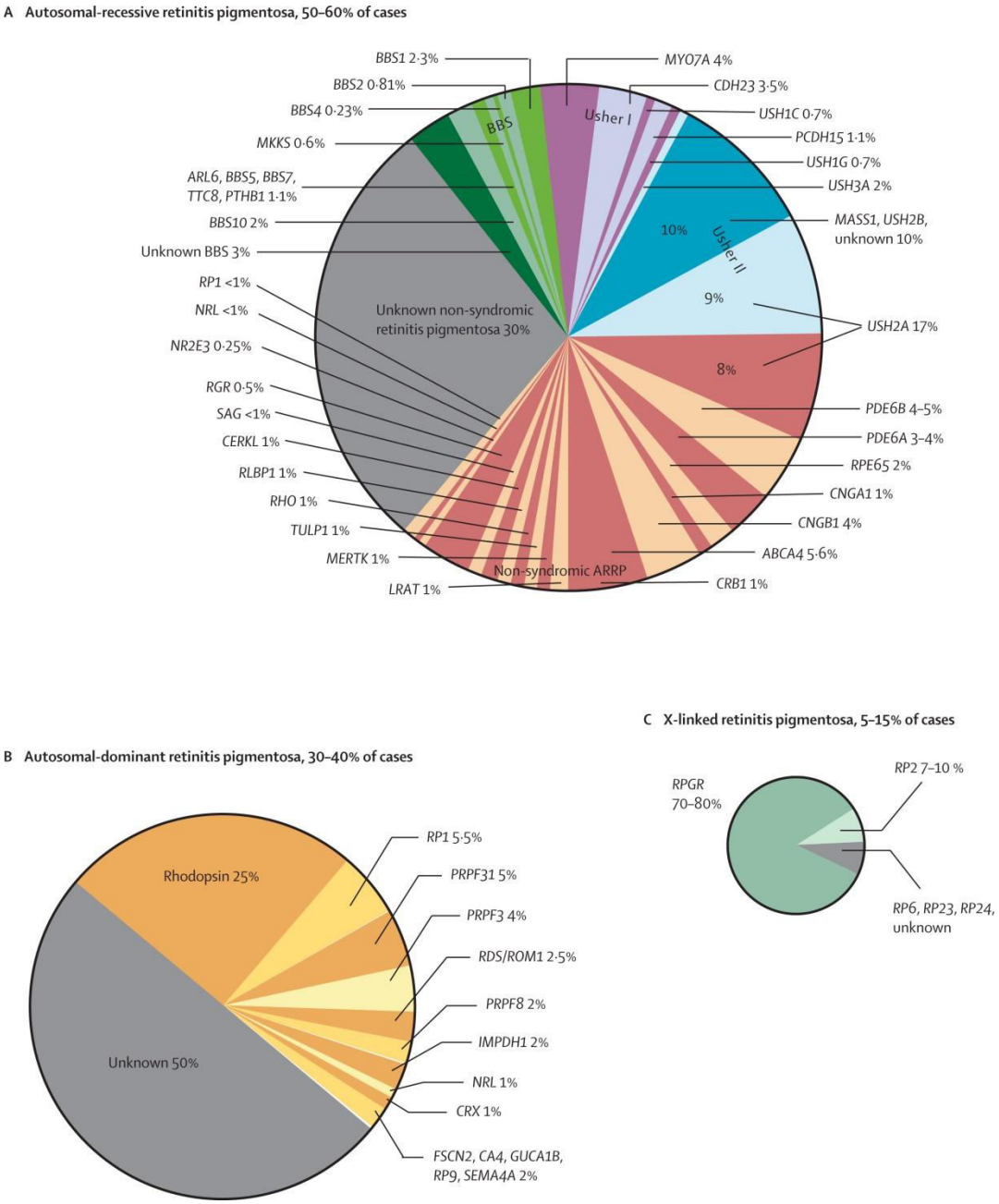


Figure 2.6 Proportion of RP cases caused by known and unknown genes displayed for each respective mode of inheritance. Reprinted from Hartong *et al.,* (2006) with copyright permission.

variants identified in *RP2*, *RP6*, *RP23*, *RP24* and *RPGR* (70-80% of X-linked cases). Autosomal dominant RP (adRP) accounts for 30-40% of cases. There are more than 22 genes identified where mutations cause adRP. The most common gene is *RHO*, encoding rhodopsin, and accounting for approximately 25% of adRP cases. The remaining identified adRP genes each account for small percentages (<5%), while a significant proportion (~50%) of adRP remains unsolved. Finally, autosomal recessive RP (arRP) accounts for the majority of cases of RP, with approximately 50-60% of cases. A substantial number of genes (over 30 genes and loci) account for arRP. Usher Syndrome Type II (RP with hearing loss) accounts for a large fraction of arRP, and *USH2A* mutations account for approximately 17% of arRP. *USH2A* is an example of clinical heterogeneity, as it can cause both syndromic RP (Usher syndrome) and non-syndromic RP. Approximately 9% of arRP cases are accounted for by *USH2A* causing Usher Syndrome, while 8% of arRP cases is accounted for by *USH2A*, causing non-syndromic RP. The majority of strictly non-syndromic RP genes individually account for a small percentage of arRP (<5%), but collectively explain the largest percentage (Figure 2.6). Finally, the genetic etiology of up to 30% of non-syndromic RP is unknown.

In humans, *MERTK* mutations are a rare cause of RP, accounting for approximately 1% of cases. Based on findings in the RCS rat, the first *MERTK* mutations found in humans were described by Gal *et al*., (2000). Gal and coauthors screened 328 DNA samples from IRD cases and identified three *MERTK* mutations in RP families. Since the initial discovery, there have been more than 33 pathogenic

missense, 12 nonsense, 12 splicing, 14 indels and 3 duplications (Audo *et al.,* 2018). Furthermore, three large *MERTK* deletions have been reported in RP families (Mackay *et al.,* 2010; Ostergaard *et al.,* 2011; Siemiatkowska *et al.,* 2011).

2.2 Study Aims & Objectives

In this study, we hypothesized the existence of a pathogenic variant segregating in a large family with non-syndromic arRP from Newfoundland. The study aim was to uncover this causal variant to provide a genetic diagnosis, and gain insight into the molecular pathogenesis of RP. The primary study objective, therefore, was to explore and interpret whole exome and genetic linkage data in this family. To prove causality, the next objective was to validate, and test segregation of candidate variants in the family. The final objective was to characterize the pathogenic variant, and perform a comprehensive clinical review to assist clinicians and researchers in understanding the manifestations of this rare disorder.

2.3 Author Contributions

The following is an original research article peer-reviewed and published in Investigative Ophthalmology & Visual Science (Evans *et al.,* 2017). As first author of the study, I was responsible for interpretation of whole exome sequence data to identify the causal pathogenic variant in this family. I validated the candidate variant in DNA samples of family members using PCR and Sanger sequencing, and tested variant segregation in all family members. I reviewed and interpreted clinical data for family members and prepared the manuscript. Gordon J. Johnson (GJJ) and Jane

S. Green (JSG) ascertained patients and clinical data. GJJ performed ophthalmic examinations, interpreted clinical findings, and provided referral to genetics in collaboration with JSG. Justin French (JF) performed funduscopy of patient VI-9. JSG conceived of the study, interviewed patients, interpreted molecular findings and assisted in manuscript revision. Generation of whole exome sequence and genetic linkage data was performed under the FORGE Canada framework. Kym M. Boycott (KMB) Jacek Majewski (JM), Somayyeh Fahiminiya (SF) and Jeremy Schwartzentruber (JS) supervised and assisted in whole exome sequencing analysis, and with result interpretation. Andrew D. Patterson (AP), Christian R. Marshall (CRM), Tara A. Paton (TAP) and Nicole M. Roslin (NMR) assisted in the generation of microarray data for linkage and copy number variant (CNV) analysis. Chandree L. Beaulieu (CLB) and Wen Qin (WQ) mapped the breakpoints of the *MERTK* deletion using primer walking. Michael O. Woods (MOW) participated in the study design, supervised the collection of molecular and genetic data, and assisted in manuscript revision. All authors read and approved the final manuscript.

Co-authors were affiliated with the following institutions during the completion of this work. GJJ was affiliated with the Department of Surgery (Ophthalmology), Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada. JSG and MOW were affiliated with the Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada. JS and JM were jointly affiliated with the Department of Human Genetics, McGill University, Montreal, Quebec, Canada and McGill University

and Genome Quebec Innovation Centre, Montreal, Quebec, Canada. CLB, WQ, SF and KMB were affiliated with the Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Ontario, Canada. TAP, NMR, CRM and AP were affiliated with the Program in Genetics and Genome Biology, The Center for Applied Genomics, The Hospital for Sick Children Research Institute, and Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. CF was affiliated with the Western Memorial Regional Clinic (Eye Care Centre), Corner Brook, Newfoundland, Canada.

2.4 Abstract

RP describes a complex group of inherited retinal dystrophies with almost 300 reported genes and loci. We investigated the genetic etiology of autosomal recessive RP (arRP) in a large kindred with five affected family members, who reside on the island of Newfoundland, Canada. Genetic linkage analysis was performed on 12 family members (Infinium HumanOmni2.5-8 BeadChip). Whole exome sequencing analysis (Illumina HiSeq) was performed on one affected individual. A custom pipeline was applied to call CNVs, annotate, and filter variants. FishingCNV was used to scan the exome for rare CNVs. Candidate CNVs subsequently were visualized from microarray data (CNVPartition v.3.1.6.). *MERTK* breakpoints were mapped and familial cosegregation was tested using Sanger Sequencing. We found strong evidence of linkage to a locus on chromosome 2 (logarithm of the odds [LOD] 4.89 [θ = 0]), at an interval encompassing the *MERTK* gene. Whole exome sequencing did not uncover candidate point mutations in *MERTK*, or other known RP genes.

Subsequently, CNV analysis of the exome data and breakpoint mapping revealed a 25,218 bp deletion of *MERTK*, encompassing exons 6 to 8, with breakpoints in introns 5 (chromosome 2:112,725,292) and 8 (chromosome 2:112,750,421). A 48 bp insertion sequence was buried within the breakpoint; 18 bps shared homology to *MIR4435-2HG* and *LINC00152*, and 30 bp mapped to *MERTK*. The deletion cosegregated with arRP in the family. This study describes the molecular and clinical characterization of an arRP family segregating a novel 25 kb deletion of *MERTK*. These findings may assist clinicians in providing a diagnosis for other unsolved RP cases.

2.5 Introduction

RP is a broad term meant to describe a complex group of hereditary retinal dystrophies. These dystrophies are characterized by the progressive degeneration of photoreceptor cells and/or RPE, which typically leads to severe visual impairment or blindness by the fourth decade of life (Hartong *et al.,* 2006). Symptoms of RP typically include early difficulty with night vision and dark adaptation, progressive loss of peripheral and/or central fields of vision, and declining visual acuity in the later stages. On examination of the fundus, RP often manifests with a triad of clinical features: a waxy pallor of the optic disc, attenuation of retinal blood vessels, and pathognomonic pigmentary aggregates called bone spicules (Hamel, 2006; Pruett, 1983). There can be great variability in the clinical manifestation of RP, owing to the large number of causative genes and metabolic pathways affected (Hartong *et al.,* 2006). Moreover, carriers of pathogenic RP mutations sometimes display phenotypic

variability within families, a phenomenon likely due to a combination of genetic modifiers and environmental influences (Chang *et al.,* 2011; Leroy *et al.,* 2007; Passerini *et al.,* 2007). Depending on the causal gene or mutation(s), patients with RP can experience onset of visual symptoms in a range from early childhood to mid-adulthood, with the progression and severity of disease influenced by the particular molecular pathways involved. The genetic etiology of inherited retinal dystrophies is complex, with nearly 300 genes and loci currently reported (RetNet, http:www.sph.uth.tmec.edu/RetNet/; in the public domain). (Daiger *et al.,* 2013). The investigative approaches to uncovering these causal genes have been well-described (Siemiatkowska *et al.,* 2014). Mutations of the *MERTK* gene explain approximately 1% of cases of autosomal recessive RP (arRP) (Tschernutter *et al.,* 2006). Pathologic mutations of this gene were first described in the RCS rat (D'Cruz *et al.,* 2000) and subsequently in human RP patients (Gal *et al.,* 2000).

MERTK has an essential role in maintaining retinal photoreceptor viability by facilitating phagocytosis of shed photoreceptor outer segments (Qingxian *et al.,* 2010; Strick and Vollrath, 2010). In humans, *MERTK* mutations are a well-documented cause of arRP (Brea-Fernández *et al.,* 2008; Coppieters *et al.,* 2014; Ebermann *et al.,* 2007; McHenry *et al.,* 2004; Shahzadi *et al.,* 2010; Srilekha *et al.,* 2015; Tada *et al.,* 2006; Thompson *et al.,* 2002). RP often displays a phenotype involving early onset of symptoms (night blindness) within the first decade of life and a rapid progression and severity of disease. Mutations of *MERTK* occur along the entire length of the gene (Jinda *et al.,* 2016). Several point mutations and small

indels in *MERTK* have been described, in addition to three large exonic deletions. These deletions include a 91 kb deletion encompassing exons 1 to 7 (Mackay *et al.,* 2010), a 9.86 kb deletion of exon 8 (Ostergaard *et al.,* 2011) and a 1.73 kb deletion of exon 15 with a complex re-arrangement (Siemiatkowska *et al.,* 2011).

We present findings of a fourth exonic deletion of *MERTK*; a 25 kb deletion encompassing exons 6 to 8, which was identified using whole exome data supplemented by CNV and linkage analyses. The deletion segregates in a large consanguineous arRP family with five affected individuals, who live on the island of Newfoundland, Canada (Figure 2.7). Affected individuals have an early-onset form of RP that affects peripheral and central fields of vision, and leads to a rapid decline in visual acuities and eventual blindness.

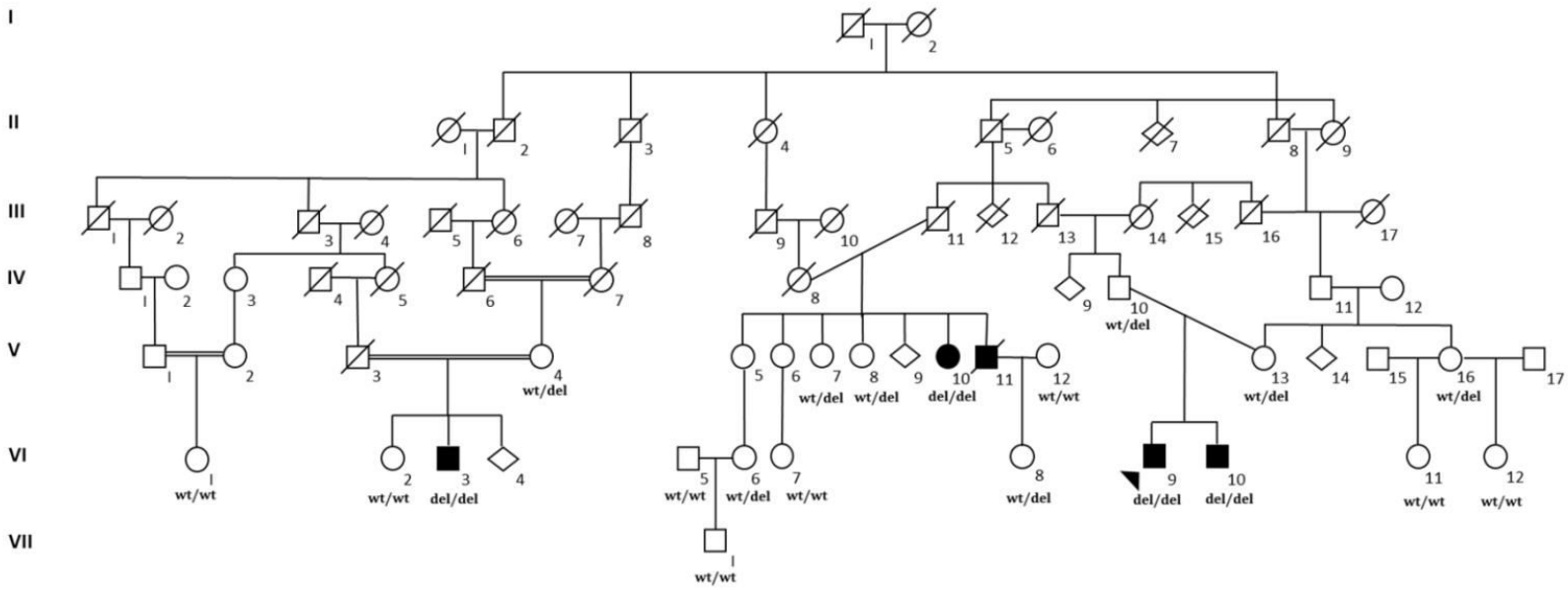


Figure 2.7 A large arRP family from Newfoundland, Canada, segregating the 25 kb (c.845-1450del; p.Ala282_His483del) deletion of *MERTK*. Reprinted from (Evans *et al.,* 2017) with copyright permission under creative commons licensing. Shaded symbols indicate clinically manifested RP. Screening for the deletion shows deletion carriers (wt/del), affected homozygotes (del/del), and wild-type non-carriers (wt/wt).

2.6 Materials and Methods

2.6.1 Patient Recruitment

Following ethical approval (HIC# 11.060), family members gave informed consent in adherence with tenets of the Declaration of Helsinki for participation in a research study. Genomic DNA was extracted from peripheral leukocytes using standard protocols for whole blood DNA extraction. The family was enrolled in the Finding of Rare Disease Genes in Canada (FORGE Canada) Consortium study.

2.6.2 Genotyping and Linkage Analysis

Genomic DNA samples of 12 family members (Figure 2.7; individuals IV-10, V-4, V-7, V-8, V-10, V-12, V-13, VI-2, VI-3, VI-8, VI-9, VI-10) were genotyped for 2,379,855 SNPs, using the Infinium HumanOmni2.5-8 v1.0 BeadChip (Illumina, Inc., San Diego, CA, USA) according to the manufacturer's protocols at The Centre for Applied Genomics (Toronto, Canada). Briefly, 200 ng of DNA (4 µL at 50 ng/µL) was independently amplified, labeled, and hybridized to BeadChip microarrays and then scanned with default settings using Illumina iScan. Analysis and intrachip normalization of the resulting image files was performed using Illumina's GenomeStudio Genotyping Module software (v.2011) with default parameters. Genotype calls were generated using the Illumina-provided genotype cluster definitions file (HumanOmni2.5-8v1_C.egt) with a Gencall cutoff of 0.15. The CNVPartition v.3.1.6 module in GenomeStudio was used for CNV and (copy number neutral) loss of heterozygosity analysis with default parameters.

Following quality control and filtering protocols, 17,718 SNPs across the autosomes and X chromosome were suitable for linkage analysis. We assumed a recessive model of inheritance with a disease allele frequency of 0.1 and penetrance of 0.2, 0.2, and 99% for 0, 1, and 2 copies of the disease-causing allele, respectively. This corresponds to a disease prevalence of approximately 1%. A multipoint linkage analysis then was performed using a two-stage approach. First, the pedigree was broken down into two subpedigrees and then analyzed using an exact multipoint calculation. Subsequently, regions showing evidence of linkage were further analyzed using approximate methods and the full pedigree.

2.6.3 Whole Exome Analysis and FishingCNV

The genomic DNA of one affected patient (Figure 2.7; individual VI-9) was chosen for whole exome sequencing analysis. Whole exome capture and high-throughput sequencing was performed at the McGill University and Genome Québec Innovation Centre (Montréal, Canada). The Agilent SureSelect (Agilent Technologies, Santa Clara, CA, USA) 50 Mb All Exon Kit (V3) was used for target exome enrichment, and high-throughput sequencing was achieved using the Illumina HiSeq platform. An in-house annotation pipeline was applied to call and annotate sequence variants as previously described for other FORGE projects (Fahiminiya *et al*., 2014). Candidate homozygous variants were reviewed manually for evidence of pathogenicity. CNVs were identified in the whole exome dataset using FishingCNV (Shi and Majewski, 2013). Briefly, this program compares exon-level read counts in a test sample against a population of control samples (from FORGE) to establish rare CNVs. Exon-

level read counts were determined using GATK (v1.0) depth of coverage for individual VI-9, with comparison against 150 in-house control exomes. These exomes consisted of DNA samples that were sequenced during other rare disease gene projects carried out by FORGE, using the same target exome enrichment and sequencing protocols as per individual VI-9.

2.6.4 Breakpoint Analysis and Carrier Screening

Following the identification of a candidate CNV in *MERTK*, the genomic position of the deletion breakpoints were mapped using primer walking. Primer sequences were designed using Primer3 (Untergasser *et al.,* 2012). The primer sequences used to identify deletion breakpoints are as follows: Forward 1 primer (ACCTTCATCCTCACCACACC), Reverse 1 primer (TCAATTCCACAGCAGACACC), Forward 2 primer (CCTGCCCTCATCCATAAAGA), and Reverse 2 primer (GACACTGAAGCAGGGAGGTC). A PCR protocol was developed to capture deletion-specific amplicons in patients carrying the *MERTK* deletion (Appendix A: Supporting Materials for arRP Study). Twenty family members were screened by this custom PCR protocol, to distinguish noncarriers from heterozygous and homozygous deletion carriers (Supp Figure 2). To mitigate the possibility of PCR reactions yielding false-negative results, we repeated each reaction in triplicate for every individual tested. Sequence chromatograms for deletion carriers were inspected visually for alignment to regions of *MERTK* according to the GRCh38.p3 human reference genome.

2.7 Results

Affected patients of this family were initially seen by GJJ and JSG in 1983, after referral to an Ocular Genetics clinic. Referrals initially were received for three separate families; however, genealogical analysis later revealed a shared and complex ancestry (Figure 2.7). Individual V-11 (Figure 2.7) was first seen in 1975 at age 28 with central vision loss. Visual acuities were 20/120 in both eyes at age 29, and by age 30, vision was reduced to hand movements only. At age 35, posterior capsular cataracts developed in the left eye, and funduscopy revealed pigmentary clumping, but no pallor of the optic discs or attenuation of retinal blood vessels. At 36, his fundus had a beaten-bronze appearance, with granular gray and whitish flecks in the midperiphery. Many bone spicules also were seen in the periphery. Individual VI-10 (Figure 2.7) reported that he was always night blind. At age 11 (1984), he had temporal constriction of the left visual field and a full field of vision in the right eye. Visual acuities were 20/20 in both eyes, and he had a color vision defect with 10/20 errors using the Ishihara test. His fundus had an albinoid appearance, with a marked decrease in chorioretinal pigmentation, and no bone spicules were found on examination. The following year at age 12, visual acuities decreased to 20/25 in both eyes, and his fundus had a salt and pepper appearance. Findings suggestive of delayed dark adaptation and nyctalopia were noted at this time. At age 16, visual acuities in both eyes further decreased to 20/40, and he reported difficulties in school, having trouble seeing the blackboard. At this time, his color vision continued to deteriorate, such that he scored 19/20 errors on the

Ishihara test, and had strong red-green and blue-yellow defects, seeing only one plate using the Hardy, Rand, and Rittler Pseudoisochromatic plate test. By age 19, he had a large central scotoma in both eyes, with peripheral vision only (nasally). By age 37, he had central scotomas to 30° to 40° in both eyes, and visual acuities were confined to hand movements only. The fundus displayed extensive chorioretinal degeneration. Alternating exotropia was noted. Individual VI: 3 (Figure 2.7) first experienced a decrease in central vision at age 7. The defect rapidly progressed, such that by age 12 the patient was enrolled at a school for the blind. Visual acuities continued to decrease, and color vision deficits were noted, as well as decreased night vision. The patient reported being sensitive to light for as long as he could remember. In 1980, at the age of 30, the patient had typical changes of RP, with macular degeneration as well. Visual acuities were reduced to hand movements only in the right eye, and light perception in the left. At age 35, fundi showed dense bone spicules at the equator, with an atrophic area in the mid periphery with grayish spots. At the macula, there was a small atrophic area with the appearance of choroidal sclerosis. Optic discs had a gray pallor and retinal blood vessels were narrowed. Visual acuities further decreased to light perception in both eyes. The patient had a marked exotropia, mainly of the left eye, which did alternate at times. Individual VI-9 (Figure 2.7) was first seen in 1986 (age 14) as his younger brother (VI-10) already was experiencing symptoms. At this time, VI-9 had 20/20 visual acuities in both eyes, and decreased chorioretinal pigmentation and stippling was noticed on funduscopic examination. He stopped playing hockey at age 15 due to

deteriorating vision, and reported he was always night blind. By age 17, uncorrected visual acuities were 20/80 in the right eye and 20/40 in the left. He had difficulties with classes and left school by grade 12. At age 20, he had developed a significant loss of central vision, such that only peripheral islands of vision (nasally) remained. Visual acuity was reduced to counting fingers only by age 30, and the patient subsequently lost all vision. Funduscopic surveillance of this patient at age 44 highlights the classic features of RP, including bone spicules, attenuation of retinal vessels, and pallor of the optic disc (Figure 2.8). A summary of the clinical findings of this family is provided in (Appendix A, Supplementary Tables & Figures, Supp Table 1). A genetic linkage analysis of 12 family members (Figure 2.7: individuals IV-10, V-4, V-7, V-8, V-10, V-12, V-13, VI-2, VI-3, VI-8, VI-9, VI-10) was performed to search for evidence of a disease-susceptibility locus in the family. Owing to the large size of the pedigree, we first performed an exact multipoint linkage analysis of subpedigrees, which revealed six regions on five chromosomes displaying logarithm of the odds (LOD) scores above 1.5 (Appendix A, Supplementary Tables & Figures). A maximum LOD score of 3.71 ($\theta = 0$) was observed on chromosome 2 encompassing markers rs7584136 and rs10177102 (chromosome 2:105833093-121145352). An approximate linkage analysis of these chromosomes subsequently was performed using the full pedigree, which increased the maximal LOD score to 4.89 ($\theta = 0$), indicating strong evidence of linkage at chromosome 2: 107493123-117426163 (Appendix A, Supp Figure 3, Supp Table 3). The region covered markers rs4676093 and rs4525709 (chromosome 2: 107493123 - 117426163), a 7.5 cM (10

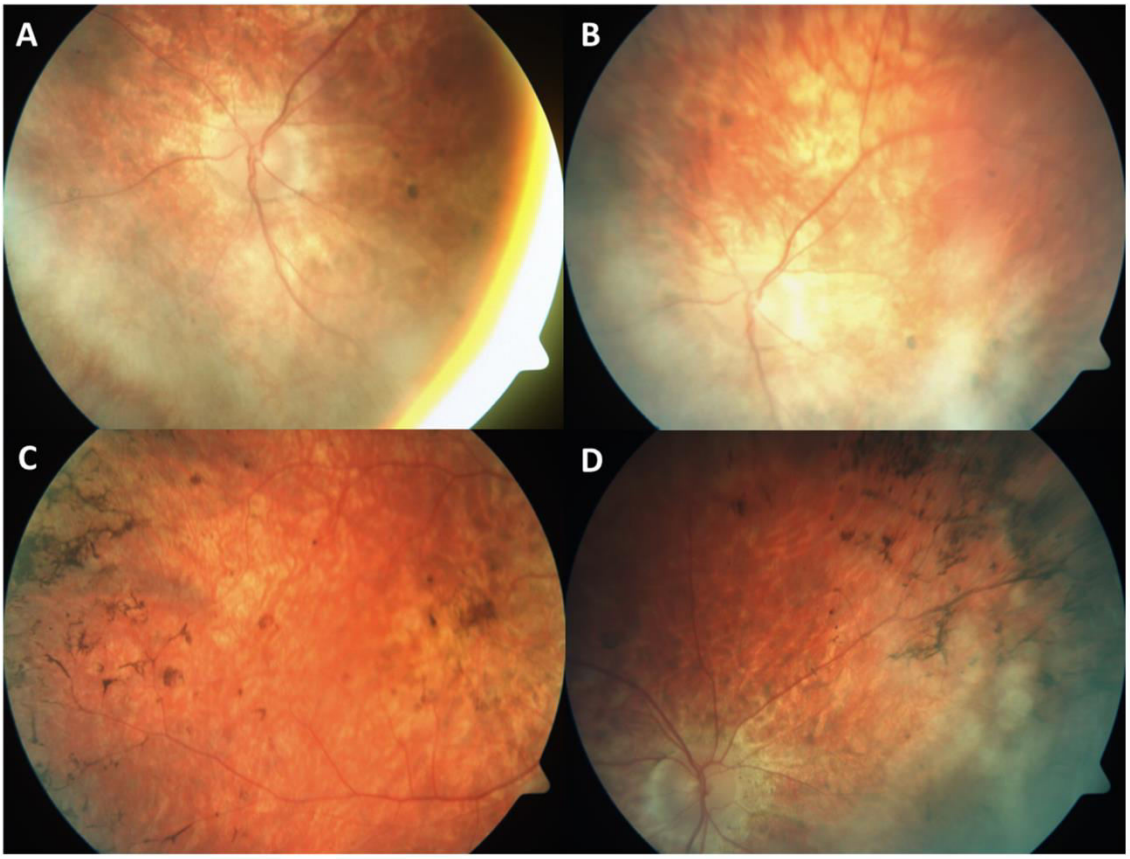


Figure 2.8 Fundus imaging of RP patient (VI-9), showing characteristic features of RP, including bone spicules, attenuation of retinal blood vessels, and pallor of the optic disc, with significant atrophy of the retina. (A, B) Patient's left eye. (C, D) Patient's right eye. Reprinted from Evans *et al.,* (2017) with Copyright Permission under creative commons licensing.

Mb) interval that includes the *MERTK* gene.

Next, we sequenced the exome of one affected patient (Fig. 1; individual VI-9), to screen for point mutations, small indels, and CNVs. In filtering the whole exome data, we included variants with minor allele frequency <5% and excluded variation that was synonymous or in untranslated regions (UTRs). This strategy produced 541 candidate variants; 32 of which were homozygous. Each homozygous variant then was assessed for evidence of pathogenicity by reviewing protein function, predicted mutation consequences, and evolutionary conservation scores. None of the 32 variants was in genes known to cause RP, in the interval with maximal LOD score, or appeared functionally relevant to an ocular phenotype (data not shown). Therefore, our exome filtering strategy did not produce any obvious candidate point mutations.

To screen the exome for candidate CNVs, we next applied FishingCNV. This program yielded 9 candidate CNVs after Holm-Bonferroni correction ($P < 0.05$). These candidates included amplifications or deletions in *ZFHX3*, *MERTK, AAK1*, *SMARCAL1*, *CCDC7*, *GIT2*, *NAALAD2*, *ZNF658*, and *FGGY* (Appendix, Supp Table 4). Among these, the CNV in *MERTK* was the only candidate functionally relevant to an ocular phenotype and also situated in a linked interval of LOD ≥ 1. This was a significant finding, given that aberrations of *MERTK* are known to cause RP, and this locus had the strongest evidence of linkage in the family, demonstrating a significant LOD score of (4.89 [$\theta = 0$]). The CNV in *MERTK* is a large 25 kb deletion spanning chromosome 2:112725714-112740570 and encompassing exons 6, 7, and 8. It is a

novel deletion, which to our knowledge, is not reported in the literature or online databases. Thus, we identified a strong candidate deletion for arRP in this family; a novel in-frame deletion of 25 kb in the *MERTK* gene (c.845-1450del; p.Ala282_His483del), which removes 201 amino acids from the encoded protein, corresponding to two fibronectin type-III domains.

To further investigate this mutation, we screened other family members for the deletion, by re-examining microarray data generated during the linkage analysis. This corroborated a heterozygous *MERTK* deletion in obligate carriers (Figure 2.7: individuals IV-10, V-4, V-13), and a homozygous deletion in the affected individuals with RP (Figure 2.7; V-10, VI-3, VI-9, VI-10), and no unaffected individuals were homozygous for the deletion. We then proceeded to map the breakpoints of the segregating *MERTK* deletion using primer walking. The 5′ breakpoint is located in intron 5 (chromosome 2:112,725,292), while the 3′ deletion breakpoint is in intron 8 (chromosome 2:112,750,421). The sequence surrounding the 5′ breakpoint maps to an L2B repeat of the LINE-2 family, while the 3′ breakpoint maps to a MER21A repeat sequence. The deletion spans a total of 25,218 base pairs according to the GRCh38.p3 human reference genome.

We next optimized a custom PCR protocol to validate the *MERTK* deletion and screen additional family members, by amplifying deletion-specific fragments (Appendix A Supplementary Methods). In brief, the PCR protocol is able to differentiate DNA samples with wild-type breakpoint sequences from heterozygous or homozygous deletion carriers (Appendix Supp Figure 2), and, therefore, may

prove useful for others who wish to screen for this deletion in unsolved arRP families. A total of 20 family members were screened and validated in this manner. As expected, the deletion was homozygous in affected individuals and heterozygous in obligate carriers (Figure 2.7), thereby replicating our previous findings. Screening identified a total of 8 carriers in the family, who did not have a phenotype similar to their homozygous affected cousins, and were not routinely followed by ophthalmology.

Interestingly, Sanger sequencing of the deletion-specific PCR amplicons revealed a 48 bp insertion sequence within the deletion breakpoints (Figure 2.9). A Basic Local Alignment Search Tool (BLAST) search of this sequence (ATTACTAGGTGAAGCACAGTGGAGCACATGGCTTGGTATAGGAGACCC), showed that the terminal 30 bp (GTGGAGCACATGGCTTGGTATAGGAGACCC) have a 97% identity to intron 5 of *MERTK*. The remaining proximal 18 bp of the insertion (ATTACTAGGTGAAGCACA) mapped with 100% identity to intronic regions of *MIR4435-2HG* and *LINC00152* (Figure 2.10). *MIR4435-2HG* is a microRNA approximately 700 kb upstream of *MERTK*, while *LINC00152* is a long noncoding RNA on the opposite arm of chromosome 2.

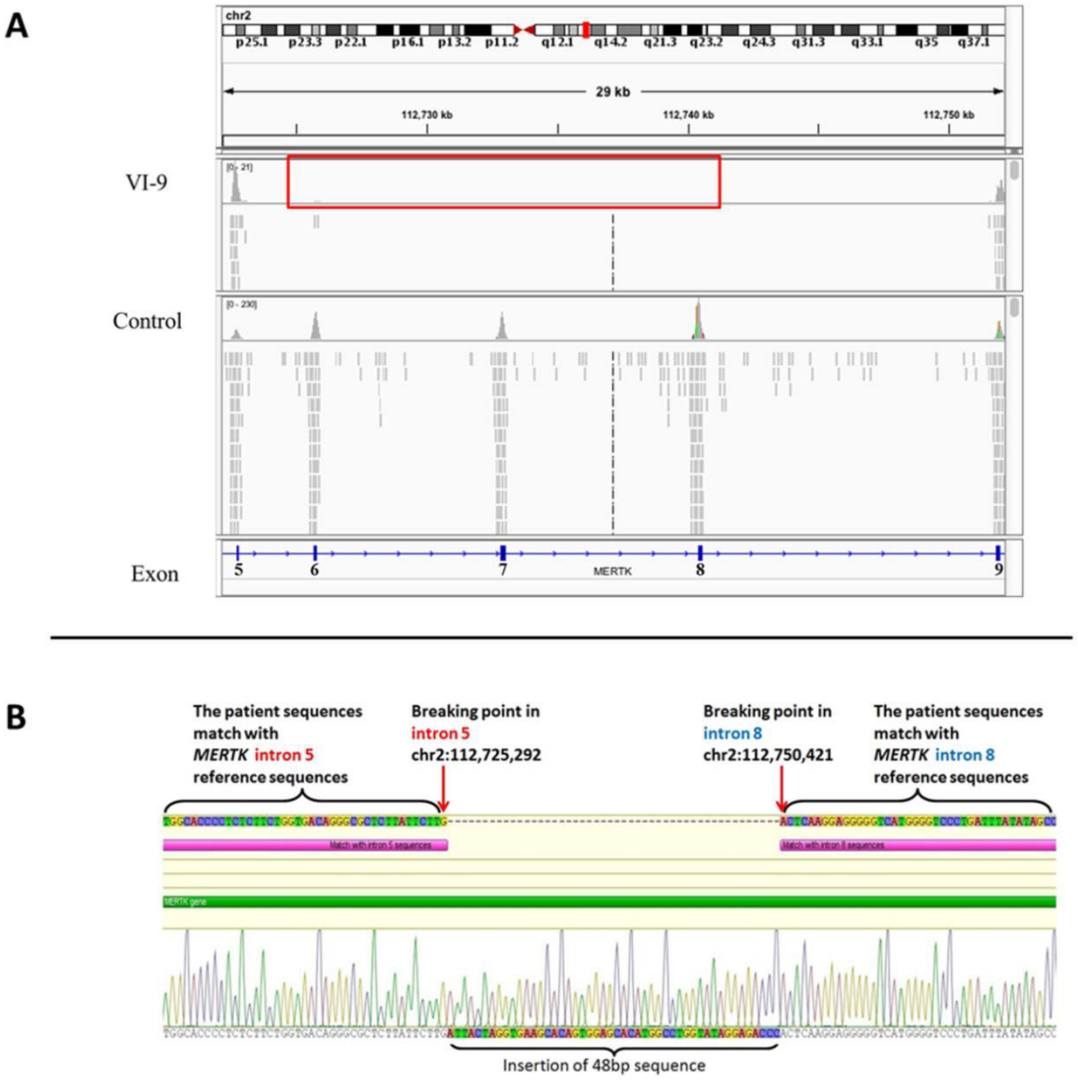


Figure 2.9 Novel 25 kb deletion of *MERTK* detected in WES data (A) and validated using Sanger sequencing (B). Validation of the deletion breakpoints (B) revealed a 48 bp insertion sequence in this affected family member. Reprinted from Evans *et al.,* (2017) with copyright permission under creative commons licensing.

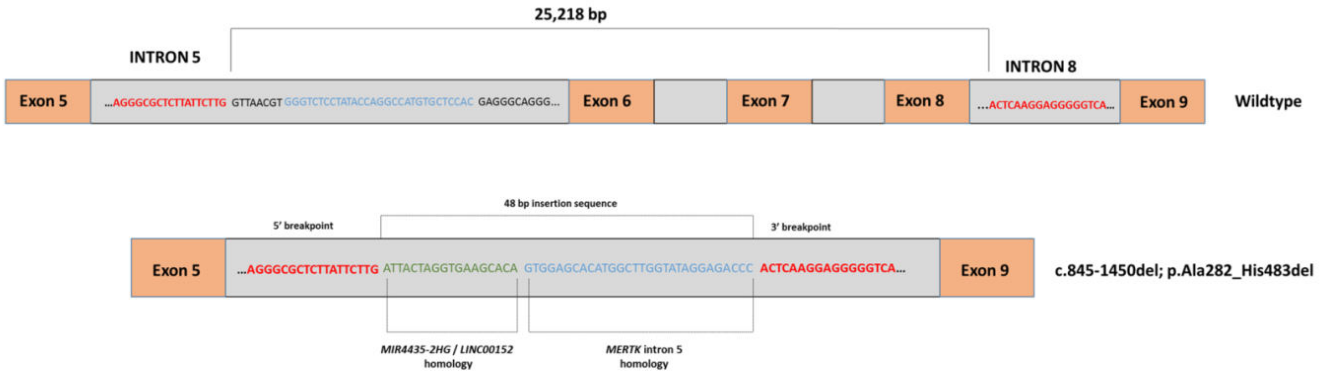


Figure 2.10 Diagrammatic representation of 25 kb *MERTK* (c.845-1450del, p.Ala282_His483del); an in-frame deletion encompassing exons 6 to 8, with a 48 bp insertion sequence at the breakpoints. Red font, the 5′ and 3′ breakpoints of the deletion; green font, 18 bp of insertion with homology to *MIR4435-2HG*; *blue font*, 30 bp region of homology to intron 5 of *MERTK*. Reprinted from Evans *et al.,* (2017) with copyright permission under creative commons licensing.

2.8 Discussion

This study provided compelling evidence for the discovery of a novel 25 kb deletion of *MERTK* causing arRP in this large family from the island of Newfoundland. We discuss our rationale for these findings below. First, we observed strong evidence of linkage disequilibrium at an interval that encompasses *MERTK* (LOD 4.89 [θ = 0]). Meanwhile, our whole exome filtering strategy did not uncover strong candidate mutations in known or putative RP genes. Subsequently, CNV analysis then uncovered a large exonic deletion of *MERTK* within the interval with the strongest linkage signal. Analysis of genotype data and a custom PCR protocol corroborated the segregation of this deletion in a pattern typical of autosomal recessive inheritance. Moreover, deletion of exon 8 is known to cause RP (Mackay *et al.,* 2010) and the deletion identified here encompasses exons 6 to 8. The large size of our study family provides important insights into the clinical manifestation of arRP caused by *MERTK* mutations.

The family is characterized by an early onset form of RP. Affected individuals all reported decreased or absent night vision early in life, and subsequently experienced a rapidly progressive deterioration of central and peripheral vision. Most individuals were blind by their 20th to 30th year. Deterioration of color vision also was noted in these patients, and funduscopy revealed classic changes of RP, with some variability in the detection of bone spicules (Appendix Supp Table 1). Indeed, other patients with *MERTK* mutations often experience earlier onset of symptoms compared to other forms of arRP.

This family segregates only the fourth exonic deletion of *MERTK* to be reported to our knowledge. The 25 kb deletion (c.845-1450del; p.Ala282_His483del), encompasses exons 6 to 8 and overlaps portions of two other *MERTK* deletions reported in the literature: one encompassing exons 1 to 7 (91 kb) (Ostergaard *et al.,* 2011) and the other a deletion of exon 8 (9.86 kb) (Mackay *et al.,* 2010).

Interestingly, some clinical overlap is seen in each of these families carrying *MERTK* deletions. For example, patients carrying the 91 kb deletion encompassing exons 1 to 7 also experienced onset of symptoms in the first decade of life, which led to deterioration of central and peripheral vision; a finding that also was seen in our patients. As well, authors of that study noted interindividual variation within their family with respect to retinal vessel attenuation and optic disc pallor. Interindividual variation is seen within our family as well, as individual V-11 (Figure 2.7) did not display any obvious attenuation of retinal vessels or pallor of the optic disc, while these features were prominent in individual VI-3. It remains unclear if these observations are due to differences in disease manifestation or thoroughness of surveillance. We also observed variability between two affected individuals of the same sibship (Figure 2.7: individuals VI-10, VI-9). Individual VI-10, the younger brother, appeared to have a more severe onset and progression of disease, as he began noticing decreased peripheral vision by age 11. Meanwhile, his older brother did not clinically manifest symptoms until age 14. Moreover, individual VI-3, a more distant relative, manifested symptoms at age 7 and

experienced significant visual deterioration such that he was registered at a school for the blind by age 12. Thus, the phenotype seen in our family is one that is severe and early in onset, with first reported decline in visual acuities often by the second decade, ranging from age 7 to age 30. In patients segregating the *MERTK* exon 8 deletion (Mackay *et al.,* 2010), the authors noted macular involvement, as the proband of their study had central vision deficits by age 13. Additionally, affected family members of their study experienced progressive color vision defects. For example, the proband of their study failed a color vision test with 24/24 errors at age 26, while the younger sibling, who also was affected, began experiencing mild generalized dyschromatopsia by age 8. Early color vision deficits also were noted in our study, as individual VI-10 experienced significant deficits at age 11 (10/20 errors), which markedly progressed by age 16 (19/20 errors). Moreover, individual IV-3 of our study also had early color vision deficits, documented at age 12, suggestive of early rod and cone involvement.

A peculiar finding of this study was the discovery of a 48 bp insertion sequence buried within the deletion breakpoint. In patients with the previously reported exon 8 deletion of *MERTK*, the authors uncovered a complete AluY element insertion, leading them to hypothesize nonhomologous recombination between two AluY elements as the mechanism of the deletion. Intrigued by those findings, we performed a BLAST search of our 48 bp insertion sequence, but did not find homology to Alu elements. We did, however, find homology to three intronic regions on chromosome 2. Shown in Figure 2.10, 30 bps of the insertion sequence map to a

stretch of intron 5, which is deleted in these patients, and appears to be the reverse complement compared to wild-type intron 5 reference sequences. The remaining 18 bp portion of the insertion sequence maps to two other intronic loci on chromosome 2: *MIR4435-2HG* and *LINC00152*. One possible explanation for these findings could be the insertion of a transposable element buried within one of these genes. The etiology of these exonic *MERTK* deletions appears complex, given that the third reported *MERTK* deletion is a removal of exon 15 (1732 bp), with a duplication and inversion event (Siemiatkowska *et al.,* 2011) with no mechanism postulated for those findings.

The deletion uncovered in this study (c.845-1450del; p.Ala282_His483del) removes 201 amino acids from encoded MERTK protein, corresponding to the removal of two fibronectin type-III domains. These domains function in the extracellular region of the protein, and together with the immunoglobulin domains, define *MERTK* within the TAM receptor tyrosine kinase family (Lemke, 2013). This deletion does not disrupt the reading frame, given that the 484th amino acid is retained as a glycine, rendering nonsense-mediated decay unlikely. It is likely, however, that the removal of the fibronectin domains could structurally alter the extracellular region of the MERTK protein, impairing its ability to bind GAS6 ligands. Binding of GAS6 ligands to *MERTK* is functionally important, as it is known to stimulate phagocytosis (Hall *et al.,* 2001; Qingxian *et al.,* 2010), and loss of this ability could lead to the observed accumulation of photoreceptor outer segment debris in the sub-retinal space.

This study highlights the importance of using alternative methods to screen arRP patients who have early onset and severe forms of RP. *MERTK* currently is believed to account for 1% of arRP cases; however, these numbers were estimated using a sequencing study that screened a cohort of 96 arRP cases (Tschernutter *et al.,* 2006). Therefore, it remains possible that the true proportion of arRP caused by *MERTK* mutations may differ when taking potentially undetected CNVs into account. Therefore, investigators should consider screening this gene for larger deletions in unsolved arRP families. The identification of novel pathogenic mutations and structural alterations remains an important process, particularly for the *MERTK* gene, as ongoing clinical trials continue to highlight the potential for the future application of gene therapy (Ghazi *et al.,* 2016), which will first require patients to have a genetic diagnosis.

We concluded that the discovery of this 25 kb *MERTK* deletion, abolishing exons 6 to 8, is a novel cause of arRP which ultimately leads to an early onset and rapid progression of visual symptoms seen in this family. We provided an optimized PCR protocol (Appendix A, Supplementary Methods) that others might use to detect this deletion in unsolved arRP families. The clinical histories provided here are of benefit to clinicians and investigators in understanding the clinical manifestation of *MERTK* deletions, and this is one of the largest families segregating a *MERTK* deletion currently reported in the literature. Moving forward, we encourage others to consider screening *MERTK* thoroughly for larger CNVs in the investigation of unsolved arRP cases.

2.9 References

Arshavsky, V. Y., Lamb, T. D., & Pugh, E. N. (2002). G Proteins and Phototransduction. *Annual Review of Physiology*, *64*(1), 153–187. https://doi.org/10.1146/annurev.physiol.64.082701.102229

Audo, I., Mohand-Said, S., Boulanger-Scemama, E., Zanlonghi, X., Condroyer, C., Démontant, V., … Zeitz, C. (2018). MERTK mutation update in inherited retinal diseases. *Human Mutation*, *39*(7), 887–913. https://doi.org/10.1002/humu.23431

Brea-Fernández, A. J., Pomares, E., Brión, M. J., Marfany, G., Blanco, M. J., Sánchez-Salorio, M., … Carracedo, A. (2008). Novel splice donor site mutation in MERTK gene associated with retinitis pigmentosa. *The British Journal of Ophthalmology*, *92*(10), 1419–1423. https://doi.org/10.1136/bjo.2008.139204

Cayouette, M., Smith, S. B., Becerra, S. P., & Gravel, C. (1999). Pigment epithelium-derived factor delays the death of photoreceptors in mouse models of inherited retinal degenerations. *Neurobiology of Disease*, *6*(6), 523–532. https://doi.org/10.1006/nbdi.1999.0263

Chang, S., Vaccarella, L., Olatunji, S., Cebulla, C., & Christoforidis, J. (2011). Diagnostic challenges in retinitis pigmentosa: genotypic multiplicity and phenotypic variability. *Current Genomics*, *12*(4), 267–275. https://doi.org/10.2174/138920211795860116

Coppieters, F., Van Schil, K., Bauwens, M., Verdin, H., De Jaegher, A., Syx, D., … De

Baere, E. (2014). Identity-by-descent-guided mutation analysis and exome sequencing in consanguineous families reveals unusual clinical and molecular findings in retinal dystrophy. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *16*(9), 671–680. https://doi.org/10.1038/gim.2014.24

Daiger, S. P., Sullivan, L. S., & Bowne, S. J. (2013). Genes and mutations causing retinitis pigmentosa. *Clinical Genetics*, *84*(2), 132–141. https://doi.org/10.1111/cge.12203

D'Cruz, P. M., Yasumura, D., Weir, J., Matthes, M. T., Abderrahim, H., LaVail, M. M., & Vollrath, D. (2000). Mutation of the receptor tyrosine kinase gene Mertk in the retinal dystrophic RCS rat. *Human Molecular Genetics*, *9*(4), 645–651.

Ebermann, I., Walger, M., Scholl, H. P. N., Issa, P. C., Lüke, C., Nürnberg, G., … Bolz, H. J. (2007). Truncating mutation of the DFNB59 gene causes cochlear hearing impairment and central vestibular dysfunction. *Human Mutation*, *28*(6), 571–577. https://doi.org/10.1002/humu.20478

Euler, T., Haverkamp, S., Schubert, T., & Baden, T. (2014). Retinal bipolar cells: elementary building blocks of vision. *Nature Reviews Neuroscience*, *15*(8), 507–519. https://doi.org/10.1038/nrn3783

Evans, D. R., Green, J. S., Johnson, G. J., Schwartzentruber, J., Majewski, J., Beaulieu, C. L., … FORGE Canada Consortium. (2017). Novel 25 kb Deletion of MERTK Causes Retinitis Pigmentosa With Severe Progression. *Investigative Ophthalmology & Visual Science*, *58*(3), 1736–1742.

https://doi.org/10.1167/iovs.16-20864

Fung, B. K., Hurley, J. B., & Stryer, L. (1981). Flow of information in the light-triggered cyclic nucleotide cascade of vision. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(1), 152–156.

Gal, A., Li, Y., Thompson, D. A., Weir, J., Orth, U., Jacobson, S. G., … Vollrath, D. (2000). Mutations in MERTK, the human orthologue of the RCS rat retinal dystrophy gene, cause retinitis pigmentosa. *Nature Genetics*, *26*(3), 270–271. https://doi.org/10.1038/81555

Ghazi, N. G., Abboud, E. B., Nowilaty, S. R., Alkuraya, H., Alhommadi, A., Cai, H., … Alkuraya, F. S. (2016). Treatment of retinitis pigmentosa due to MERTK mutations by ocular subretinal injection of adeno-associated virus gene vector: results of a phase I trial. *Human Genetics*, *135*(3), 327–343. https://doi.org/10.1007/s00439-016-1637-y

Hall, M. O., Prieto, A. L., Obin, M. S., Abrams, T. A., Burgess, B. L., Heeb, M. J., & Agnew, B. J. (2001). Outer Segment Phagocytosis by Cultured Retinal Pigment Epithelial Cells Requires Gas6. *Experimental Eye Research*, *73*(4), 509–520. https://doi.org/10.1006/exer.2001.1062

Hamel, C. (2006). Retinitis pigmentosa. *Orphanet Journal of Rare Diseases*, *1*, 40. https://doi.org/10.1186/1750-1172-1-40

Hargrave, P. A. (2001). Rhodopsin Structure, Function, and Topography The Friedenwald Lecture. *Investigative Ophthalmology & Visual Science*, *42*(1), 3–9.

Hartong, D. T., Berson, E. L., & Dryja, T. P. (2006). Retinitis pigmentosa. *The Lancet*, *368*(9549), 1795–1809. https://doi.org/10.1016/S0140-6736(06)69740-7

Jinda, W., Poungvarin, N., Taylor, T. D., Suzuki, Y., Thongnoppakhun, W., Limwongse, C., … Atchaneeyasakul, L. (2016). A novel start codon mutation of the MERTK gene in a patient with retinitis pigmentosa. *Molecular Vision*, *22*, 342–351.

Joselevitch, C. (2008). Human retinal circuitry and physiology. *Psychology & Neuroscience*, *1*(2), 141–165. https://doi.org/10.3922/j.psns.2008.2.008

Lemke, G. (2013). Biology of the TAM receptors. *Cold Spring Harbor Perspectives in Biology*, *5*(11), a009076. https://doi.org/10.1101/cshperspect.a009076

Leroy, B. P., Kailasanathan, A., De Laey, J., Black, G. C. M., & Manson, F. D. C. (2007). Intrafamilial phenotypic variability in families with RDS mutations: exclusion of ROM1 as a genetic modifier for those with retinitis pigmentosa. *The British Journal of Ophthalmology*, *91*(1), 89–93. https://doi.org/10.1136/bjo.2006.101915

Mackay, D. S., Henderson, R. H., Sergouniotis, P. I., Li, Z., Moradi, P., Holder, G. E., … Moore, A. T. (2010). Novel mutations in MERTK associated with childhood onset rod-cone dystrophy. *Molecular Vision*, *16*, 369.

Mayhew, T. M., & Astle, D. (1997). Photoreceptor number and outer segment disk membrane surface area in the retina of the rat: stereological data for whole organ and average photoreceptor cell. *Journal of Neurocytology*, *26*(1), 53–61.

McHenry, C. L., Liu, Y., Feng, W., Nair, A. R., Feathers, K. L., Ding, X., … Thompson, D. A. (2004). MERTK Arginine-844-Cysteine in a Patient with Severe Rod–Cone

Dystrophy: Loss of Mutant Protein Function in Transfected Cells. *Investigative Ophthalmology & Visual Science*, *45*(5), 1456–1463. https://doi.org/10.1167/iovs.03-0909

Nangia, V., Jonas, J. B., Khare, A., & Sinha, A. (2012). Prevalence of retinitis pigmentosa in India: the Central India Eye and Medical Study. *Acta Ophthalmologica*, *90*(8), e649-650. https://doi.org/10.1111/j.1755-3768.2012.02396.x

Ostergaard, E., Duno, M., Batbayli, M., Vilhelmsen, K., & Rosenberg, T. (2011). A novel MERTK deletion is a common founder mutation in the Faroe Islands and is responsible for a high proportion of retinitis pigmentosa cases. *Molecular Vision*, *17*, 1485–1492.

Pagon, R. A. (1988). Retinitis pigmentosa. *Survey of Ophthalmology*, *33*(3), 137–177.

Panda-Jonas, S., Jonas, J. B., & Jakobczyk-Zmija, M. (1995). Retinal photoreceptor density decreases with age. *Ophthalmology*, *102*(12), 1853–1859.

Passerini, I., Sodi, A., Giambene, B., Menchini, U., & Torricelli, F. (2007). Phenotypic intrafamilial variability associated with S212G mutation in the RDS/peripherin gene. *European Journal of Ophthalmology*, *17*(6), 1000–1003.

Photoreception - Structure and function of photoreceptors. (n.d.). Retrieved August 1, 2018, from https://www.britannica.com/science/photoreception

Pruett, R. C. (1983). Retinitis pigmentosa: clinical observations and correlations. *Transactions of the American Ophthalmological Society*, *81*, 693–735.

Qingxian, L., Qiutang, L., & Qingjun, L. (2010). Regulation of phagocytosis by TAM

receptors and their ligands. *Frontiers in Biology*, *5*(3), 227–237.

Shahzadi, A., Riazuddin, S. A., Ali, S., Li, D., Khan, S. N., Husnain, T., … Riazuddin, S. (2010). Nonsense mutation in MERTK causes autosomal recessive retinitis pigmentosa in a consanguineous Pakistani family. *The British Journal of Ophthalmology*, *94*(8), 1094–1099. https://doi.org/10.1136/bjo.2009.171892

Shi, Y., & Majewski, J. (2013). FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data. *Bioinformatics (Oxford, England)*, *29*(11), 1461–1462. https://doi.org/10.1093/bioinformatics/btt151

Shichida, Y., & Morizumi, T. (2007). Mechanism of G-protein Activation by Rhodopsin†. *Photochemistry and Photobiology*, *83*(1), 70–75. https://doi.org/10.1562/2006-03-22-IR-854

Siemiatkowska, A. M., Arimadyo, K., Moruz, L. M., Astuti, G. D. N., de Castro-Miro, M., Zonneveld, M. N., … Collin, R. W. J. (2011). Molecular genetic analysis of retinitis pigmentosa in Indonesia using genome-wide homozygosity mapping. *Molecular Vision*, *17*, 3013–3024.

Siemiatkowska, A. M., Collin, R. W. J., den Hollander, A. I., & Cremers, F. P. M. (2014). Genomic Approaches For the Discovery of Genes Mutated in Inherited Retinal Degeneration. *Cold Spring Harbor Perspectives in Medicine*, *4*(8). https://doi.org/10.1101/cshperspect.a017137

Srilekha, S., Arokiasamy, T., Srikrupa, N. N., Umashankar, V., Meenakshi, S., Sen, P., …

Soumittra, N. (2015). Homozygosity Mapping in Leber Congenital Amaurosis and Autosomal Recessive Retinitis Pigmentosa in South Indian Families. *PLOS ONE*, *10*(7), e0131679. https://doi.org/10.1371/journal.pone.0131679

Strauss, O. (2005). The Retinal Pigment Epithelium in Visual Function. *Physiological Reviews*, *85*(3), 845–881. https://doi.org/10.1152/physrev.00021.2004

Strick, D. J., & Vollrath, D. (2010). Focus on Molecules: MERTK. *Experimental Eye Research*, *91*(6), 786–787. https://doi.org/10.1016/j.exer.2010.05.006

Tada, A., Wada, Y., Sato, H., Itabashi, T., Kawamura, M., Tamai, M., & Nishida, K. (2006). Screening of the MERTK gene for mutations in Japanese patients with autosomal recessive retinitis pigmentosa. *Molecular Vision*, *12*, 441–444.

Thompson, D. A., McHenry, C. L., Li, Y., Richards, J. E., Othman, M. I., Schwinger, E., … Gal, A. (2002). Retinal Dystrophy Due to Paternal Isodisomy for Chromosome 1 or Chromosome 2, with Homoallelism for Mutations in RPE65 or MERTK, Respectively. *American Journal of Human Genetics*, *70*(1), 224–229.

Tschernutter, M., Jenkins, S. A., Waseem, N. H., Saihan, Z., Holder, G. E., Bird, A. C., … Webster, A. R. (2006). Clinical characterisation of a family with retinal dystrophy caused by mutation in the Mertk gene. *The British Journal of Ophthalmology*, *90*(6), 718–723. https://doi.org/10.1136/bjo.2005.084897

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115. https://doi.org/10.1093/nar/gks596

Wright, A. F., Chakarova, C. F., Abd El-Aziz, M. M., & Bhattacharya, S. S. (2010).

Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. *Nature Reviews. Genetics*, *11*(4), 273–284. https://doi.org/10.1038/nrg2717

Chapter 3. Novel Mutation Discovered for Weill-Marchesani Syndrome in a Newfoundland Family Using Whole Exome Sequencing and Homozygosity Mapping.

3.1 Background: Weill-Marchesani Syndrome

As the most abundant tissue of the body, connective tissues play key roles in body structure, protecting organs and maintaining homeostasis. Four main categories of connective tissues are the connective tissue proper (including loose and dense connective tissue), cartilage, bone and blood. The connective tissue proper has three core components: ground substance, fibers and cells (Figure 3.1). Ground substance is a viscous fluid primarily composed of proteoglycans and cell adhesion proteins, which allows the flow of molecules between surrounding cells and capillaries. Meanwhile, the connective tissue fibers (i.e. collagen, elastic fibers and reticular fibers) provide structural support. Together with the ground substance, these fibers make up the extracellular matrix (ECM).

Elastic fibers are an important component of the ECM. They are made of cross-linked elastin and fibrillin proteins, and are important for load-bearing tissues, such as the heart, lungs, kidneys, blood vessels, and components of the eye. Fibrillins are large glycoproteins (~350 kDa) that polymerise to form the predominant component of microfibrils. The mechanism of microfibril assembly is poorly understood. Microfibrils are seen in two forms – in loose association with elastic fibers (i.e. in the skin, arteries and lungs), or alternatively, in tightly bound, non-elastic microfibril bundles (i.e. in ciliary zonules, tendons, glomeruli) (Figure 3.2).

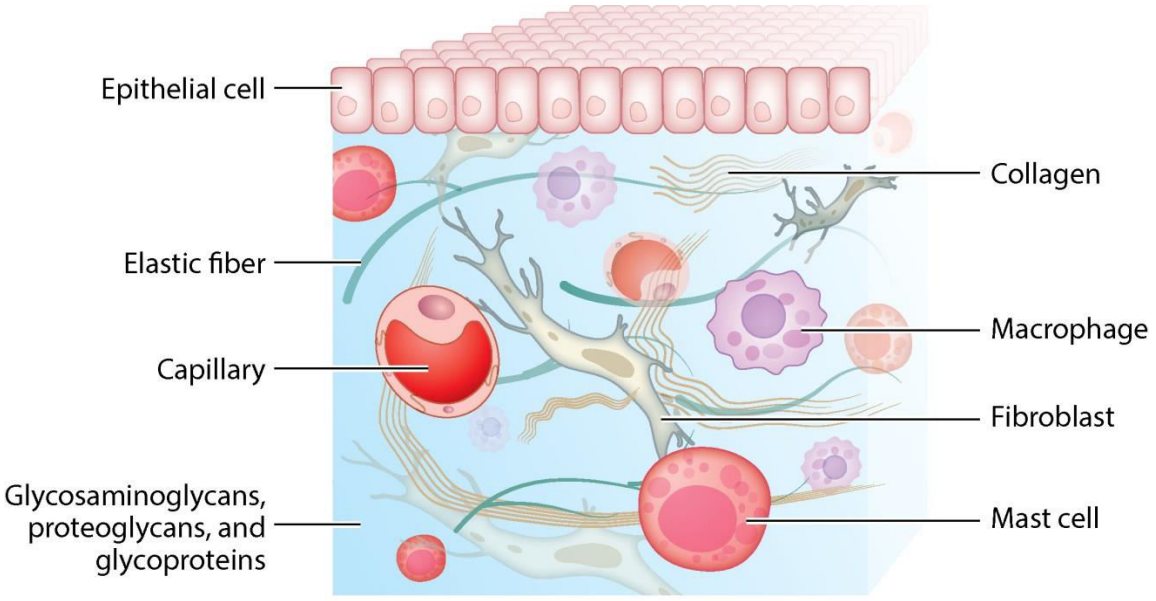


Figure 3.1 Basic components of connective tissue. Together, the glycosaminoglycans (ground substance) and fibers (i.e. collagen and elastic fibers) make up the ECM. Reprinted from Vanakker *et al.,* (2015) with copyright permission.

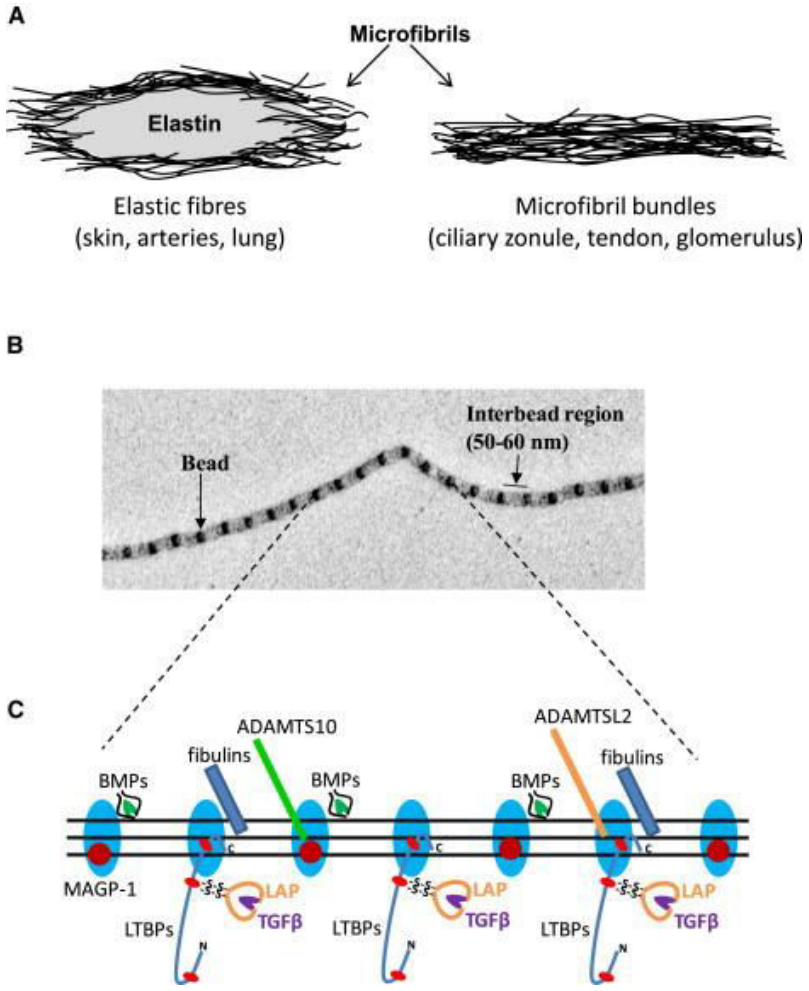


Figure 3.2 The structure of connective tissue microfibrils, highlighting the complex protein-protein interaction of microfibril bundles. Reprinted from Jensen *et al.,* (2012) with copyright permission.

Microfibrils confer pliability to the skin as a loose association of fibers and elastin, running parallel to the epidermis (Ramirez *et al.,* 2004). Similarly, in the aorta, elastin and micofibril bundles provide critical elasticity to a tissue undergoing large hydrostatic pressures.

Meanwhile, non-elastic bundles of micofibrils are dynamic molecules that interact with and are re-modelled by a host of proteins within the extracellular matrix. They also play a key role in sequestration of latent TGF-β (Zeyer and Reinhardt, 2015). In the eye, microfibril bundles play a key structural role in the ciliary zonules (i.e. Zonules of Zinn), which are suspensory ligaments connecting the ciliary body of the eye to the ocular lens (Figure 3.3). If the ciliary zonules are structurally compromised, then the lens can either partially or fully dislocate (ectopia lentis) (Figure 3.4). Bilateral ectopia lentis can be pathognomic for certain hereditary connective tissue disorders, for example in Marfan Syndrome and Weill-Marchesani Syndrome (WMS) when identified with supporting systemic features.

The critical role of microfibril stability in tissues such as the skin, aorta, long bones, and ciliary zonules was demonstrated in Marfan Syndrome. In 1991, shortly after fibrillin-1 protein was isolated from microfibrils, the critical role of the *FBN1* gene was demonstrated by Dietz *et al.,* who found a pathogenic variant p.R239P in a region of strong genetic linkage (LOD 3.9) (Dietz *et al.,* 1991).

FBN1 mutations produce substantial phenotypic heterogeneity -- an intriguing example of how mutations in a single gene can produce opposite phenotypes. Collectively, Mendelian disorders caused by mutations in *FBN1* are

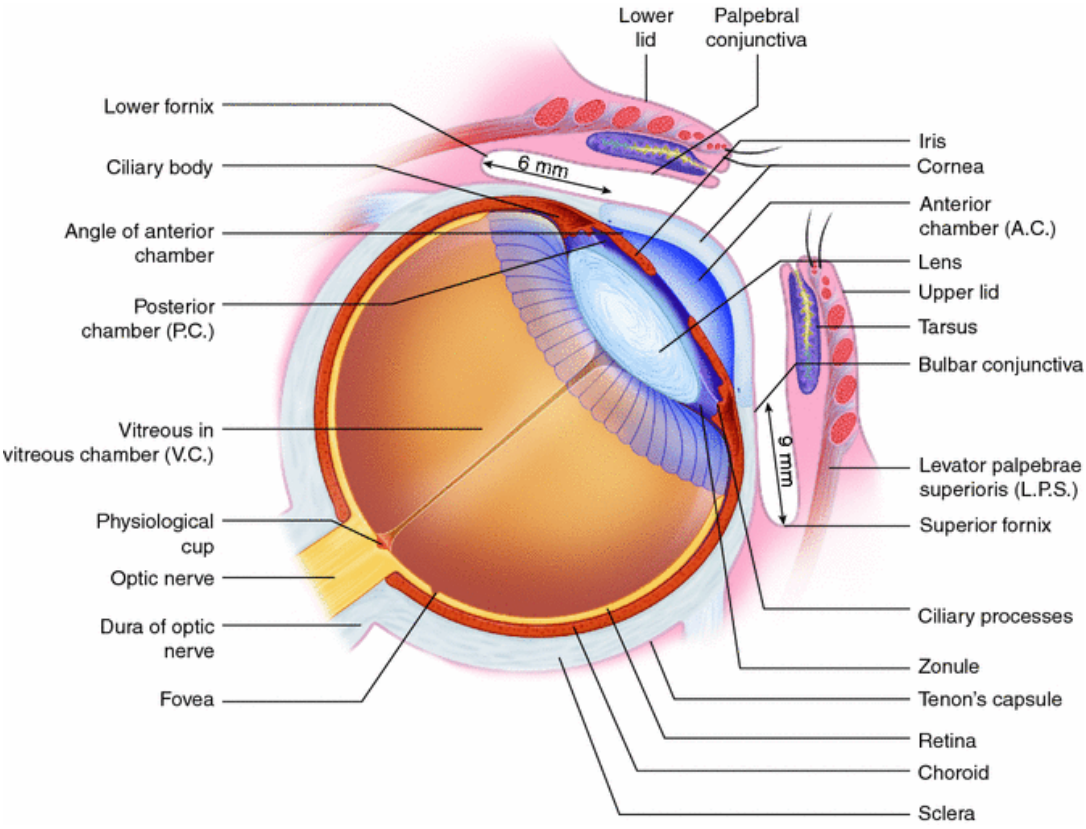


Figure 3.3 Structure of the eye demonstrating the ciliary zonules which regulate diameter of the lens. Reprinted from Ansari & Nadeem, (2016) with copyright permission.

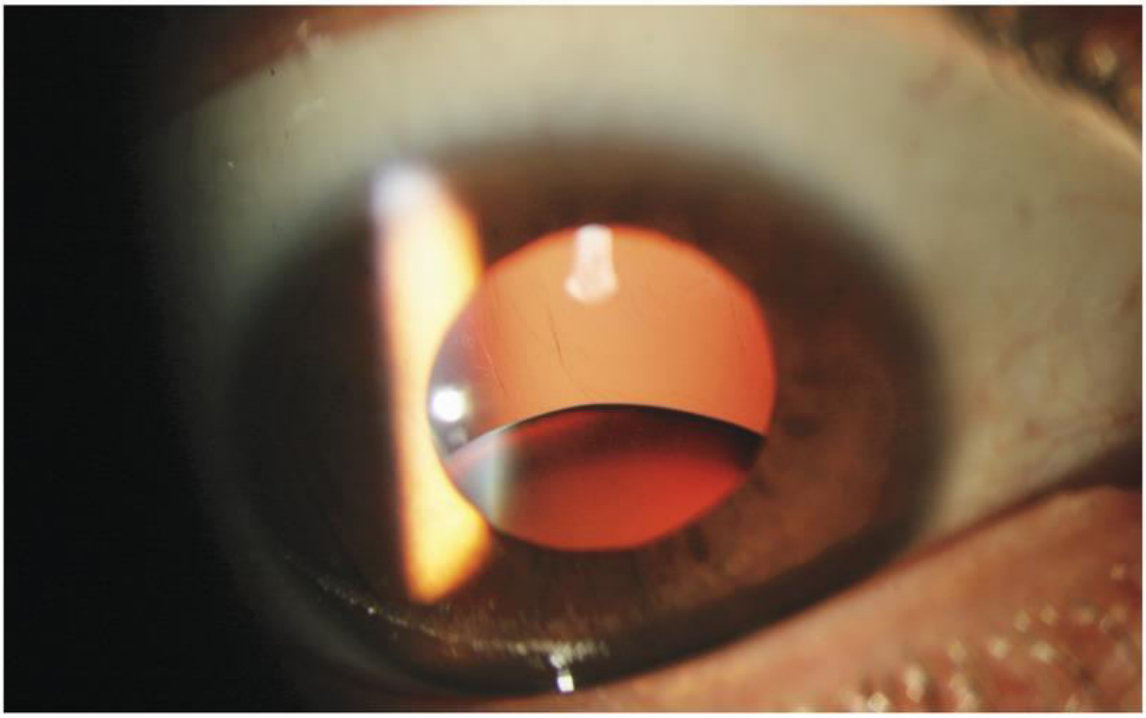


Figure 3.4 The clinical appearance of ectopia lentis viewed using slit lamp microscopy. Reproduced with permission from Sridhar and Chang, (2017), copyright Massachusetts Medical Society.

called fibrillinopathies, and they demonstrate the importance of fibrillin-1 proteins in connective tissue stability. The fibrillinopathy phenotypes show both key differences as well as overlapping features. Broadly, fibrillinoapthies can cause a tall stature phenotype or a short stature phenotype (Figure 3.5). Whereas Marfan Syndrome is a fibrillinopathy manifesting with increased height and disproportionately long limbs or digits (arachnodactyly) and joint laxity, the acromelic dysplasias (WMS, Geleophysic dysplasia & Acromicric dysplasia) and Myhre syndrome are united by their short stature, joint stiffness, and short stubby hands and feet (brachydactyly) (Figure 3.5).

WMS (formerly spherophakia-brachymorphia syndrome) is a fibrillinopathy defined by short stature, muscular build, skin stiffness, brachydactyly, microspherophakia, ectopia lentis, severe myopia and risk for glaucoma. WMS was initially defined by Georges Weill (Weill, 1932) and later redefined by Oswald Marchesani (Marchesani, 1939). The prevalence is estimated to be 1 in 100,000. Figure 3.6 demonstrates a 19-year old male WMS patient with the typical muscular build, short stature and brachydactyly – joint stiffness is evident by pallor around the knuckles while extending the hand. Meanwhile, Figure 3.7 shows the small, rounded lens characteristic of microspherophakia. As WMS is inherited in autosomal dominant and autosomal recessive forms, Faivre and colleagues studied a cohort of WMS patients published in the literature to identify any clinical evidence to distinguish between dominant or recessive forms (Faivre *et al.,* 2003). They reviewed published literature (from 1932-2001) of 128 WMS patients, consisting of

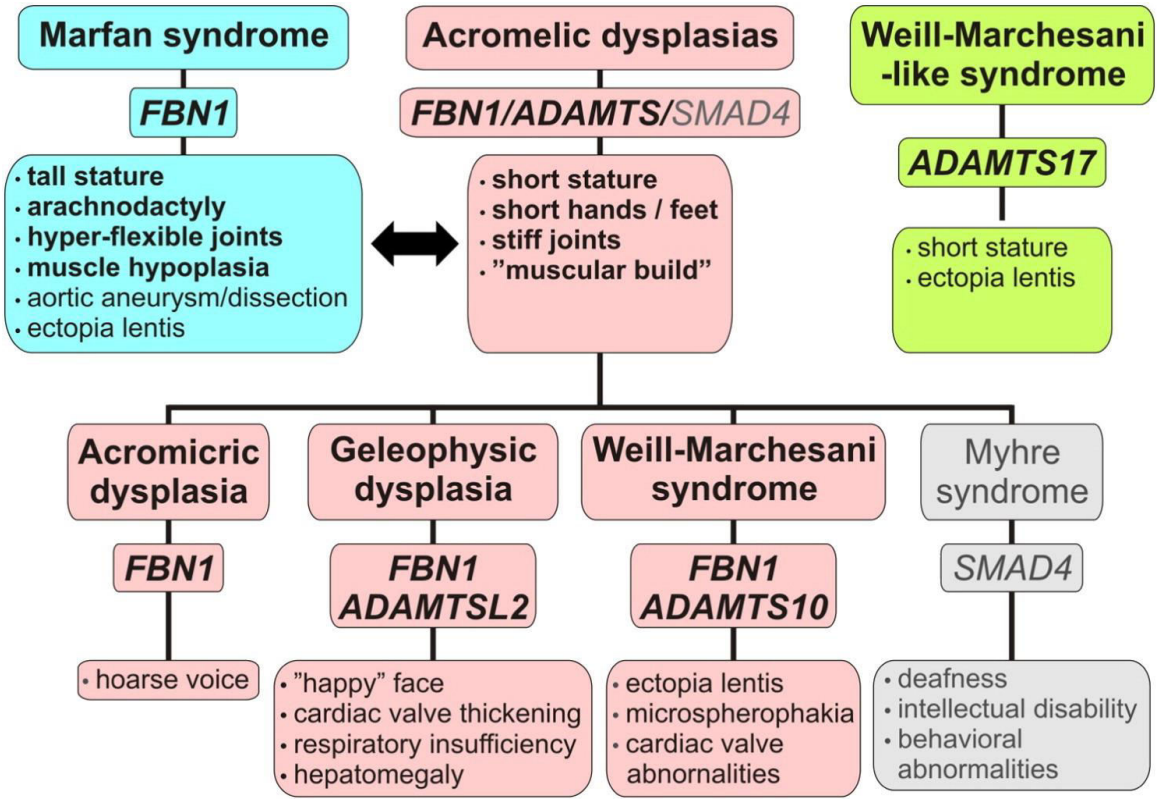
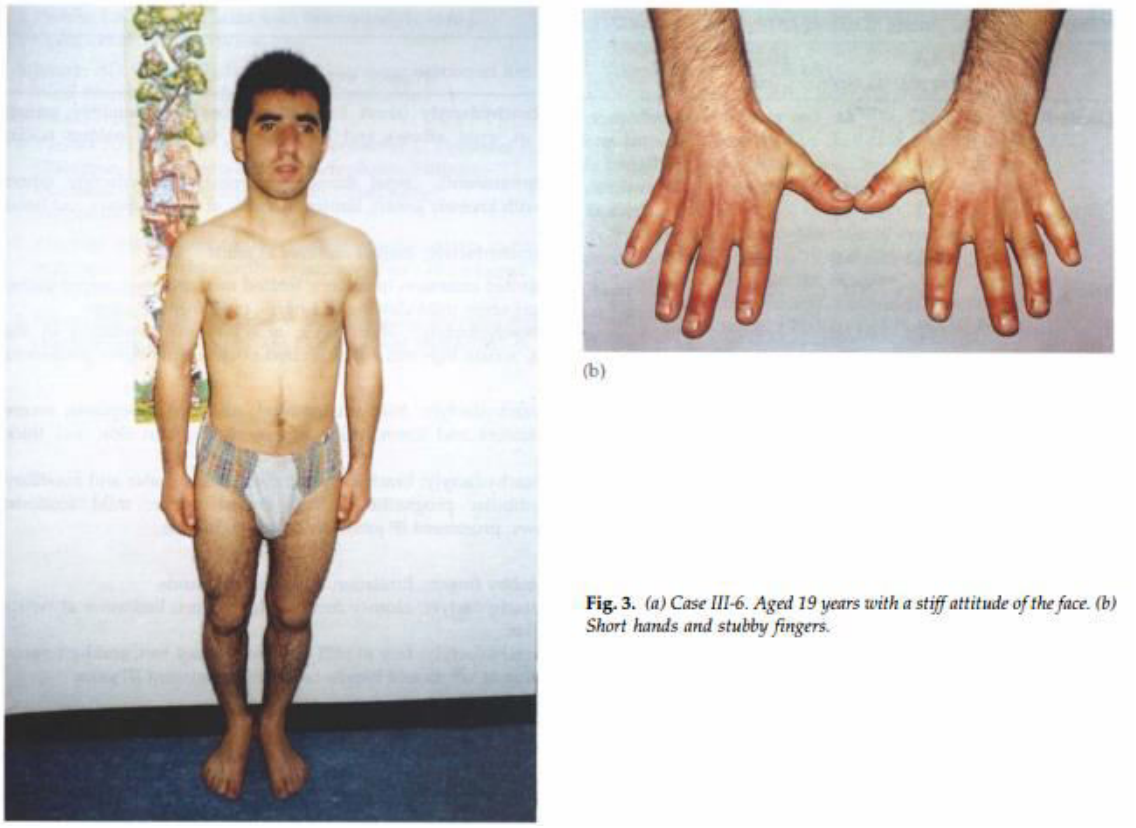


Figure 3.5 The clinical features differentiating several fibrillinpathies and their causal genes. Reprinted from Hubmacher & Apte (2015) with copyright permission.



(b)

Fig. 3. *(a) Case III-6. Aged 19 years with a stiff attitude of the face. (b) Short hands and stubby fingers.*

Figure 3.6 Clinical manifestations of WMS. A short stature with muscular build and brachydactyly seen in this patient with WMS. Reprinted from Evereklioglu *et al.,* (1999) with copyright permission.

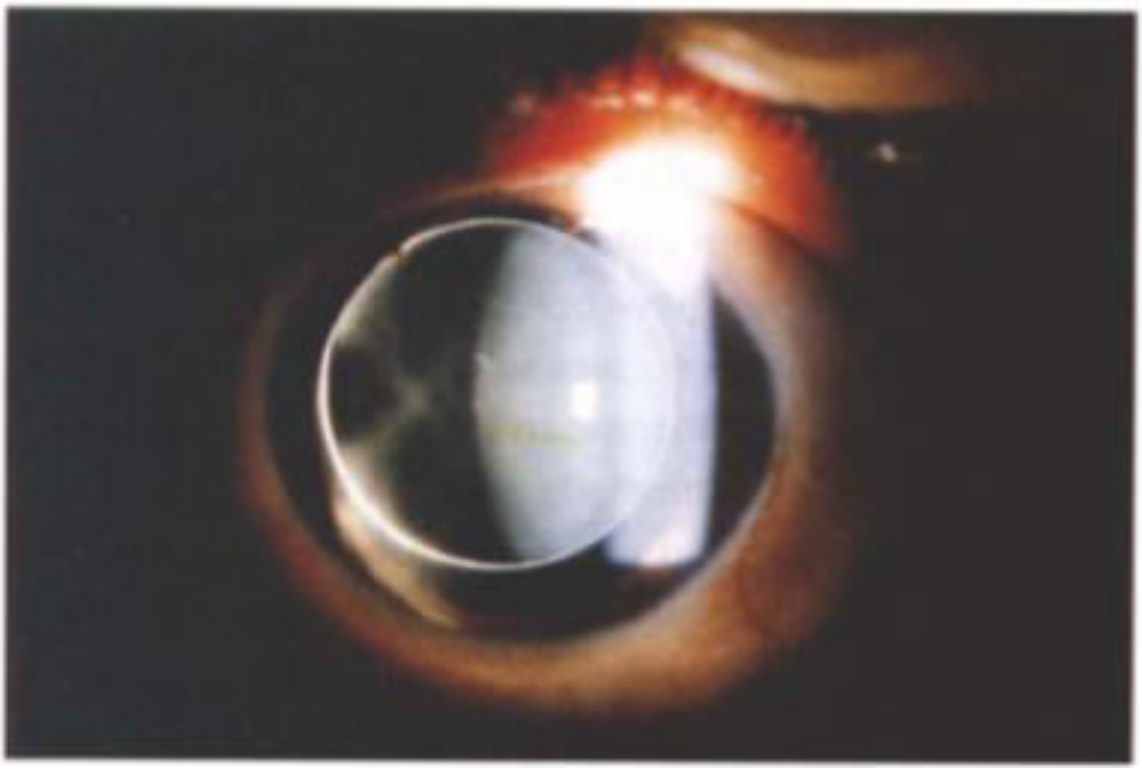


Figure 3.7 Microspherophakia in a lens of a patient with WMS with subluxation also seen. Reprinted from (Evereklioglu *et al.,* (1999) with copyright permission.

30 autosomal recessive, 12 autosomal dominant and 21 sporadic families. The cohort included 58 males and 56 females with mean age of 24 years. All autosomal dominant WMS patients had short stature (49/49), and nearly all autosomal recessive patients (55/56) manifested short stature too. Similarly, all autosomal dominant cases manifested brachydactyly (49/49), as did 96% of autosomal recessive cases (54/56). Nearly all autosomal recessive patients 96% (51/53) had myopia, while 88% of autosomal dominant (38/43) cases with available data suggested myopia. A majority of autosomal recessive and autosomal dominant cases developed glaucoma-- 75% and 80% respectively. Meanwhile, cataracts were much less common 21% (AR) and 28% (AD) respectively. Neither pattern of inheritance could be distinguished by clinical findings alone, but there were some significant differences in the data. Faivre and colleagues noted statistically significant differences in microspherophakia, ectopia lentis, joint limitation and cardiac anomalies in autosomal recessive verses autosomal dominant cases (Table 3.1).

In terms of the genetics of WMS, autosomal dominant forms are caused by *FBN1* mutations, while recessive forms are caused by *LTBP2*, *ADAMTS10* and *ADAMTS17*. *FBN1* was the first disease gene mapped causing WMS. Wirtz *et al*. studied two extensive families with three generations of affected individuals, and identified linkage to chromosome 15q21.1 (Wirtz *et al.,* 1996). Given the clinical similarities between Marfan Syndrome and WMS (ectopia lentis, microspherophakia, risk for thoracic aneurysms), they proposed *FBN1* was a candidate gene for WMS. This prompted Faivre *et al*. to investigate two autosomal

Table 3.1 The clinical manifestations of 128 WMS cases published in the literature. Reprinted from Faivre *et al.,* (2003) with copyright permission.

	Total	AR	AD	Sporadic	Fischer
Number of families	63	30	12	21	
Number of patients	128	57	50	21	
Mean age (years)	24	28	21.5	26.5	
Sex	58M/56F				
Ophthalmologic examination					
Microspherophakia	95/113 (84%)	49/52 (94%)	32/43 (74%)	14/18 (78%)	0.007
Ectopia lentis	92/126 (73%)	35/55 (64%)	42/50 (84%)	15/21 (71%)	0.016
Myopia	110/117 (94%)	51/53 (96%)	38/43 (88%)	21/21 (100%)	NS
Glaucoma	95/119 (80%)	38/51 (75%)	44/50 (80%)	13/18 (72%)	NS
Cataract	24/103 (23%)	11/53 (21%)	9/32 (28%)	4/18 (22%)	NS
Other features					
Short stature	124/126 (98%)	55/56 (98%)	49/49 (100%)	20/21 (95%)	NS
Brachydactyly	124/126 (98%)	54/56 (96%)	49/49 (100%)	20/20 (100%)	NS
Joint limitations[a]	54/87 (62%)	18/37 (49%)	30/39 (77%)	6/11 (55%)	0.010
Thick skin[a]	55/81 (68%)	23/35 (66%)	28/38 (74%)	4/8 (50%)	NS
Muscular build[a]	46/68 (68%)	21/30 (70%)	17/25 (68%)	8/13 (62%)	NS
Cardiac anomalies[a]	24/99 (24%)	15/38 (39%)	6/48 (13%)	3/13 (23%)	0.004
Mental retardation	16/128 (13%)	6/57 (11%)	9/50 (18%)	1/21 (5%)	NS

Cardiac abnormalities include pulmonary valve stenosis, mitral valve insufficiency, aortic valve stenosis, ductus arteriosus, and ventricular septal defect. F, female; M, male; AR, autosomal recessive; AD, autosomal dominant.
[a]These numbers and percentages should be considered with caution due to insufficient clinical description in some reports.

dominant WMS families – one being the previous family with linkage to chromosome 15 – by sequencing all 65 exons of *FBN1* (Faivre *et al.,* 2003b). In the family with linkage to chromosome 15, the authors identified a heterozygous 24 nucleotide deletion in exon 41 – an in frame deletion (p. 5074_5097del) affecting the latent TGF beta binding domain of fibrillin-1. The next WMS gene identified was *ADAMTS10*. Faivre *et al*. studied two large families with autosomal recessive WMS from Lebanon and Saudi Arabia. They found strong evidence of linkage (LOD 5.99) to chromosome 19p13.3-p13.2 – a 12.4 cM interval (Faivre *et al.,* 2002). Dagoneau and colleagues examined the over 100 positional candidate genes located within this interval (Dagoneau *et al.,* 2004). On the basis of function within the ECM, they selected five candidate genes for sequencing – *COL5A3*, *ICAM1*, *SMARCA4*, *TUBB5* and *ADAMTS10*. COL5A3 revealed no candidate pathogenic variants, however sequencing all 24 exons of *ADAMTS10* identified 3 separate variants: two splice site variants (c.1190+1G>A and c.810+1G>A) and the nonsense variant (p.R237X). The family from Lebanon was homozygous for (p.R237X), while the family from Saudi Arabia was homozygous for the c.1190+1G>A spice site variant, predicted to induce a frameshift. Finally, the sporadic individual was a compound heterozygote for c.810+1G>A and p.R237X.

In 2012, Haji-Seyed-Javadi proposed *LTBP2* as a novel candidate gene for autosomal recessive WMS (Haji-Seyed-Javadi *et al.,* 2012). They screened 30 unrelated Iranian probands with ectopia lentis – 4 patients attributable to WMS, 13 to Marfan Syndrome and 13 isolated ectopia lentis cases. They identified candidate

pathogenic variants in two unrelated families. One family segregated p.Val1177Met in *LTBP2* in a manner consistent with autosomal recessive inheritance, while the second family segregated a candidate heterozygous variant in *LTBP2* (p.Arg548*) which was digenically inherited with a *FBN1* variant (p.Cys129Gly). The significance of this inheritance pattern in the second family is unknown.

Finally, pathogenic variants in *ADAMTS17* were first described in 2009 (Morales *et al.,* 2009). Morales *et al*. investigated 13 patients belonged to seven unrelated WMS families – all from Saudi Arabia. Screening *ADAMTS10* and *FBN1* in the cohort identified two different homozygous *ADAMTS10* pathogenic variants (p.G518D, p.G700C) each in one unrelated family, while no causal *FBN1* variants were identified. They proceeded with another unsolved autosomal recessive WMS family called 'Family A'. Multipoint linkage analysis identified a disease locus on 15q26.3 (LOD 3). The critical interval encompassed 2.05 Mb, and the authors selected 6 positional candidate genes based on gene function (*IGR1R, SYNM, TTC23, MEF2A, LYSMD4, ADAMTS17*). They identified c.2458_2459insG (p.E820GfsX23), a homozygous insertion in exon 18 of *ADAMTS17,* segregating as expected for an autosomal recessive pattern of inheritance. Next they sequenced ADAMTS17 in additional cases, and identified a homozygous truncating variant (p.Q254X) in an additional family (Family E), and a homozygous splice site variant c.1721+1 G>A, in a third family, a sporadic case (WMS-D1). Clinical features and pedigrees for these WMS families are shown in Figure 3.8. An interesting observation among these families, all from the Saudi Arabian peninsula, was that each had only partial

features of WMS – short stature, ectopia lentis and microspherophakia, and they all lacked brachydactyly and joint stiffness. This led Morales and colleagues to propose that *ADAMTS17* variants cause a sub-phenotype which could be referred to as 'WMS-like' syndrome. Other studies then identified other pathogenic *ADAMTS17* variants. In 2012, Khan and colleagues studied two siblings from one consanguineous family, also from Saudi Arabia (Khan *et al.,* 2012). Homozygosity mapping in the two siblings identified *ADAMTS17* as the most likely shared disease-causing gene. Sequencing *ADAMTS17* identified a novel homozygous pathogenic variant (p.Asp218Thr*fs*X41) in each sibling. Both siblings were short of stature, had spherophakia and no joint stiffness or brachydactyly; consistent with the previous findings of Morales *et al* (2009). Radner *et al* then investigated 3 Tunisian families with WMS features (with brachydactyly and joint stiffness) as well as features of congenital ichthyosis (Radner *et al.,* 2013). Using genotyping and homozygosity mapping, they identified a large homozygous genomic deletion encompassing 100kb. The deletion encompassed all of *FLJ42289*, and exon 13 of *CERS3*, while deleting all of the 5'UTR and first three exons of *ADAMTS17*. Finally, Shah and colleagues studied one consanguineous autosomal recessive WMS family from India (Shah *et al.,* 2014). Whole exome sequencing and a custom filtering protocol for homozygous variants identified a homozygous splice variant in ADAMTS17 (c.873+1 G>T). The pathogenic variant causes skipping of exon 5.

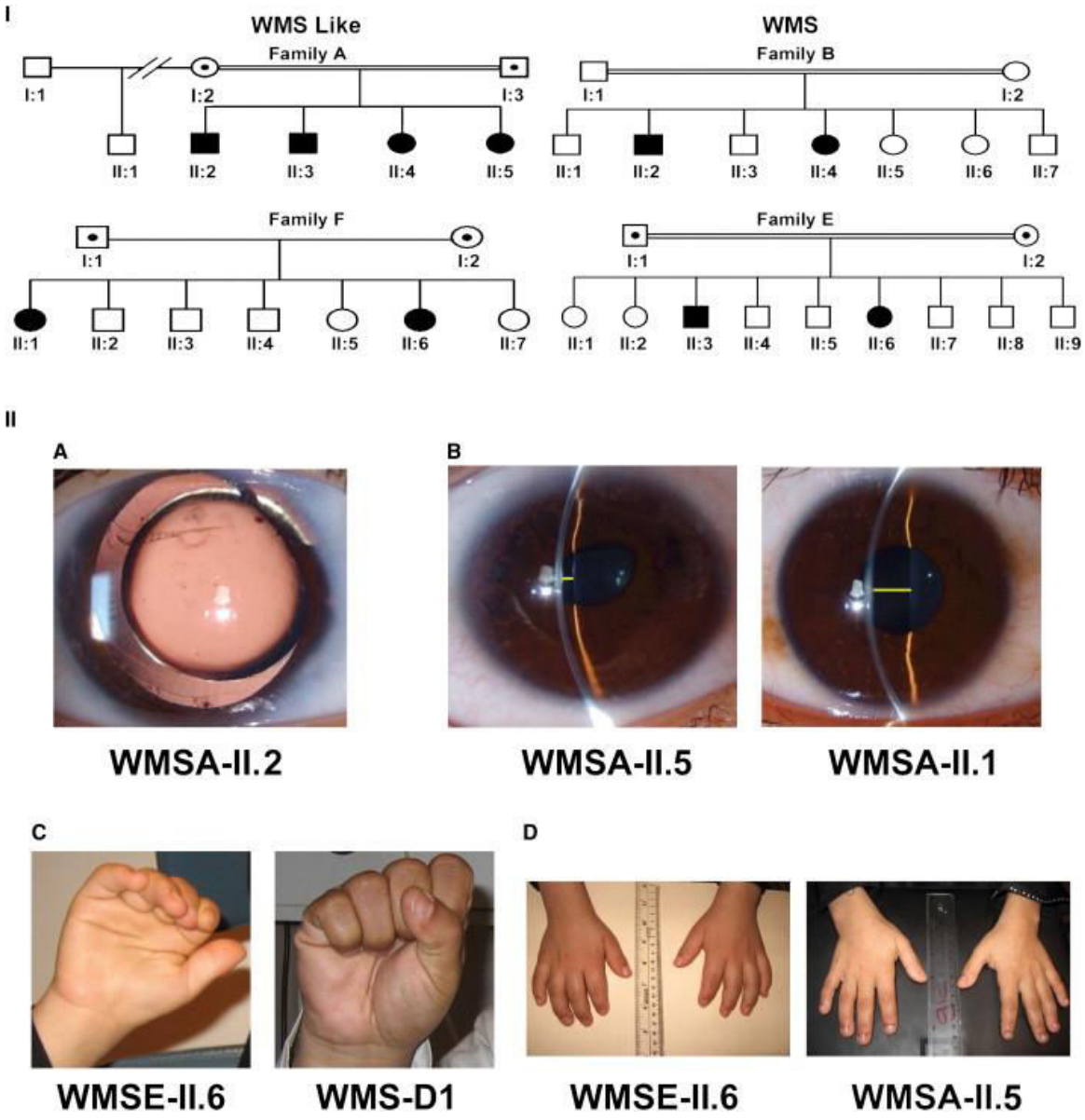


Figure 3.8 Family structures and clinical manifestation of WMS-like families identified by Morales *et al.,* (2009). Reprinted from (Morales *et al.,* 2009) with copyright permission.

3.2 Study Aims

In our study, we hypothesized that a pathogenic variant segregated in a family with autosomal recessive WMS from Newfoundland. The study aim was to uncover the causal pathogenic variant(s) to provide a genetic diagnosis and gain insight into the molecular pathogenesis of WMS in this family. The primary study objective, therefore, was to explore and interpret whole exome and homozygosity mapping data in this family. To prove causality, the next objective was to validate, and test segregation of the candidate variant in the family. The final objective was to characterize the pathogenic variant, and perform a comprehensive clinical review to assist clinicians and researchers in understanding the manifestation of this rare disorder.

3.3 Author Contributions

The following is original research in preparation for submission to an academic journal. As first author of the study, I interpreted candidate variants in the whole exome sequencing data. I then validated candidate variants in the family and tested population controls. I reviewed and interpreted clinical information in the family, and prepared the manuscript.

JSG conceived of the study, assisted in manuscript revision and collected patient records and consent. KMB, JM and SF aided in the collection and analysis of high throughput sequencing data. Bridget A. Fernandez (BAF) assisted in the collection of radiological scans and interpretation of the clinical data. Matthew A.

Deardorff (MAD) assisted in the measurement and interpretation of radiological scans. GJJ and JHW performed ophthalmic examination of individuals and assisted with clinical interpretation of disease. Dirk Hubmacher (DH) and Suneel Apte (SA) performed cell culture and protein assay of the candidate *ADAMTS17* variant. MOW participated in the study design, supervised the collection of molecular and genetic data, assisted in the analysis of genetic data and helped draft the manuscript. All authors read and approved the final manuscript.

Co-authors had the following affiliations during the time of this work. MOW, JSG and BAF were affiliated with the Discipline of Genetics, Memorial University of Newfoundland, Faculty of Medicine, St. Johns, NL, A1B 3V6 Canada. SF and JM were each affiliated with the Department of Human Genetics, McGill University, Montreal, QC, H3A 1B1 Canada as well as McGill University and Genome Quebec Innovation Centre, Montreal, QC, H3A 0G1 Canada. MAD was affiliated with the Division of Genetics, Children's Hospital of Philadelphia, Department of Pediatrics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania, PA 19104 U.S.A. GJJ was affiliated under Care of JSG, Discipline of Genetics, Memorial University of Newfoundland, Faculty of Medicine, St. Johns, NL, A1B 3V6, Canada. JHW was affiliated with Memorial University of Newfoundland, Faculty of Medicine, Discipline of Surgery (Ophthalmology) St. Johns, NL, A1B 3V6 Canada. DH was affiliated with Orthopaedic Research Laboratories, Leni and Peter W. May Department of Orthopaedics, Icahn School of Medicine at Mt. Sinai, New York, NY, 10029, USA. SA was affiliated with Department of Biomedical Engineering, Cleveland

Clinic Lerner Research Institute, Cleveland, OH, 44195, USA.

3.4 Abstract

WMS is a rare hereditary connective tissue disorder manifesting with short stature, brachydactyly and limitation of joint movement, as well as ocular features that prominently include microspherophakia and ectopia lentis. We identified WMS clinically in a large kindred from Newfoundland & Labrador, Canada. Therefore, the genetic etiology of WMS in five affected individuals of this apparent autosomal recessive family was investigated. Whole exome sequencing and homozygosity mapping revealed a novel homozygous missense variant in *ADAMTS17* (c. 3068G>A, p. C1023Y) which co-segregated with WMS in the family. The pathogenic variant is absent in 150 ethnically matched controls, and resides in a Thrombospondin-1 repeat domain of ADAMTS17. Transfection of mutant ADAMTS17 (p.C1023Y) expression plasmids into HEK293T cells revealed significantly reduced secretion of ADAMTS17 into the extracellular matrix, supporting pathogenicity of this variant. We characterized the clinical presentation in our WMS patients, with longitudinal data spanning more than 40 years. Assessment of metacarpophalangeal bones demonstrated variable dysmorphia in the hands of two patients examined. This work expands and clarifies the phenotype of this rare disorder, further delineating the genetic etiology of WMS caused by *ADAMTS17* mutations, and exploring the subsequent phenotype provided by this unique family.

3.5 Introduction

Genetic aberrations in fibrillin-1 (*FBN1*) (OMIM *134797) demonstrate significant clinical heterogeneity, with eight OMIM phenotypes (i.e. fibrillinopathies) currently known. Marfan syndrome (OMIM #154700), a well-known example; is a hereditary connective tissue disorder involving increased height and disproportionately long limbs/digits, as well as other clinical features (Dietz, 1993). Interestingly, WMS is a fibrillinopathy that displays the opposite phenotype (i.e. short stature and short digits) with overlapping clinical features (i.e ectopia lentis).

WMS is a rare disorder, as the prevalence is estimated to be about 1 in 100,000 (Tsilou and MacDonald, 1993). One Danish national survey of congenital ectopia lentis (ECL) observed a 6.4 per 100,000 prevalence of ECL, with readily identifiable WMS patients accounting for only 0.7% of cases (Fuchs and Rosenberg, 1998). While there are no absolute diagnostic criteria for WMS, the key manifestations are well described. These include systemic features: short stature, thick skin, joint stiffness, short digits (brachydactyly); as well as ocular findings: microspherophakia, ectopia lentis, severe myopia and glaucoma (Le Goff *et al.,* 2011). Occasionally, WMS patients develop pulmonary or aortic valve stenosis. Typically, an ophthalmologist ascertains WMS patients after noticing early onset myopia and a microspherophakic lens and/or ectopia lentis -- in patients with short stature. Management focuses on the consequent myopia, glaucoma and cataracts that may result.

As there is risk for secondary angle-closure glaucoma due to subluxation of

the lens in WMS patients, management of intraocular pressures is especially important in mitigating risk of damage to the optic nerve, and subsequent blindness. Certainly, one cohort of patients (*N*=80) ascertained for microspherophakia (including 18 WMS patients) revealed that blindness due to glaucoma approached 20-30% in this group (Senthil *et al.,* 2014). Measurement of intraocular pressures is also complicated by increased corneal thickness in some WMS patients, which should be taken into account (Razeghinejad *et al.,* 2009). Medical treatments can be ineffective in treating rising intraocular pressure for WMS patients (Taylor, 1996). Surgical treatments for WMS patients seek to relieve pupillary block, and have included iridectomy, iridotomy, trabeculotomy, anterior vitreotomy and lensectomy (Guo *et al*., 2015; Halpert & BenEzra, 1996; Harasymowycz & Wilson, 2004; Yang *et al.,* 2016).

WMS is inherited in autosomal dominant or autosomal recessive forms. Faivre et al. reviewed 128 WMS patients reported in the literature and found that the genetic etiology of WMS patients cannot be differentiated based on clinical findings alone (Faivre *et al.,* 2003). Variable expressivity of clinical manifestations is also seen in siblings or among extended family members with WMS (Evereklioglu *et al.,* 1999; Kloepfer and Rosenthal, 1955). Identifying the genetic cause for WMS is a critical and worthwhile endeavor. For example, a recent case report highlights the diagnostic odyssey of a 30 year old patient, who was being treated for glaucoma, when her diagnosis of WMS only became apparent after microspherophakia was identified following trabeculotomy (Guo *et al.,* 2015). Securing a diagnosis early is

beneficial for these patients, to monitor their intraocular pressures closely and hopefully mitigate optic nerve damage, prognosticate on patient outcomes, and to facilitate family planning. Investigation of the etiology of WMS is also integral to scientific discoveries. Notably, genotype-phenotype correlations of WMS patients have previously led to stunning insights into molecular mechanisms of the extracellular matrix (Faivre *et al.,* 2003; Hubmacher & Apte, 2011).

WMS is genetically heterogeneous and caused by highly penetrant pathogenic variants. Four genes currently explain WMS. These include *FBN1*, *ADAMTS10*, and *ADAMTS17. LTBP2* has also been identified as a candidate gene based on homozygosity mapping. When caused by *FBN1*, WMS segregates in an autosomal dominant pattern (OMIM #608328) (Faivre *et al.,* 2003), while WMS caused by *ADAMTS10* and *ADAMTS17* and *LTBP2* segregate in autosomal recessive form (OMIM #277600) (Dagoneau *et al.,* 2004; Haji-Seyed-Javadi *et al.,* 2012; Morales *et al.,* 2009).

In this study, we investigate the genetic etiology of autosomal recessive WMS in a large family from Newfoundland, Canada, which includes five affected siblings (Figure 3.9). We explore extensive longitudinal clinical data available for this family, and examine variable metacarpophalangeal deformities present in siblings. Finally, we detail the discovery of a novel germline pathogenic missense variant, localizing

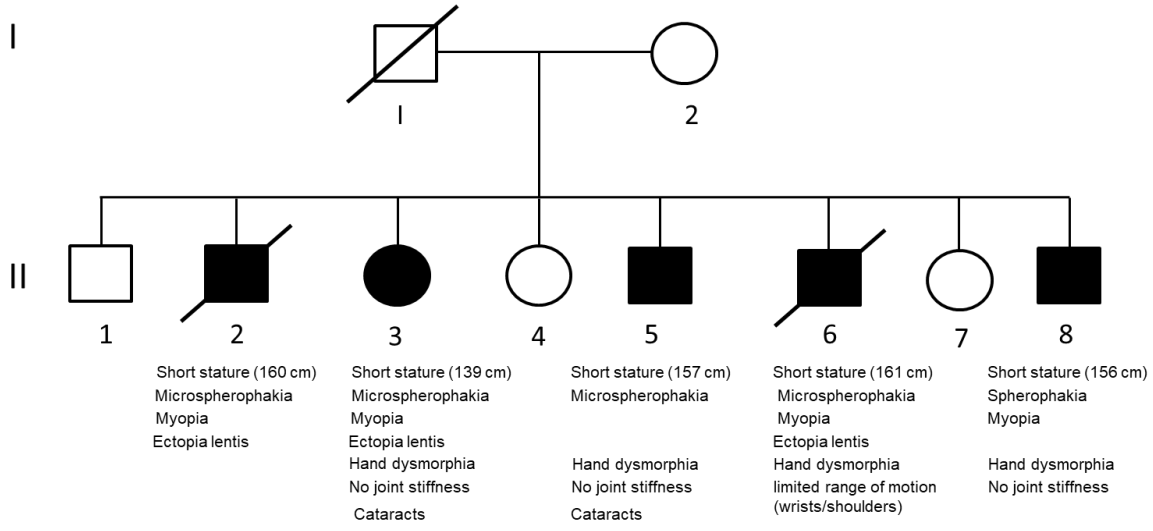


Figure 3.9 The clinical manifestations of an autosomal recessive WMS family living in Newfoundland. Figure created by the author.

This large family with apparent autosomal recessive inheritance consists of five affected individuals with variable presentation of WMS manifestations including short stature, myopia, microspherophakia, ectopia lentis, cataracts and variable dysmorphic hand features. The majority of affected individuals did not develop joint stiffness, while a single individual (II-6) reported limited range of motion in wrists and shoulders, and had a winged scapula.

to the thrombospondin-1 repeat domain of ADAMTS17, and have characterized its impact on secretion of ADAMTS17 into the extracellular matrix.

3.6 Materials and Methods

3.6.1 Patient Ascertainment

The family is from a small rural community on the island of Newfoundland, Canada. Patients were first seen at an Ophthalmology clinic prior to 1983 by GJJ, and were the subject of a case report in 1983 (Johnson and Bosanquet, 1983). Subsequently, they were referred to the Ocular Genetics Clinic and were followed by JSG. In 2007, family members gave written and informed consent to participate in an ocular genetics study. In 2012, they consented to participate in the FORGE Canada Consortium study to employ whole exome sequencing to their DNA.

3.6.2 Whole Exome Sequencing, Filtering and Homozygosity Mapping

Whole blood was drawn and genomic DNA was extracted from peripheral leukocytes by standard protocols. Whole exome sequencing (Illumina Hiseq 2000) was performed according to protocols at the Genome Quebec Innovation Center, Montreal, Canada (Fahiminiya *et al.,* 2014; McDonald-McGinn *et al.,* 2013). Briefly, 3 µg of genomic DNA from two affected individuals (Figure 3.9: II-3, II-5) was used for library preparation, and exome capture and enrichment was accomplished using the SureSelect Human Exome Kit V.4 (Agilent Technologies, Inc., Santa Clara, CA). Alignment, variant calling, and annotation were performed as per previous FORGE projects (Dyment *et al.,* 2013; Smith *et al.,* 2014). Mapping of high quality paired-end

reads (100 bp) against the UCSC Genome hg19 was performed using BWA (v. 0.5.9) (Li and Durbin, 2010). The Genome Analysis Toolkit (GATK) (McKenna *et al.,* 2010) was used for local realignment around indels (short insertions and deletions) and coverage assessment. Samtools (v. 0.1.17) mpileup (Li *et al.,* 2009) and ANNOVAR (Wang *et al.,* 2010) were then used for variant calling and annotation, respectively. A custom filtering protocol was applied to the whole exome data and sequence variants were excluded if there were i) < 3 reads supporting the alternative variant ii) there was variant allele frequency of > 0.05 in ExAC database and > 30 individuals with the variant in our database (~1000 exomes sequenced previously in our center), iii) variants were homozygous in ExAC database, iv) variants were synonymous, v) variants were within 5' UTRs or were intronic variants outside of splice-site boundaries.

Finally, homozygosity mapping was performed using in-house scripts written in Perl, to identify regions of homozygosity (ROH). An ROH was defined as a span of at least 25 SNPs where no more than two were called as heterozygous and the remainder were homozygous.

3.6.3 Validation of Candidate Variant, Segregation Analysis and Population Controls

To validate the candidate *ADAMTS17* variant, primers were designed using Primer3 (Untergasser *et al.,* 2012), capturing exon 21 (forward primer: TCCCTTGACCTCA; reverse primer: CACCGTCAGGGAG). PCR conditions are available upon request. Sanger sequencing of PCR products was accomplished using an ABI 3130XL Genetic

Analyzer, and chromatograms were interpreted using Sequencher 5.2.3.

In this manner, familial segregation was then tested in individuals I-2, II-3, II-4, II-5 and II-8 (Figure 3.9). Finally, a cohort of 150 geographically matched population controls, from a previous study (Green *et al.,* 2007) were then screened for the *ADAMTS17* variant.

3.6.4 Measurement of Metacarpophalangeal Lengths

Roentgenograms of both hands of an unaffected sibling (Figure 1: Individual II-4) and two affected siblings (Figure 1: II-5, II-3) were obtained in 2016. To examine for evidence of dysmorphia, measurements of metacarpal bones, and proximal, middle and distal phalanges were taken. These measurements were compared against their expected proportions (Poznanski, 1984), taking each person's age and sex into consideration.

3.6.5 Cloning

The expression plasmids for full-length ADAMTS17 and the proteolytically inactive mutant (ADAMTS17-EA) were described previously (Hubmacher *et al.,* 2017). To introduce a variant of interest for functional analysis, a synthetic DNA fragment spanning nucleotide 2799 – 3352 of *ADAMTS17* (Ref.Seq.XM_017021975) was ligated with the gel-purified 8335 bp SrfI x AfeI fragment obtained from ADAMTS17 and ADAMTS17-EA, using the NEBuilder® HiFi DNA Assembly Kit (New England Biolabs, Ipswich, MA). Positive clones were screened via an additional ScaI

restriction site introduced by the mutation and verified by DNA sequencing (Cleveland Clinic Lerner Research Institute Genomics Core).

3.6.6 Cell Culture, Transfection, and Collection of Serum-free Medium and Cell Lysate

Human embryonic kidney cells (HEK293T, ATCC, Manassas, VA) were maintained in DMEM supplemented with 10% FBS, 100 units/ml penicillin, and 100 μ g/ml streptomycin in a 5% CO2 atmosphere in a humidified incubator at 37 °C. HEK293T cells were seeded in 12-well plates (BD Bioscience, San Jose, CA) and transfected with 1 μg plasmid DNA using Lipofectamine 3000 according to the manufacturer's protocol, in triplicates. After 16-24h, the medium was removed and cells were rinsed with 1 ml PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na2KPO4, 1.8 mM KH2PO4) and incubated with 600 μl serum-free DMEM for 48h. Conditioned medium was harvested, cell debris removed by centrifugation, and equal volumes subjected to SDS-PAGE under reducing conditions. The cell layer was washed with 1 ml PBS and cells were lysed in 0.1% NP40, 0.01% SDS, and 0.05% Na-deoxycholate in PBS. Cell lysates were incubated for 5 min with rotation end-over-end and cleared by centrifugation (5 min, > 20,000 g, 4 °C)

3.6.7 Western blotting

Equal volumes of conditioned medium and cell lysate were subjected to 7.5% SDS-PAGE under reducing conditions. Proteins were transferred onto PVDF membranes

(Immobilon F, EMD Millipore, Billerica, MA) for 1.5 h at 70 V at 4 °C in 25 mM Tris, 192 mM glycine and 20% methanol buffer. The membrane was blocked with 5% (w/v) milk in TBS (10 mM Tris-HCl, pH 7.2, 0.15 mM NaCl) for 1 h at room temperature, and anti-Myc (9E10, Invitrogen, 1:500) was diluted in 5% (w/v) milk in TBS + 0.1% Tween 20 (TBST) and incubated overnight at 4 °C. Membranes were washed three times with TBST for 5 min at room temperature and incubated with fluorophore-labeled IRDye goat-anti-mouse secondary antibody (1:10,000) (LI-COR Biosciences, Lincoln, NE) in 5% milk in TBST + 0.01% SDS for 1 h at room temperature. Membranes were washed three times with TBST for 5 min, one time with TBS for 5 min at room temperature, and were scanned wet on an Odyssey CLx scanner (LI-COR Biosciences, Lincoln, NE). The fluorescent signal was quantified using the Image Studio software.

3.7 Results

WMS in this family segregated in an autosomal recessive pattern, with five of eight siblings affected (Figure 3.9). Their unaffected parents are from a small rural community and shared the same surname prior to marriage. Clinical features in affected patients were first described in a 1983 case report (Johnson and Bosanquet, 1983). The clinical data encompassing a 40-year period is provided in Table 3.2 and is summarized below.

The oldest affected individual (Figure 3.9: II-2) first presented with myopia at 13 years of age. He had iridodonesis with lenses displaced nasally and microspherophakia. Shallow anterior chambers were noted. Peripheral iridectomies

were recommended for this individual, but he died accidentally at age 30 before any treatment took place. Other family members were examined. His younger sister (Figure 3.9: II-3) had high myopia at age 10 with increasing intraocular pressures (IOP). She subsequently received bilateral peripheral iridectomies at age 28. She was short of stature and somewhat stocky; her approximate height was 149 cm in her early thirties. She did not report any difficulties with range of motion in her fingers or wrists. Visually, her hands appeared wider and stubbier than normal, with a lateral curvature of both fifth fingers at the distal interphalangeal joints. Like her older brother, she had microspherophakia with both lenses positioned only slightly nasally (subluxation). At age 48 she had a right retinal tear and detachment repair, and at ages 50 and 51 she had cataract extractions with placement of intraocular lenses (IOL). At 58 years old, she had extraction of a dislocated left IOL, pars plana vitrectomy and lens replacement with an anterior chamber IOL. Her visual acuity returned to R: $20/25^{+}$ and L: $20/40^{-}$ following surgery. There were no cardiac abnormalities reported. She did not report any new stiffness in her joints, specifically when asked about her ability to make a closed fist. Her younger brother (Figure 3.9: II-5) was also assessed. He developed a traumatic cataract from a penetrating right eye injury at age 2. He had microspherophakia identified by age 18, and subsequently had a left peripheral iridectomy due to increasing IOP at age 26. In his late twenties, his build was short and stocky, with an approximate height of 157 cm. He had a normal echocardiogram and no cardiac abnormalities. His hands and feet were broad, and fingers and toes appeared short and wide. He had a

shortened first metacarpal on his left hand. There was no restriction of joint mobility. He had cataract extraction and replacement with an anterior chamber IOL (age 49). Coincidentally, at age 53 he had colon cancer with several polyps, as did his mother. The next sibling was a brother (Figure 3.9: II-6), who had myopia and microspherophakia noted at age 14. Marked iridodonesis was observed, and due to increasing IOP, he had bilateral peripheral iridectomies at age 22. His approximate height was 161 cm in his early twenties. In 1982, hand x-rays demonstrated bilateral shortening of the fourth and fifth metacarpals. He died at age 43. Finally, the youngest brother (Figure 3.9: II-8) was first examined at age 10. He had myopia with spherical lenses. As his bilateral IOPs increased, he was first treated with timolol, and then had bilateral peripheral iridectomies at age 17. He has cognitive impairment (thought to be from birth trauma) and had a heart murmur as a child. He was diagnosed with pulmonary stenosis by a cardiologist (grade I/VI) at this time. A subsequent echocardiogram performed in 1982 indicated the pulmonary stenosis was not clinically significant. Roentgenograms in 1982 demonstrated shortening of the first, fourth and fifth metacarpals in both hands. He was 156 cm in his mid-to late teens. On physical examination, his hands were described as short and spade-like with stubby fingers, and with prominent knuckles and knobby interphalangeal joints. He had markedly decreased limitation of movement in wrists and shoulders, and a winged scapula.

To further explore evidence of variable hand dysmorphology in this family, we obtained new roentgenograms of the hands of two affected family members (II-3,

II-5) and compared them against their unaffected sibling (II-4) (Figure 3.10A). Visually, both hands of the affected siblings (II-5 and II-3) appear smaller compared to their unaffected sibling (II-4). The hands of sibling II-3 are broader and stubbier compared to sibling II-5. Metacarpophalangeal measurements were taken to allow comparison between affected and unaffected siblings, while adjusting for expected lengths given respective sexes and ages (Poznanski, 1984). Analysis of the measurements revealed key differences between affected and unaffected siblings (Figure 3.10B). For example, in the affected siblings, most metacarpophalangeal measurements were 2.0 standard deviations or more below expected values, as compared to their unaffected sibling, whose metacarpophalangeal measurements were mostly one standard deviation or less. She had relative bilateral shortening of the first proximal phalanx, with her right hand 1.5 standard deviations below, and left hand 1.8 standard deviations below normal. Hands in the affected siblings were overall smaller (2.0 standard deviations) and showed evidence of unilateral shortening. In II-5, the first metacarpal bone of the left hand was unilaterally shorter than expected by seven standard deviations. Moreover, his affected sister (II-3) also showed unilaterally shortened metacarpals, as she had shortening of her right first proximal phalanx, measuring 5.3 standard deviations below the mean. These intra-familial differences were interesting given the shared genetic etiology in this family.

Whole exome sequencing analysis and homozygosity mapping were then pursued. Germline DNAs of two affected family members (II-5 and II-3) were

Table 3.3 Longitudinal clinical features of autosomal WMS family from Newfoundland. Table created by the author.

Clinical Feature	II-2	II-3	II-5	II-6	II-8
Relationship	Oldest brother	Sister	Brother	Brother	Youngest Brother
Early myopia	Yes (13)	high myopia (10)	Yes (glasses in 1972)	yes (14)	Yes (10)
Microspherophakia	yes (13)	yes (10)	Yes (18)	yes (14)	spherophakia (10)
Ectopia lentis	Iridodonesis (13), both lenses displaced nasally (13)	both lenses displaced only slightly nasally (10), iridodonesis (2000)	Subluxation of left lens (2003)	Iridodonesis (14)	None
Increasing IOPs	yes (13)	yes (10)	yes (26)	yes (22)	Yes (10)
Glaucoma	No	BL (28) (1981)	Secondary glaucoma (1981)	No	No
Cataracts	No	Yes. Extractions (50, 51)	Traumatic cataract right eye (2), cataract extraction (L) (49)	No	No
Other eye features	No	right retinal tear and detachment (48),	Penetrating right eye injury (2)	No	No
Height	160cm	149 cm (~30s)	157 cm (late 20s)	161 cm	155.5 (mid to late teens)
Hand dysmorphia	NA	BL lateral curvature of 5th fingers, 2016: unilateral right first proximal phalanx shortened -5.3 SD	shortened first metacarpal on his left hand (1983? 26?) (2016: L first metacarpal shortened -7 SD)	relative BL shortening fourth and fifth metacarpals (1982)	BL shortening of the first, fourth and fifth metacarpals
Joint stiffness	No	None (30s)	None	limited movement wrists, shoulders	None
Cardiac defects	NR	NR	Normal echocardiogram	No	Murmur (as child), Grade I/IV (age), Pulmonary stenosis, echocardiogram reveals no significant stenosis (1982)
Treatments	Advised to have PI but died accidentally	Timolol, BL PI (28), Left lensectomy, pars plana vitrectomy and lens replacement with an anterior chamber IOL (58) (Timoptic and truspot BID for glaucoma 2000)	Timolol (1981), Left peripheral irridectomy (26), anterior chamber IOL (49)	BL PI (22) (1982)	Timolol ineffective, BL Pis (17) (1982)
Other systemic features	No	Endometriosis 1978-1979, L ovarian cystectomy (1979), total Hysterectomy (1980), BL salpingectomy with L oophorectomy	CRC and polyps (53)	Winged scapula	None
Notes	d.30	NA	NA	d. 43	Cognitive impairment (birth trauma)

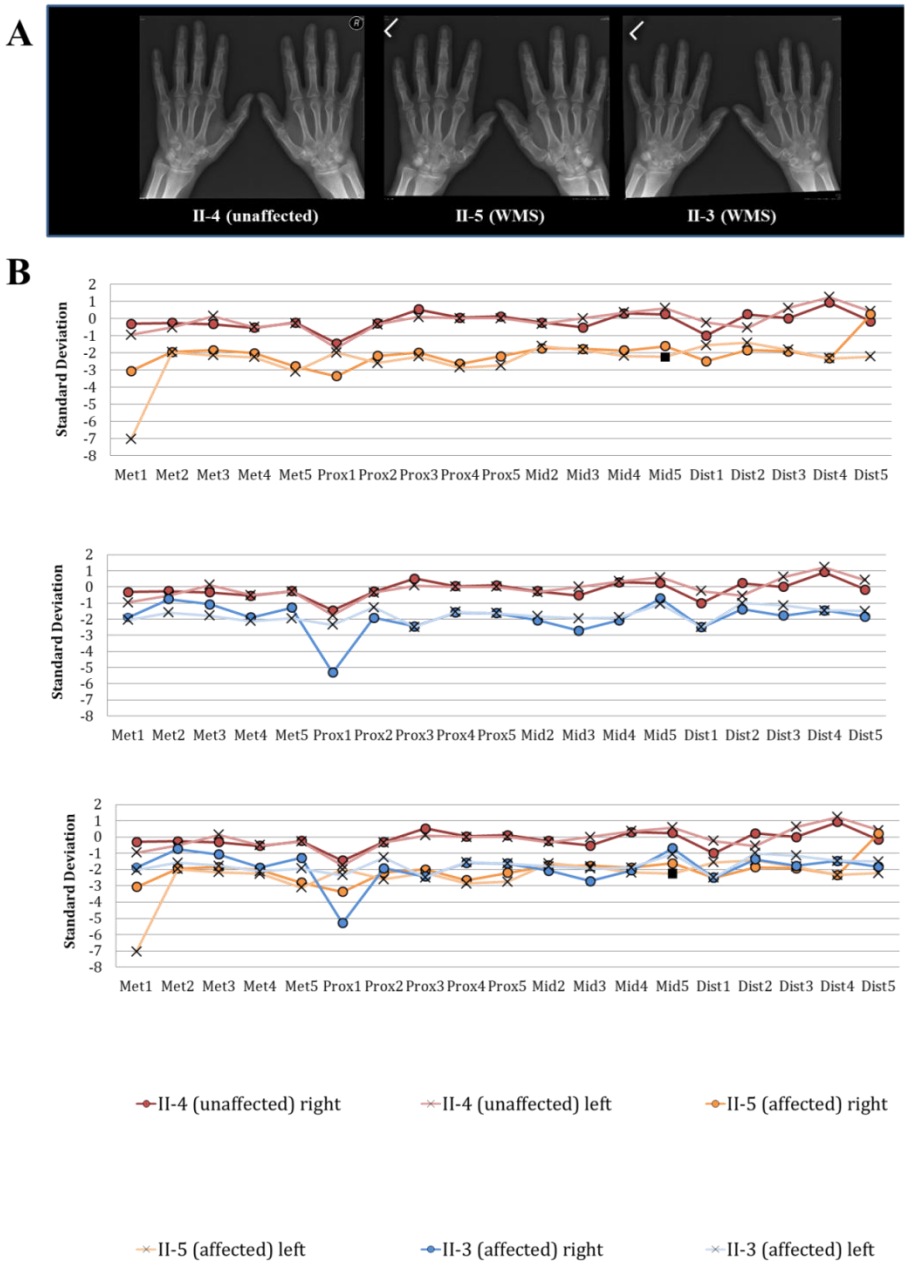


Figure 3.10 The variable metacarpophalangeal measurements between affected and unaffected siblings in a Newfoundland family with WMS. Figure created by the author.

Part A: Bilateral hand xray comparison of unaffected sibling (II-4; left panel) with two affected siblings (II-5; center panel, II-3; right panel) which demonstrates hands that visually appear small and stubbier compared to their unaffected sibling, which also display shortened metacarpal bones. Part B: Metacarpophalangeal measurements identify bilateral shortening of hands in affected patients compared to unaffected siblings and demonstrate the metacarpals which are much shorter than expected for their relative age and sex. II-5 displays a left 1st metacarpal bone that is markedly shortened, and while II-3 displayed a shorted right 1st proximal phalanx.

sequenced and achieved a mean coverage of 140X and 144X, respectively. Our variant filtering protocol yielded 262 (II-5) and 234 (II-3) rare coding variants, respectively. Of these variants, 120 were shared between both affected individuals. Homozygosity mapping revealed a homozygous missense variant (NM_139057.3: c.3068G>A: p.C1023Y) in *ADAMTS17*, being the only variant passing filtering criteria and also residing within a region of homozygosity (approximately 2.5 Mb in length), common to both affected siblings (Figure 3.11). The global minor allele frequency (MAF) of this variant (c.3068G>A: p.C1023Y) was extremely low -- as it has not reported in the two largest population databases available (GnomAD and ExAC browser), with one hit in the NHLBI Exome Variant Server database. The variant is heterozygous in 1 out of 4173 alleles (0.0237%) in an African American population, with combined MAF estimated at 0.0081% in the NHLBI database.

Familial co-segregation analysis corroborated a pathogenic role for this variant, as the variant segregated as expected for fully penetrant autosomal recessive inheritance (Figure 3.12). This variant was homozygous in affected family members (II-3, II-5, II-8), and was not carried in their unaffected sibling (II-4). Their mother (I-2) was heterozygous, and their father, now deceased, was an obligate carrier. Screening of 300 ethnically matched population control alleles from NL demonstrated this variant is not a common in this isolated population, given that no additional alleles were detected in controls.

Bioinformatics analysis identified that the missense *ADAMTS17* variant alters a cysteine residue in a region predicted to have high evolutionary conservation

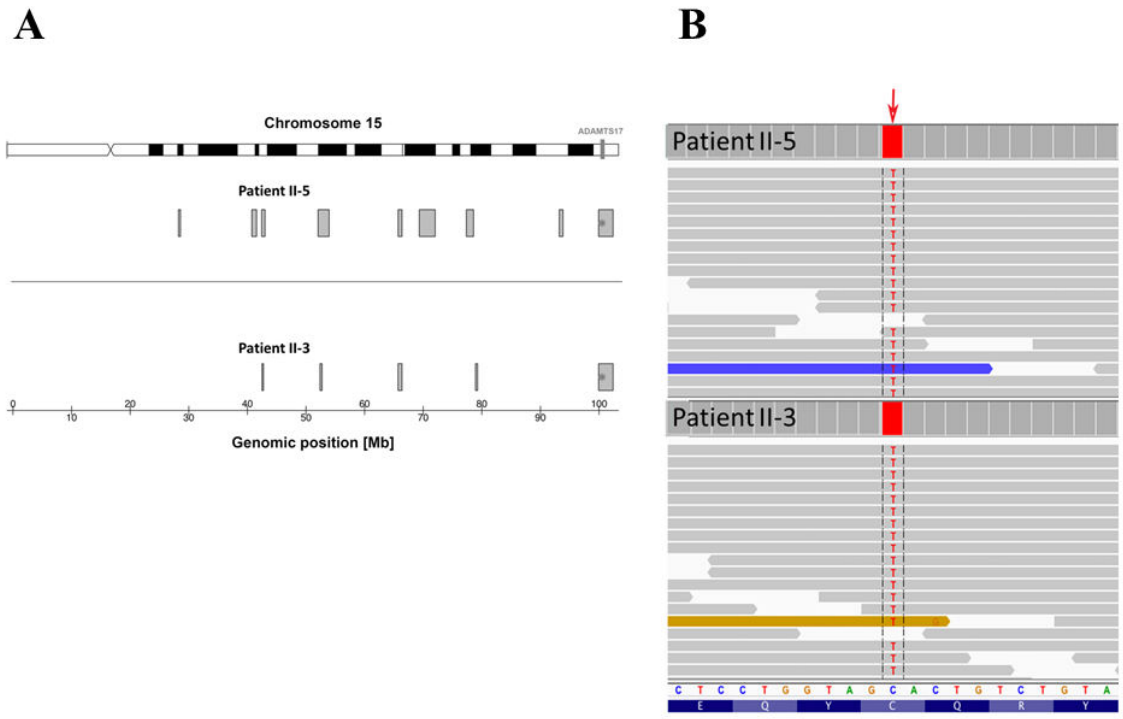


Figure 3.11 Whole exome sequencing and homozygosity mapping of affected WMS patients (II-5, II-3) identifies novel ADAMTS17 variant. Figure created by the author.

Part A demonstrates the homozygosity mapping on Chromosome 15. Gray bars indicate regions of homozygosity in each patient. Stars indicate the position of the homozygous (c. 3068G>A, p. C1023Y) variant in *ADAMTS17* in both patients which is located within an identical region of approximately 2.5Mb, shared in both patients. Part B shows sequence reads of the rare homozygous missense variant in *ADAMTS17*, which was the only variant identified by whole exome sequencing that was within the region of homozygosity.

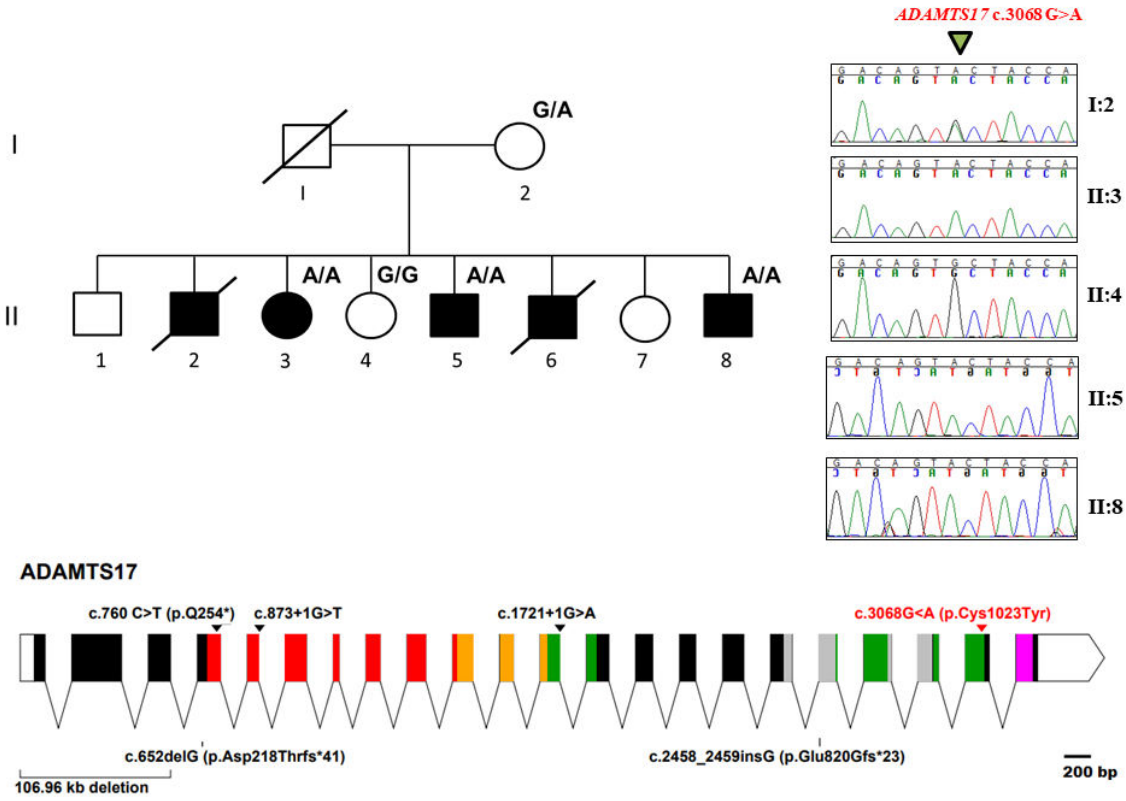


Figure 3.12 *In silico* analysis of the homozygous *ADAMTS17* variant (c. 3068G>A, p. C1023Y) in Newfoundland family and position of the variant within the ADAMTS17 gene compared to other previously reported pathogenic variants. Figure created by the author.

Segregation of p.C1023Y is shown in the pedigree at the top of the figure and the chromatogram traces on the right. The segregation of the variant fits with an autosomal recessive inheritance as all affected individuals are homozygous for the variant. Their unaffected sibling (II-4) was a non-carrier, while their unaffected mother (I-2) was a heterozygous carrier. Their father (I-1) was deceased at the time of the study, but did not manifest WMS features, and is assumed to be an obligate carrier. All currently reported *ADAMTS17* variants discovered in WMS patients are shown in the diagram at the bottom of the figure. The p.C1023Y variant is the first disease-causing missense variant identified in this gene date. Alternating colors demonstrate different domains in ADAMTS17. The p.C1023Y variant localizes to the 5th thrombospondin-1 repeat domain.

according to the GERP algorithm (score: 5.23). Additional *in silico* tools predicted highly deleterious consequences for this substitution. The SIFT algorithm predicted damaging effects (score: 0), as well as PolyPhen-2 (score= 1.0), MutationTaster (score: 0.99) and CADD (32) -- each indicating a highly deleterious effect. The variant encodes a residue that lies within a thrombospondin type 1 domain of ADAMTS17, a domain spanning amino acids 976-1028 of the polypeptide (Figure 3.12).

Finally, the impact of *ADAMTS17* p.C1023Y mutants on secretion of the enzyme into the extracellular matrix was observed in HEK293T cells (Figure 3.13). ADAMTS17 fluorescence intensity was measured in cell lysates compared to cell media (i.e. secreted protein). Because ADAMTS17 undergoes autoproteolysis, wild type constructs with (+) or without (-) the p.C1023Y mutation were compared to either wildtype ADAMTS17 or constructs concomittently carrying an active site mutation (p. E390A), which decreases autoproteolysis and increases the amount of fluorescent protein detectable. Figure 13.3 demonstrates that HEK293T cells (both wildtype or p.E390A constructs) with the p.C1023Y mutation (+) or without the mutation (-) have comparable expression levels in the cell lysate, as expected. For ADAMTS17 p.C1023Y (+) mutants in the cell medium however, there is a significant reduction of secreted protein in both wildtype (p= 0.0012) and p.E390A constructs (p= 0.00014), indicating that the p.C1023Y mutation results in a significant reduction of secretion compared to cells without the mutation.

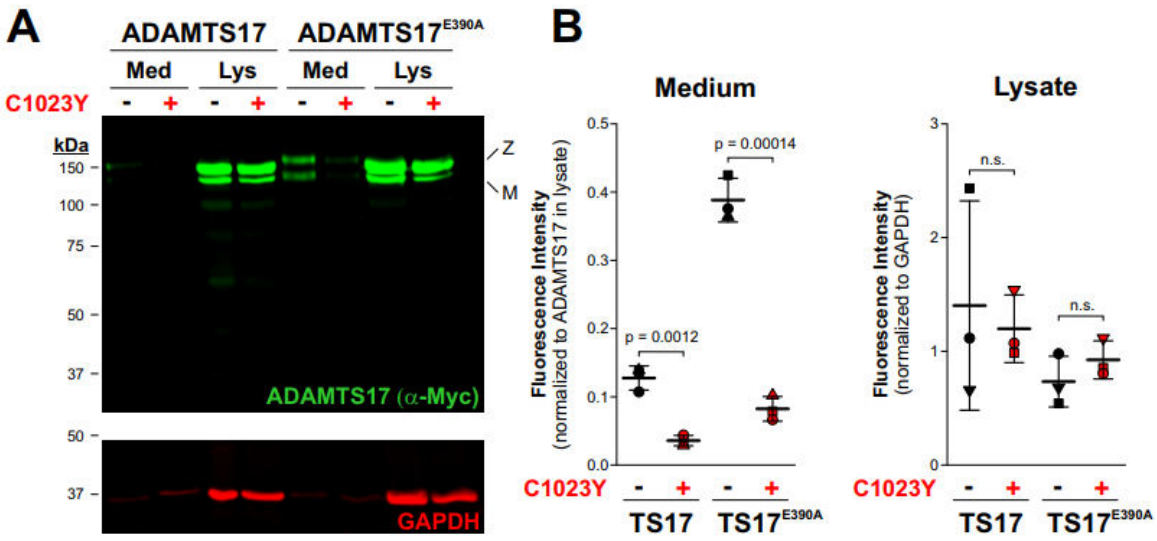


Figure 3.13 Fluorescence intensities of *ADAMTS17* p.C1023Y mutants (+) and wild type (-) in *ADAMTS17* (TS17) and active site mutant (TS17^{E390A}) constructs transfected in HEK293T cells. Figure created by the author.

Panel A shows fluorescence intensities for each construct in cell media verses lysate. There is a noticeable decrease in intensities in cell media in p.C1023Y mutants compared to wildtype in both constructs. No difference in intensities are seen in the cell lysate. Panel B demonstrates quantification of fluorescence intensities normalized to ADAMTS17 in the lysate. The decrease in intensities for ADAMTS17 in the p.C1023Y mutants is significant in the media (p=0.0012 in TS17 and p=0.00014 in TS17^{E390A}).

3.8 Discussion

Here, we conducted a clinical and molecular investigation of autosomal recessive WMS. The clinical features of this rare disorder in a large family including five affected individuals are summarized in Table 3.4. Family members were short of stature, with a mean height of 156.5 cm, and heights ranging from 149 cm to 161cm. All had myopia within a few years after their first decade of life. There was microspherophakia in four siblings, while the eyes of the youngest sibling (II-8) appeared spherophakic but not small. Ectopia lentis was seen in 4/5 siblings, with iridodonesis or partial subluxation seen in II-2, II-3 and II-6 by their early teens, while II-5 experienced a left lens subluxation at age 48. The youngest did not have any ectopic lenses when last followed. Intraocular pressures steadily increased in all five siblings, who each received treatment in their early teens to mid-twenties. Two siblings (II-3, II-5) developed glaucoma while the others did not, and both first developed cataracts at age 50 and 49 respectively. There was no joint stiffness seen in four of the affected siblings, while one sibling (II-6) had limited movement of the wrists and shoulders, and a winged scapula. No siblings had clinically significant cardiovascular defects, although II-8 had a grade I/VI systolic ejection murmur that was not significant based on echocardiogram.

With the exception of II-2, for whom there was no data available, all siblings displayed some degree of dysmorphia in their hands. Sibling II-3 had bilateral curvature of the fifth fingers and a first proximal phalanx on the right hand that was 5.3 standard deviations below expected for her age and sex. There was also

dysmorphia seen in the hands of II-5, who had shortening of the first metacarpal on the left hand (by seven standard deviations). The metacarpophalangeal lengths in the hands of these two siblings were noticeably shorter compared to their unaffected sibling. Whereas most of the measurements in these two affected siblings were between two and three standard deviations below the mean for their age and sex, the measurements for their unaffected sibling (II-4) were mostly between zero and one standard deviation below the mean (variant (NM_139057.3: c.3068G>A: p.C1023Y) in *ADAMTS17*, being the only variant passing filtering criteria

Figure 3.10B). We did not obtain new roetnograms for affected siblings II-6 and II-8, but prior roetnograms revealed bilateral shortening of the fourth and fifth metacarpals in II-6 and of the first, fourth and fifth metacarpals in II-8. These observations reveal that WMS patients with pathogenic *ADAMTS17* variants can display variable features of hand dysmorphia.

We next investigated the genetic etiology of WMS in the family. DNA from two affected siblings (II-5 and II-3) was examined using whole exome sequencing and homozygosity mapping (Figure 3.14), which revealed the missense *ADAMTS17* variant (NM_139057.3: c.3068G>A: p.C1023Y) as the only rare homozygous variant present in a shared region of homozygosity in the two siblings. The variant segregated as expected for an autosomal recessive disorder, with their other affected sibling (II-8) being a homozygous carrier and their unaffected sibling (II-4) having two wild type alleles (Figure 3.15). Their mother (I-2) was a heterozygous carrier of the p.C1023Y variant, while their deceased unaffected father is an obligate

carrier who is presumed to be heterozygous. This variant is rare in the general population databases, with 0.0081% MAF in the NHLBI exome variant server, and it is predicted to be highly deleterious with a CADD score of 32. The variant was absent in a further 300 control alleles from the NL population, demonstrating that it is not common to this population. Interestingly, while others have reported frameshift, nonsense and splice variants in *ADAMTS17*, the variant identified in our study is the first pathogenic missense variant (Figure 3.16).

The p.C1023Y point mutation localizes to the fifth thrombospondin repeat-1 domain of ADAMTS17. To functionally characterize the impact of this variant, we introduced p.C1023Y into full-length ADAMTS17 (TS17) and transfected plasmids into HEK293T cells (Figure 3.17). Additionally, given that ADAMTS17 is auto proteolytic once secreted into cell media, we also introduced the p.C1023Y variant into a proteolytically inactive *ADAMTS17* construct (ADAMTS17^{E390A}) to prevent degradation of ADAMTS17. Compared to wildtype, the p.C1023Y mutants are significantly reduced in both the full length ADAMTS17 (p= 0.0012) and TS17^{E390A} constructs (p= 0.00014) in the cell media. There were no differences seen in the cell lysate between wild type and ADAMTS17, indicating that the protein is expressed but not secreted. It is likely that the change from cysteine to tyrosine causes instability in conformation of the protein.

Taken together, we have identified and characterized a novel rare homozygous missense mutation (p.C1023Y) in *ADAMTS17*, which segregates in a large family from an isolated region of NL. The clinical information demonstrated

here shows the variable and intrafamilial variability in the phenotypic manifestations of this rare disorder, with dysmorphic hand features prominent in family members, who for the most part lack joint stiffness. The finding of this pathogenic variant may provide others with a genetic diagnosis for WMS, and our investigation provides unique insights into the molecular pathogenesis of ADAMTS17, identifying a key residue within the fifth thrombospondin-1 repeat domain.

3.9 References

Ansari, M. W., & Nadeem, A. (2016). The Eyeball: Some Basic Concepts. In *Atlas of Ocular Anatomy* (pp. 11–27). Springer, Cham. https://doi.org/10.1007/978-3-319-42781-2_2

Dagoneau, N., Benoist-Lasselin, C., Huber, C., Faivre, L., Mégarbané, A., Alswaid, A., … Cormier-Daire, V. (2004). ADAMTS10 mutations in autosomal recessive Weill-Marchesani syndrome. *American Journal of Human Genetics*, *75*(5), 801–806. https://doi.org/10.1086/425231

Dietz, H. (1993). Marfan Syndrome. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, K. Stephens, & A. Amemiya (Eds.), *GeneReviews*®. Seattle (WA): University of Washington, Seattle. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1335/

Dietz, H. C., Cutting, G. R., Pyeritz, R. E., Maslen, C. L., Sakai, L. Y., Corson, G. M., … Curristin, S. M. (1991). Marfan syndrome caused by a recurrent de novo

missense mutation in the fibrillin gene. *Nature*, *352*(6333), 337–339. https://doi.org/10.1038/352337a0

Dyment, D. A., Smith, A. C., Alcantara, D., Schwartzentruber, J. A., Basel-Vanagaite, L., Curry, C. J., … Innes, A. M. (2013). Mutations in PIK3R1 cause SHORT syndrome. *American Journal of Human Genetics*, *93*(1), 158–166. https://doi.org/10.1016/j.ajhg.2013.06.005

Evereklioglu, C., Hepsen, I. F., & Er, H. (1999). Weill-Marchesani syndrome in three generations. *Eye (London, England)*, *13 (Pt 6)*, 773–777. https://doi.org/10.1038/eye.1999.226

Fahiminiya, S., Almuriekhi, M., Nawaz, Z., Staffa, A., Lepage, P., Ali, R., … Ben-Omran, T. (2014). Whole exome sequencing unravels disease-causing genes in consanguineous families in Qatar. *Clinical Genetics*, *86*(2), 134–141. https://doi.org/10.1111/cge.12280

Faivre, L., Gorlin, R. J., Wirtz, M. K., Godfrey, M., Dagoneau, N., Samples, J. R., … Cormier-Daire, V. (2003b). In frame fibrillin-1 gene deletion in autosomal dominant Weill-Marchesani syndrome. *Journal of Medical Genetics*, *40*(1), 34–36.

Faivre, Laurence, Dollfus, H., Lyonnet, S., Alembik, Y., Mégarbané, A., Samples, J., … Cormier-Daire, V. (2003). Clinical homogeneity and genetic heterogeneity in Weill-Marchesani syndrome. *American Journal of Medical Genetics. Part A*, *123A*(2), 204–207. https://doi.org/10.1002/ajmg.a.20289

Faivre, Laurence, Mégarbané, A., Alswaid, A., Zylberberg, L., Aldohayan, N., Campos-

Xavier, B., … Cormier-Daire, V. (2002). Homozygosity mapping of a Weill-Marchesani syndrome locus to chromosome 19p13.3-p13.2. *Human Genetics*, *110*(4), 366–370. https://doi.org/10.1007/s00439-002-0689-3

Fuchs, J., & Rosenberg, T. (1998). Congenital ectopia lentis. A Danish national survey. *Acta Ophthalmologica Scandinavica*, *76*(1), 20–26.

Green, R. C., Green, J. S., Buehler, S. K., Robb, J. D., Daftary, D., Gallinger, S., … Younghusband, H. B. (2007). Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Familial Cancer*, *6*(1), 53–62. https://doi.org/10.1007/s10689-006-9104-x

Guo, H., Wu, X., Cai, K., & Qiao, Z. (2015). Weill-Marchesani syndrome with advanced glaucoma and corneal endothelial dysfunction: a case report and literature review. *BMC Ophthalmology*, *15*. https://doi.org/10.1186/1471-2415-15-3

Haji-Seyed-Javadi, R., Jelodari-Mamaghani, S., Paylakhi, S. H., Yazdani, S., Nilforushan, N., Fan, J.-B., … Elahi, E. (2012). LTBP2 mutations cause Weill-Marchesani and Weill-Marchesani-like syndrome and affect disruptions in the extracellular matrix. *Human Mutation*, *33*(8), 1182–1187. https://doi.org/10.1002/humu.22105

Halpert, M., & BenEzra, D. (1996). Surgery of the hereditary subluxated lens in children. *Ophthalmology*, *103*(4), 681–686.

Harasymowycz, P., & Wilson, R. (2004). Surgical Treatment of Advanced Chronic Angle Closure Glaucoma in Weill–Marchesani Syndrome. *Journal of Pediatric*

Ophthalmology and Strabismus, *41*(5), 295–299. https://doi.org/10.3928/01913913-20040901-08

Hubmacher, D., & Apte, S. S. (2011). Genetic and functional linkage between ADAMTS superfamily proteins and fibrillin-1: a novel mechanism influencing microfibril assembly and function. *Cellular and Molecular Life Sciences: CMLS*, *68*(19), 3137–3148. https://doi.org/10.1007/s00018-011-0780-9

Hubmacher, D., Schneider, M., Berardinelli, S. J., Takeuchi, H., Willard, B., Reinhardt, D. P., … Apte, S. S. (2017). Unusual life cycle and impact on microfibril assembly of ADAMTS17, a secreted metalloprotease mutated in genetic eye disease. *Scientific Reports*, *7*, 41871. https://doi.org/10.1038/srep41871

Jensen, S. A., Robertson, I. B., & Handford, P. A. (2012). Dissecting the Fibrillin Microfibril: Structural Insights into Organization and Function. *Structure*, *20*(2), 215–225. https://doi.org/10.1016/j.str.2011.12.008

Johnson, G. J., & Bosanquet, R. C. (1983). Spherophakia in a Newfoundland family: 8 years' experience. *Canadian Journal of Ophthalmology. Journal Canadien D'ophtalmologie*, *18*(4), 159–164.

Khan, A. O., Aldahmesh, M. A., Al-Ghadeer, H., Mohamed, J. Y., & Alkuraya, F. S. (2012). Familial spherophakia with short stature caused by a novel homozygous ADAMTS17 mutation. *Ophthalmic Genetics*, *33*(4), 235–239. https://doi.org/10.3109/13816810.2012.666708

Kloepfer, H. W., & Rosenthal, J. W. (1955). Possible genetic carriers in the spherophakia-brachymorphia syndrome. *American Journal of Human*

Genetics, *7*(4), 398–425.

Le Goff, C., & Cormier-Daire, V. (2011). The ADAMTS(L) family and human genetic disorders. *Human Molecular Genetics*, *20*(R2), R163-167. https://doi.org/10.1093/hmg/ddr361

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Marchesani, O. (1939). Brachydaktylie und angeborene Kugellinse als Systemerkrankung. *Klin. Monatsbl. Augenheilkd*, *103*, 392–406.

McDonald-McGinn, D. M., Fahiminiya, S., Revil, T., Nowakowska, B. A., Suhl, J., Bailey, A., … Jerome-Majewska, L. A. (2013). Hemizygous mutations in SNAP29 unmask autosomal recessive conditions and contribute to atypical findings in patients with 22q11.2DS. *Journal of Medical Genetics*, *50*(2), 80–90. https://doi.org/10.1136/jmedgenet-2012-101320

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Morales, J., Al-Sharif, L., Khalil, D. S., Shinwari, J. M. A., Bavi, P., Al-Mahrouqi, R. A., … Al Tassan, N. (2009). Homozygous mutations in ADAMTS10 and ADAMTS17 cause lenticular myopia, ectopia lentis, glaucoma, spherophakia, and short stature. *American Journal of Human Genetics*, *85*(5), 558–568. https://doi.org/10.1016/j.ajhg.2009.09.011

Poznanski, A. K. (1984). *The Hand in Radiologic Diagnosis: v. 2* (2nd edition). Philadelphia: W B Saunders Co Ltd.

Radner, F. P. W., Marrakchi, S., Kirchmeier, P., Kim, G.-J., Ribierre, F., Kamoun, B., … Fischer, J. (2013). Mutations in CERS3 cause autosomal recessive congenital ichthyosis in humans. *PLoS Genetics*, *9*(6), e1003536. https://doi.org/10.1371/journal.pgen.1003536

Ramirez, F., Sakai, L. Y., Dietz, H. C., & Rifkin, D. B. (2004). Fibrillin microfibrils: multipurpose extracellular networks in organismal physiology. *Physiological Genomics*, *19*(2), 151–154. https://doi.org/10.1152/physiolgenomics.00092.2004

Razeghinejad, M. R., Hosseini, H., & Namazi, N. (2009). Biometric and corneal topographic characteristics in patients with Weill-Marchesani syndrome. *Journal of Cataract and Refractive Surgery*, *35*(6), 1026–1032. https://doi.org/10.1016/j.jcrs.2009.01.029

Senthil, S., Rao, H. L., Hoang, N. T. Q., Jonnadula, G. B., Addepalli, U. K., Mandal, A. K., & Garudadari, C. S. (2014). Glaucoma in microspherophakia: presenting features and treatment outcomes. *Journal of Glaucoma*, *23*(4), 262–267.

https://doi.org/10.1097/IJG.0b013e3182707437

Shah, M. H., Bhat, V., Shetty, J. S., & Kumar, A. (2014). Whole exome sequencing identifies a novel splice-site mutation in ADAMTS17 in an Indian family with Weill-Marchesani syndrome. *Molecular Vision*, *20*, 790–796.

Smith, A. C., Mears, A. J., Bunker, R., Ahmed, A., MacKenzie, M., Schwartzentruber, J. A., … Graham, G. E. (2014). Mutations in the enzyme glutathione peroxidase 4 cause Sedaghatian-type spondylometaphyseal dysplasia. *Journal of Medical Genetics*, *51*(7), 470–474. https://doi.org/10.1136/jmedgenet-2013-102218

Sridhar, J., & Chang, J. S. (2017). Marfan's Syndrome with Ectopia Lentis. *New England Journal of Medicine*, *377*(11), 1076–1076. https://doi.org/10.1056/NEJMicm1406002

Taylor, J. N. (1996). Weill-Marchesani syndrome complicated by secondary glaucoma. Case management with surgical lens extraction. *Australian and New Zealand Journal of Ophthalmology*, *24*(3), 275–278.

Tsilou, E., & MacDonald, I. M. (1993). Weill-Marchesani Syndrome. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, K. Stephens, & A. Amemiya (Eds.), *GeneReviews*®. Seattle (WA): University of Washington, Seattle. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1114/

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115. https://doi.org/10.1093/nar/gks596

Vanakker, O., Callewaert, B., Malfait, F., & Coucke, P. (2015). The Genetics of Soft

Connective Tissue Disorders. *Annual Review of Genomics and Human Genetics*, *16*(1), 229–255. https://doi.org/10.1146/annurev-genom-090314-050039

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Weill, G. (1932). Ectopie des cristallins et malformations générales. *Ann Oculist,*169:21-24.

Wirtz, M. K., Samples, J. R., Kramer, P. L., Rust, K., Yount, J., Acott, T. S., … Godfrey, M. (1996). Weill-Marchesani syndrome--possible linkage of the autosomal dominant form to 15q21.1. *American Journal of Medical Genetics*, *65*(1), 68–75. https://doi.org/10.1002/(SICI)1096-8628(19961002)65:1<68::AID-AJMG11>3.0.CO;2-P

Yang, J., Fan, Q., Chen, J., Wang, A., Cai, L., Sheng, H., … Lu, Y. (2016). The efficacy of lens removal plus IOL implantation for the treatment of spherophakia with secondary glaucoma. *British Journal of Ophthalmology*, *100*(8), 1087–1092. https://doi.org/10.1136/bjophthalmol-2015-307298

Zeyer, K. A., & Reinhardt, D. P. (2015). Fibrillin-containing microfibrils are key signal relay stations for cell function. *Journal of Cell Communication and Signaling*, *9*(4), 309–325. https://doi.org/10.1007/s12079-015-0307-5

Chapter 4. Candidate Gene Screening Reveals Evidence *GALNT12* is a Moderate Penetrance Gene for Colorectal Cancer within the Newfoundland & Labrador Population.

4.1 Background: Hereditary Colorectal Cancer

Neoplasia of the large intestine or rectum (i.e. colorectal cancer), also known as CRC, is a common malignancy affecting 1 in 13 Canadians (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2015). CRC is the second leading cause of cancer death among Canadian men and the third leading cause among Canadian women. NL has the highest age and sex-standardized incidence of CRC reported in Canada, with 435.8 new cases per 100,000. NL has the highest familial incidence of CRC reported anywhere in the world (Green *et al.,* 2007).

Three molecular pathways define the pathogenesis of CRC (Colussi *et al.,* 2013). These are the chromosomal instability (CIN), DNA mismatch repair (MMR) and CpG island methylator phenotype (CIMP) pathways. CRC evolves through the adenoma-to-carcinoma sequence, whereby premalignant lesions (i.e. polyps) acquire somatic mutations and eventually progress to invasive malignancy over a period of 5-10 years (Kuipers *et al.,* 2015) (Figure 4.1). This growth phase is often subclinical, with as few as only 39% of CRCs detected early (Hudson *et al.,* 2012). As the second most commonly diagnosed cancer in Canada, the five-year survival rate is 64% (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2015).

Despite this, CRC is highly preventable. Early detection and polypectomy is effective in reducing CRC incidence. For example, colonoscopic surveillance has been

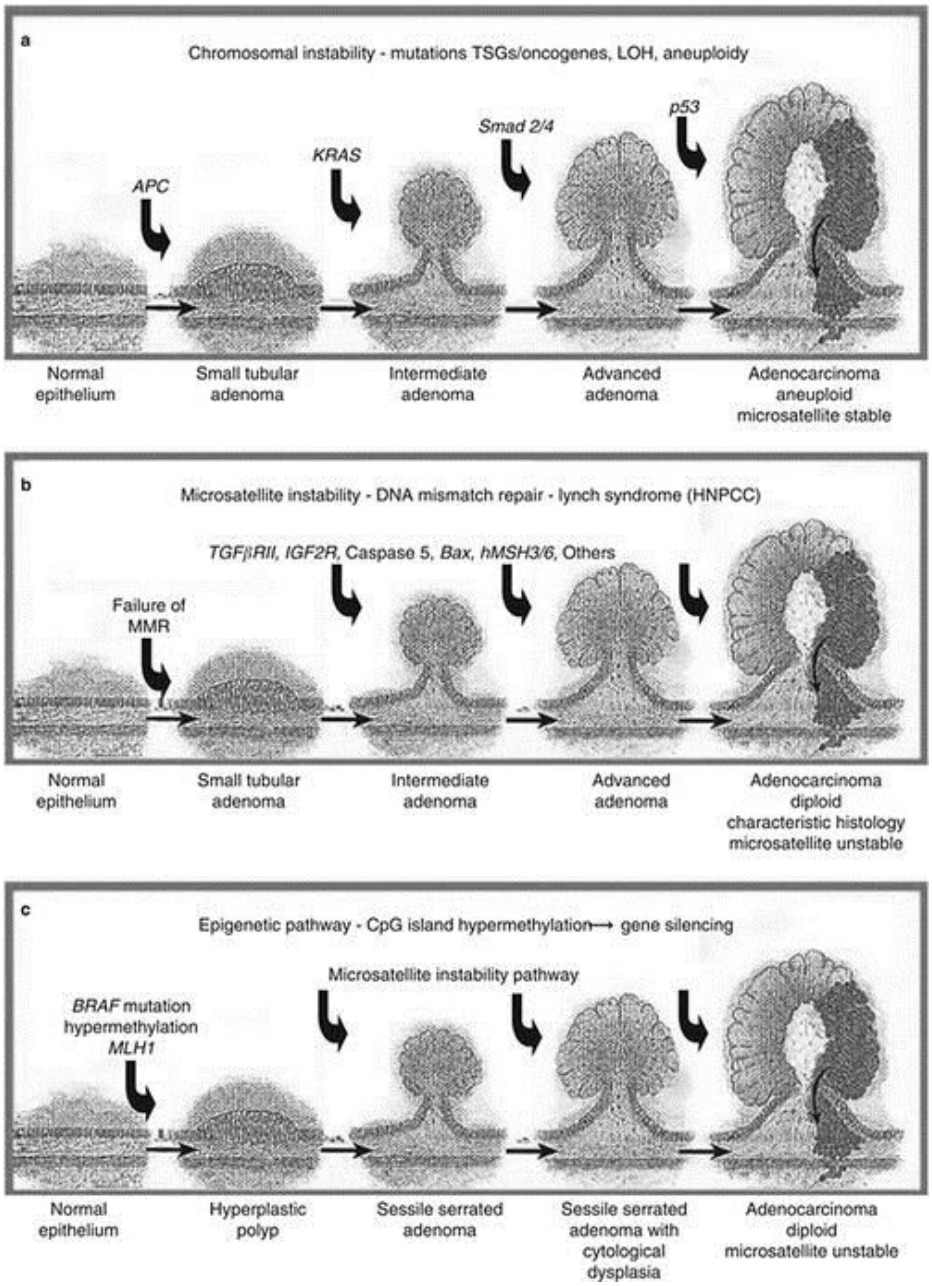


Figure 4.1 The adenoma to carcinoma sequence and common genetic abberations in each respective CRC molecular pathway. Reprinted from Ahnen (2011) with copyright permission.

shown to decrease CRC-related mortality by up to 33% (Pan *et al.,* 2016; Stanesby and Jenkins, 2017). Therefore, CRC screening programs are effective to prevent this disease. The Canadian Task Force on Preventative Health Care therefore recommends individuals at population-level risk begin CRC screening at age 50 (Canadian Task Force on Preventive Health Care, 2016). Preventative initiatives such as screening colonoscopies can therefore offer patients options including prophylactic bowel resection. Individuals with strong genetic predisposition for CRC may develop CRC earlier than age 50. For this reason, discovery of genetic markers is useful to prioritize these patients for targeted CRC screening (Lech *et al.,* 2016). Identification of novel genetic markers is therefore a high priority within the scientific community (Gonzalez-Pons and Cruz-Correa, 2015). The heritability of CRC can be up to 35% based on twin studies (Jiao *et al.,* 2014). CRC is a genetically heterogeneous disease, which demonstrates both multifactorial and monogenic etiology. Up to 30% of CRC cases have a familial component (i.e. more than one first degree relative affected), caused by a number of genetic and environmental factors (Jasperson *et al.,* 2010) (Figure 4.2). Hahn and others describe the model of genetic heterogeneity in CRC, which examines relative risk against allele frequency (Hahn *et al.,* 2016; Manolio *et al.,* 2009) (Figure 4.3). Under this model, there are rare and highly penetrant risk factors, as well as common low penetrance risk factors. Experimental methodologies such as whole exome sequencing have been beneficial in studying and identification of several highly penetrant monogenic disorders, while genome wide association studies have been successful in identifying

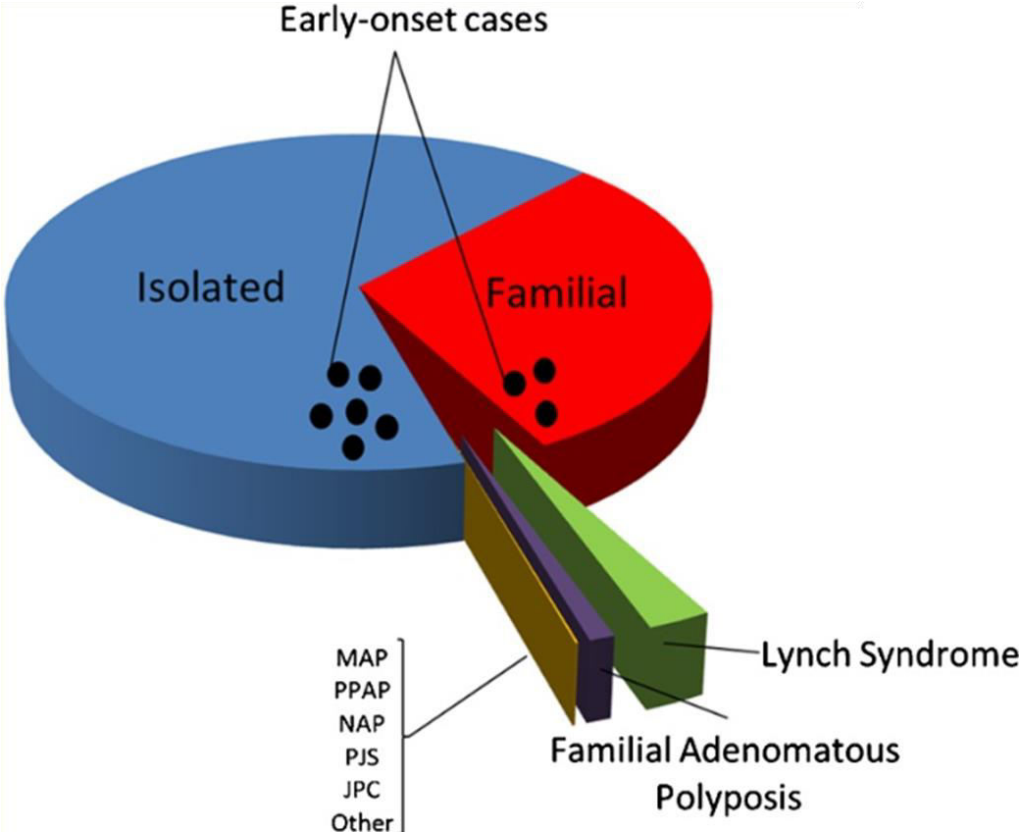


Figure 4.2 The overall proportion of hereditary, sporadic and familial CRC cases. Reprinted from Hahn *et al.,* (2016) with copyright permission.

Abbreviations: *MUTYH*-associated polyposis (MAP), polymerase proofreading-associated polyposis (PPAP), *NTHL1*-associated polyposis (NAP), Peutz-Jeghers Syndrome (PJS), juvenile polyposis coli (JPC).

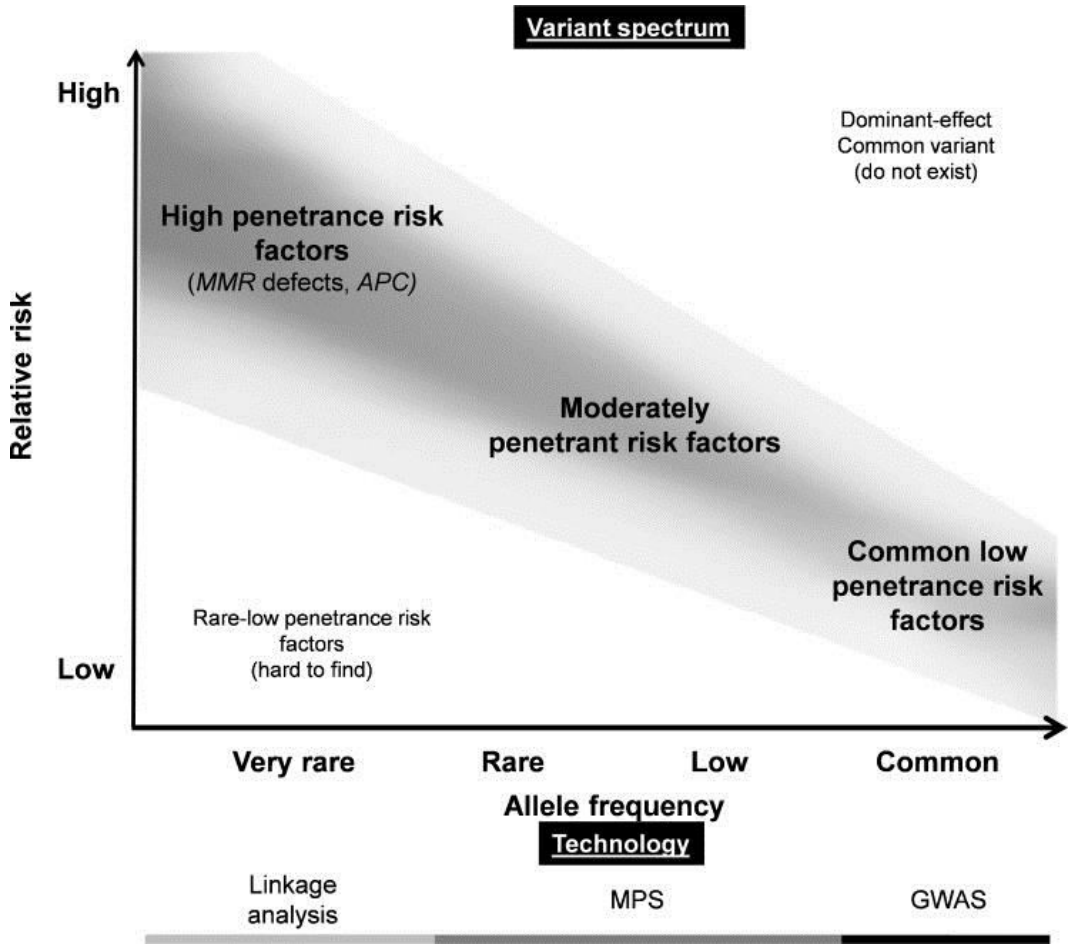


Figure 4.3 A model of the genetic heterogeneity of CRC, with variant penetrance as a function of allele frequency. Reprinted from Hahn et al. (2016) with copyright permission.

common low penetrance risk factors. Meanwhile, the rare-to-low frequency moderate penetrance risk factors have been largely unexplored in CRC.

In terms of the highly penetrant genes, about 5-10% of CRC cases are ascribed to the monogenic hereditary CRC disorders (Brandão and Lage, 2015). FAP is a hereditary CRC syndrome accounting for less than 1% of all cases (Half *et al.,* 2009). CRC is inherited in an autosomal dominant manner in FAP. Mutations in Adenomatous Polyposis Coli (*APC)* are responsible for FAP. Moreover, there is significant allelic heterogeneity, with more than 1000 pathogenic *APC* variants described (Rivera *et al.,* 2011). In FAP, CRC arises through the CIN pathway, due to disrupted signalling of the WNT pathway. *APC* is a tumor suppressor gene that regulates WNT signalling. The APC protein regulates degradation of β-catenin -- a transcription factor that induces proliferation of intestinal cell crypts (Benchabane and Ahmed, 2009). Clinically, FAP manifests with hundreds to thousands of polyps throughout the gastrointestinal tract by the second decade of life. The extracolonic manifestations of FAP include papillary thyroid carcinoma, hepatoblastoma, congenital hypertrophy of the retinal pigment epithelium, osteomas and desmoids, the occurance and frequency of these being correlated with mutations position along the *APC* gene (Groen *et al.,* 2008). FAP is almost fully penetrant, with CRC by age 40 if no screening and management is in place (Malloy & Smith, 2015). Finally, pathogenic variants in *APC* can also cause a clinical variant known as attenuated FAP, which is characterized by fewer adenomatous polyps in the gastrointestinal tract (i.e. 10-100).

Juvenile Polyposis Syndrome (JPS) is another autosomal dominant hereditary CRC syndrome caused by mutations affecting WNT signalling. JPS accounts for a small fraction of CRC cases (<1%) and arises by pathogenic variants in *SMAD4* and *BMPR1a* (Larsen Haidle & Howe, 1993). The term juvenile refers to the histological appearance of the polyps. The diagnostic criteria for JPS are satisfied if a proband is found with any of the following: 5 or more juvenile polyps in colon or rectum or any number of juvenile polyps with positive family history; multiple juvenile polyps in both upper and lower gastrointestinal tract; or a known pathogenic variant in *SMAD4* or *BMPR1a*. As the above indicates, JPS can manifest with juvenile polyps as part of JPS, however it may also manifest as JPS with hereditary hemorrhagic telangactasia (HHT) (i.e. extracolonic features). These extracolonic manifestations in JPS with HHT include epistaxis, arteriovenous malformations, digital clubbing and telangiectases (Williams *et al.,* 2012). There is a significantly increased risk for CRC in JPS patients, having a cumulative lifetime risk of 38.7% and mean age of CRC diagnosis at 43.9 years (Brosens *et al.,* 2007). While JPS with HHT is limited to patients with *SMAD4* variants, a separate syndrome from JPS is also seen in some patients with *BMPR1a* variants – called hereditary mixed polyposis syndrome (O'Riordan *et al.,* 2010). The latter is characterized by different polyp types with mixed adenomatous, hyperplastic, and atypical juvenile histologies -- and can also be caused by the *GREM1* gene (Lieberman *et al.,* 2017).

Peutz-Jeghers syndrome (PJS) is another rare autosomal dominant disorder characterized by pathogenic variants in *STK11* and characterized by polyps of

hamartomatous histology. *STK11* acts as a tumor suppressor gene restricting cell division. Hamartomatous polyps develop in 90% of PJS patients, and are found throughout the gastrointestinal tract – predominantly in the small and large intestine, but are also found in the stomach and urinary tract (Anaya *et al.,* 2008). PJS patients are at a high lifetime risk (93%) for developing CRC and extracolonic cancers, including pancreatic, gastric, esophageal, uterine, ovarian, lung, testicular and breast cancers. The presence of mucocutaneous pigmentations seen around the buccal mucosa and anus can help establish the diagnosis of PJS.

Another autosomal dominant CRC disorder is Serrated Polyposis Syndrome, again accounting for a small fraction of overall CRC cases. Polyps in this syndrome are of the hyperplastic histological type. The genetic basis for this syndrome is not yet known (Hassan *et al.,* 2016). Several criteria exist according to the WHO and can include 1.) at least 5 serrated polyps (two of which are >1 cm) located proximal to the sigmoid colon or 2.) any type of serrated polyp proximal to the sigmoid colon if there is family history of this disorder in a first degree relative, or 3.) twenty or more serrated polyps throughout the colon (Snover *et al.,* 2010). The risk for CRC in Serrated Polyposis Syndrome is lower than for other hereditary CRCs (IJspeert *et al.,* 2017).

MUTYH-associated polyposis (MAP) is the only autosomal recessive CRC syndrome and is a highly penetrant, but rare cause of CRC (Poulsen & Bisgaard, 2008). Biallelic mutations in this base-excision repair gene predispose to CRC and a number of extracolonic manifestations including: duodenal adenomas, osteoma,

desmoid cysts, esophageal cancer, thyroid cancer, breast cancer, uterine cancer, leukemia, dental cysts, tooth agenesis, basal cell carcinoma and congenital hypertrophy of the retinal pigment epithelium and leukemia and uterine cancer, among others. Patients may develop tens to hundreds or even thousands of adenomas in the gastrointestinal tract, and therefore the clinic manifestation can mirror attenuated FAP or FAP. MAP is highly penetrant, with 80% of biallelic carriers developing CRC by age 70, and mean diagnosis of CRC being age 48 (Kantor *et al.,* 2017).

HNPCC is a large umbrella term and accounts for the largest percentage of hereditary CRC cases (10%) (Kohlmann and Gruber, 1993). In HNPCC, patients develop markedly fewer adenomas than the polyposis colon cancers, but have significant risk for CRC. Most individuals with HNPCC fulfill the Amsterdam Criteria 1 (AC-1): which requires that there are three or more CRC cases in a family, with at least one being a first degree relative of the others. The age of diagnosis in one patient must be under age 50, and CRC cases must be histologically confirmed where possible. Finally, CRC must be found in two or more successive generations. The AC-1 provides a starting point to identify families suspicious for a genetic predisposition to CRC. From there, two possibilities exist: either patients have inherited defects in MMR genes (i.e. Lynch Syndrome or LS) or they meet AC-1 but have no detectable MMR defects (i.e. FCCTX).

LS is an autosomal dominant hereditary CRC disorder accounting for a relatively large percentage of CRC cases overall (~10%). Four genes are implicated

in LS and these are all involved in the MMR pathway. These genes are *MLH1*, *MSH2*, *MSH6* and *PMS2*. In addition to sequencing these genes, suggestive deficits in MMR genes can be identified by testing tumors for microsatellite instability (MSI) and for intact MMR proteins tested by immunohistochemistry. High MSI (MSH-H) indicates defective DNA repair processes; thus the tumors of LS patients have MSI and MMR defects, which will usually prompt sequencing of the MMR genes in patient DNA. LS patients have a high risk for CRC (52-82% of patients, with mean diagnosis between 44-61 years) in addition to a high risk of extracolonic cancers, most notably endometrial and ovarian cancers. Other extracolonic malignancies include gastric, urogenital, brain, biliary tract, pancreatic, skin and small bowel cancers (Bansidhar, 2012).

Families that meet AC-1, but have microsatellite stable (MSS) tumors, and intact MMR proteins by immunohistochemistry, with no germline mutations in either *APC* or MMR genes, are grouped into the heterogeneous classification called FCCTX. The mean age of CRC in FCCTX patients is later then for LS (i.e. 57 years compared to 43-45 for LS) (Dominguez-Valentin *et al.,* 2015). It is likely that numerous genes are responsible for this heterogeneous cluster of families with a strong predisposition of unknown etiology, and therefore FCCTX families are an ideal group for investigation of novel genetic causes for CRC. Several putative FCCTX genes have been identified over the years, including *BMPR1a* (Nieminen *et al.,* 2011), *SEMA4A* (Schulz *et al.,* 2014), *RPS20* (Nieminen *et al.,* 2014) and *POLE*/*POLD1* (Palles *et al.,* 2013). Among these candidates (and numerous others identified in early onset

non-FCCTX families), only mutations in the POLE gene have been widely replicated and classified into a new category (polymerase proofreading associated polyposis) (Church, 2014). Thus, while the investigation of FCCTX families has yielded a number of putative candidate genes -- many have not been globally replicated and therefore represent CRC syndromes for a very limited number of families or are false assertions. An FCCTX study of the NL population (*N*=30 unrelated FCCTX probands) screened *BMPR1a* for pathogenic variants, and did not identify any mutations of this gene in FCCTX (Evans *et al.,* 2017). Furthermore, whole exome sequencing of 30 NL FCCTX probands has not identified pathogenic variants in any of the aforementioned candidate FCCTX genes (Woods & Evans, unpublished data).

The paucity of novel CRC genes that describe CRC in the way that *APC* or the MMR genes do has led some to hypothesize that the highest impact genes or those that are easiest to find (i.e. the low hanging fruit) have all been discovered. Furthermore, what remain of any highly penetrant CRC predisposition genes are likely to explain only a small proportion of families within geographically restricted regions. Indeed, a key article by Chubb *et al.,* studied 8000 early onset familial CRCs (one of the largest early onset CRC cohorts assembled to date), and demonstrated that no highly penetrant CRC genes could be found (Chubb *et al.,* 2016). As well, the authors found that at most only 16% of familial CRCs could be ascribed to the known

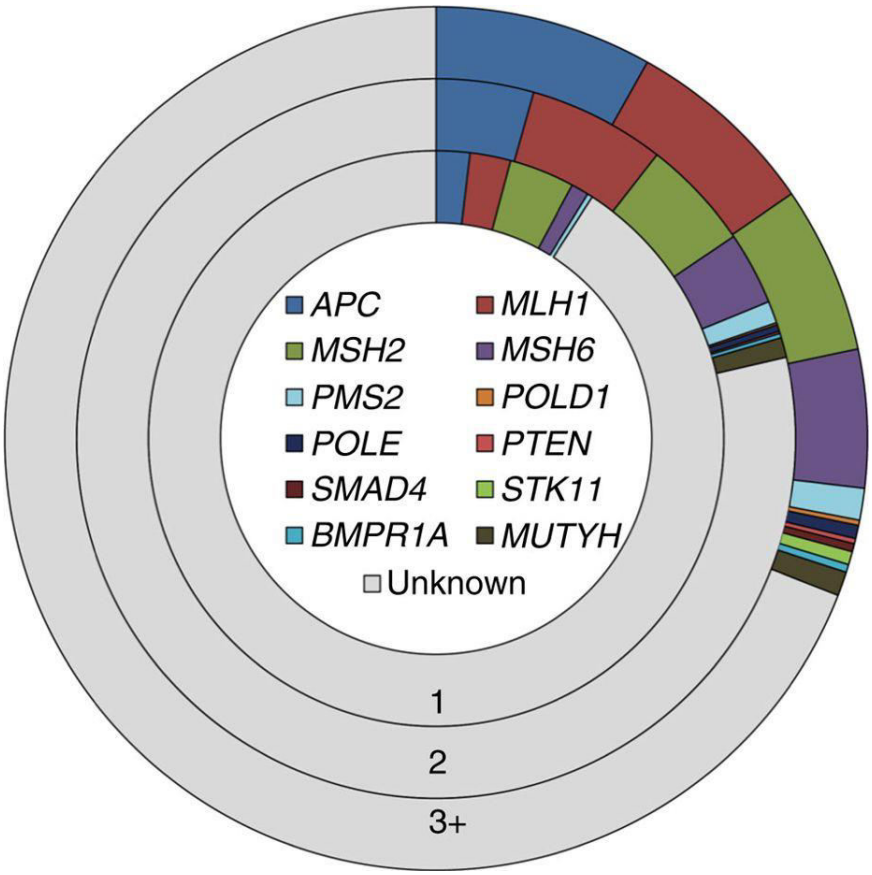


Figure 4.4 Contribution of rare variants in known CRC predisposition genes in a cohort of early onset familial CRCs. Reprinted from Chubb *et al.,* (2016) with copyright permission.

Variants were considered under different categories. Class 1 indicates disruptive variants (nonsense and frameshift). Class 2 indicates variants predicted to be damaging using *in silico* tools. Class 3+ indicates all non-synonymous variants and previously reported pathogenic splice variants according to InSight or Clinvar.

highly penetrant genes in this cohort of early onset familial CRC cases (Figure 4.4). Given that their study had over 80% statistical power to detect a novel highly penetrant CRC predisposition gene, the authors hypothesized that the likelihood of finding a global novel high penetrance predisposition gene is becoming increasingly unlikely.

As we appear to be confronted with a paucity of global high penetrance CRC predisposition genes, the research community has identified a significant knowledge gap in our understanding of rare moderate penetrance alleles, which could represent a significant fraction of the missing heritability (Valle, 2017). In recent years, limited examples of moderate penetrance variants have been recognized in the typically highly penetrant known CRC genes, reflecting the increasing reality that clinicians face with respect to CRC variant interpretation (Tung *et al.,* 2016).

To that end, *GALNT12* is a candidate CRC gene which has been investigated for several years. Several genetic linkage and association studies demonstrated the existence of a CRC-susceptibility locus spanning 9q22.32-31.1(Gray-McGuire *et al.,* 2010; Skoglund *et al.,* 2006; Wiesner *et al.,* 2003) and *GALNT12* resides in the 6.5cM critical interval (markers D9S1851-D9S277) within this region (Kemp *et al.,* 2006). Guda *et al.* first studied *GALNT12* as a candidate CRC predisposition gene (Guda *et al.,* 2009). Sequencing 30 MSS CRC cell lines, they identified two somatic variants in *GALNT12*, which showed significantly reduced enzyme activity compared to wildtype protein (Figure 4.5). Further sequencing in germline DNAs of 272 CRC

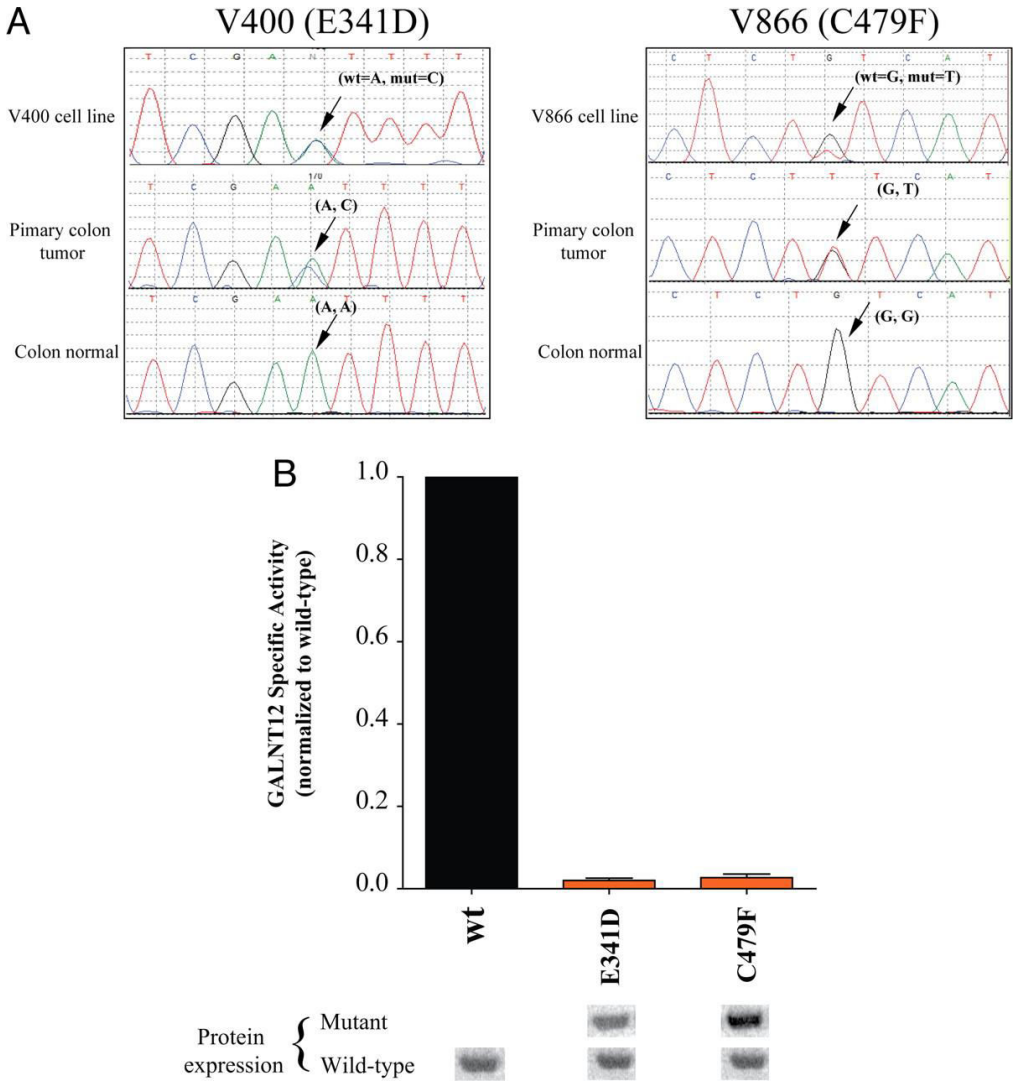


Figure 4.5 Somatic *GALNT12* variants identified in 30 MSS CRC cell lines. Reprinted from Guda *et al.,* (2009) with copyright permission.

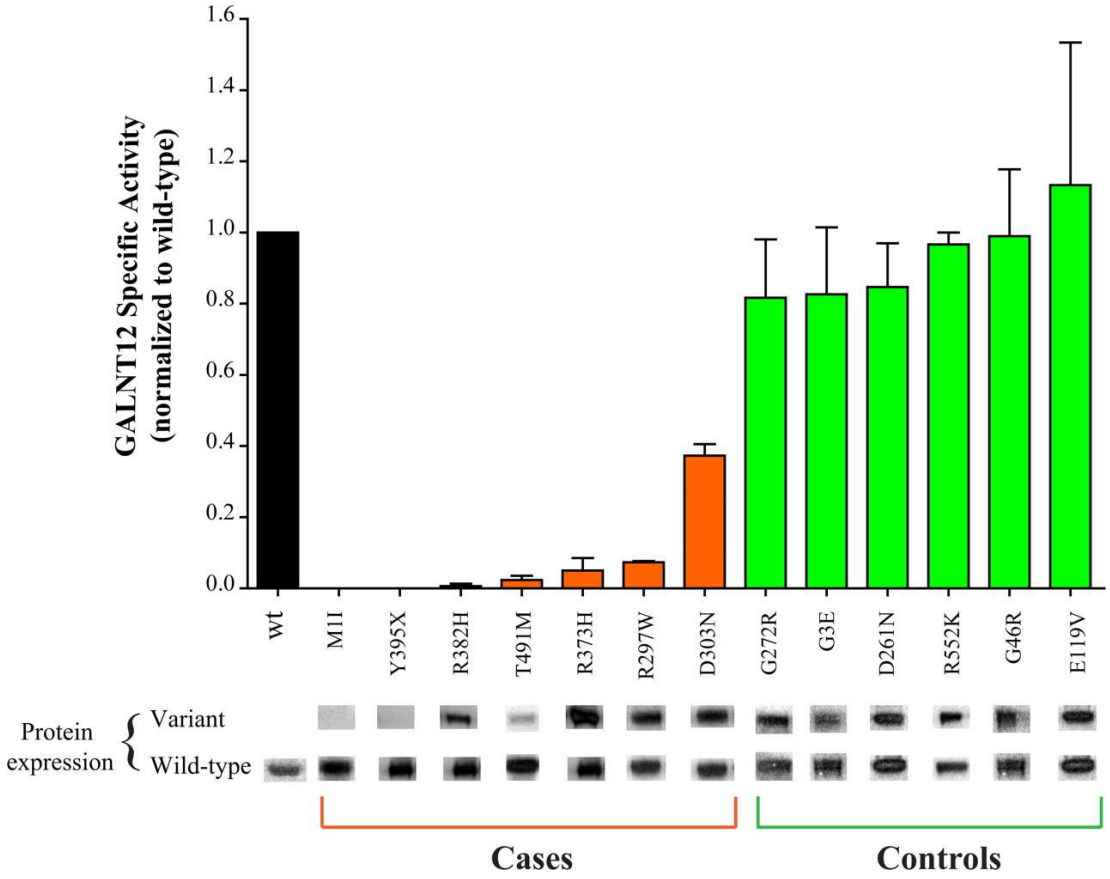


Figure 4.6 Functional assay of germline *GALNT12* variants in CRC cases and controls. Reprinted from Guda *et al.,* (2009) with copyright permission.

patients compared to 192 controls revealed six additional *GALNT12* variants exclusive to CRC cases – which all had significantly impaired GALNT12 enzyme activity compared to wildtype, and variants that were exclusive to controls (Figure 4.6).

Following this discovery of inactivating somatic and germline *GALNT12* variants in MSS CRC cells and patient germline DNAs, Clarke and colleagues then investigated *GALNT12* in a clinic-based discovery cohort of 118 probands from NL with strong CRC predisposition (Clarke *et al.,* 2012). Clarke and colleagues identified four instances of rare and deleterious *GALNT12* variants in germline DNA of probands from high risk CRC families (Table 4.1). They found three unrelated families harboring the p.D303N variant (also previously seen in the Guda et al. study), and one family harboring the p.Y396C variant. The p.D303N segregated in two CRC families, with a SISA prediction that suggested a 1.56% probability that co-segregation occurred by chance (Figure 4.7). Following the Clarke *et al*. study, Seguí and coauthors screened 103 FCCTX cases for *GALNT12* variants but did not identify any candidates for further pursuit (Seguí *et al.,* 2014). There remains continued clinical uncertainty as to the relevance of *GALNT12.* Previously, this was considered a moderate penetrance CRC predisposition gene in early versions of the National Comprehensive Cancer Network (NCCN) guidelines, however this status was removed in the 2017 edition over uncertainty and paucity in the highly penetrant FCCTX families (Benson *et al.,* 2017) . Additional analyses are required to further understand the impact of rare variation in this gene.

Table 4.1 Rare *GALNT12* Variants in clinic-based cohort of 118 high-risk NL CRC cases. Reprinted from Clarke *et al.,* (2012) with copyright permission.

Study #	Variant[a]	Proband age of diagnosis (CRC)	Location of CRC	MSI testing result[b]	IHC testing result (MLH1, MSH2, MSH6)	Bethesda criteria[c]	PolyPhen[d]	SIFT[e]
1117	c.907G>A p.Asp303Asn	72	Rectum	MSS	Intact	2, 4	PD	NT
20444	c.907G>A p.Asp303Asn	35	Rectum	ND	ND	4, 5	PD	NT
20896	c.907G>A p.Asp303Asn	72	Transverse Colon	MSS	Intact	2, 5	PD	NT
1728	c.1187A>G p.Tyr396Cys	21	Cecum	MSS	Intact	1	PD	NT

[a]Variant nomenclature is based upon the reference sequence NM_024642.4. Nucleotide numbering reflects cDNA numbering with +1 corresponding to the A of the ATG translation initiation codon in the reference sequence.
[b]MSS, microsatellite stable; ND, not determined; but DNA sequencing of *MLH1*, *MSH2*, and *MSH6* did not reveal any deleterious variants;
[c]revised Bethesda criteria 1, CRC diagnosis at less than 50 years of age; 2, synchronous or metachronous tumors (colorectal or HNPCC-related tumor); 4, one first-degree relative with HNPCC-related tumor under 50 years of age; 5, two first- or second-degree relatives with HNPCC-related tumors at any age.
[d]PD, probably damaging.
[e]NT, not tolerated. Additional nonpathogenic variants were also identified (Supp. Table S1).

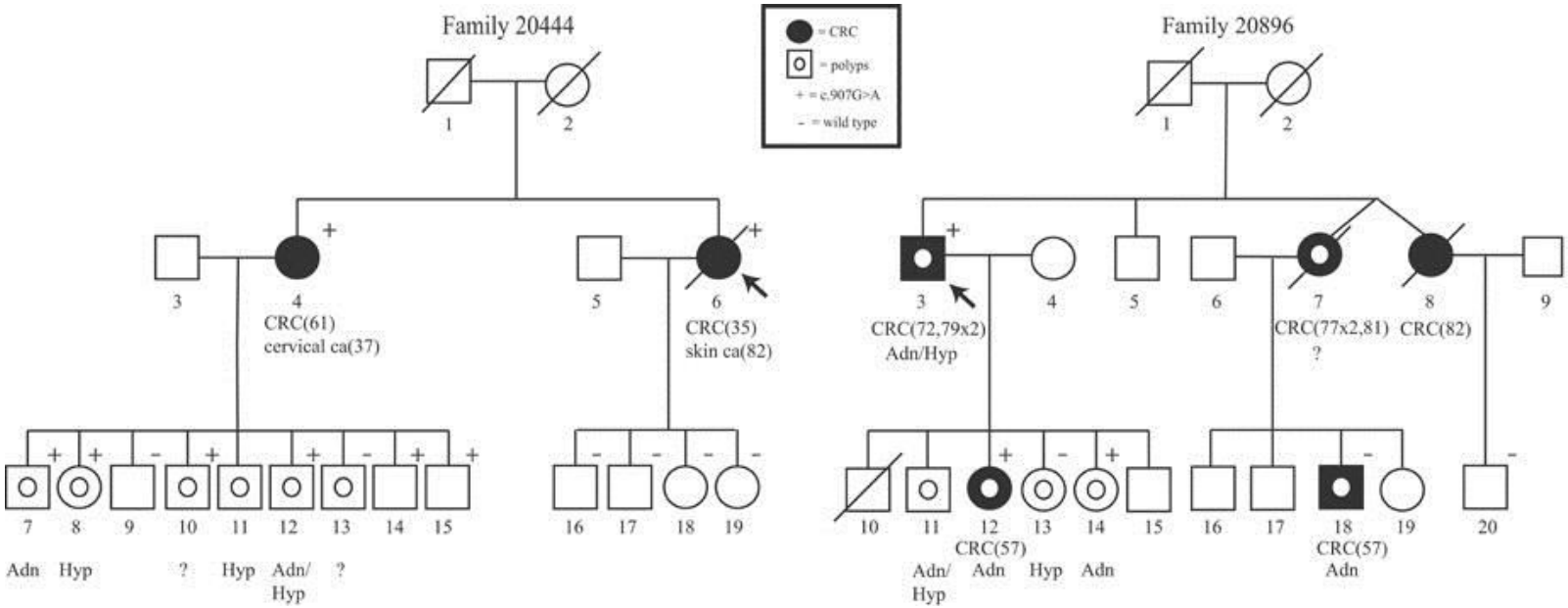


Figure 4.7 Segregation of *GALNT12* p.D303N variant in two NL families from high risk CRC cohort. Reprinted from Clarke *et al.,* (2012) with copyright permission.

4.2 Study Aims & Objectives

GALNT12 is a strong candidate for CRC susceptibility based on prior genetic and functional studies. In this study, it was hypothesized *GALNT12* is a moderate penetrance gene for CRC susceptibility, explained by rare and deleterious variants. The aim of the study was to examine a large population-based cohort (*N*=479) of cases from NL (i.e. Newfoundland Colorectal Cancer Registry – NFCCR), to identify rare *GALNT12* variants, and replicate the findings of the prior Clarke *et al.* study.

Therein, the main study objective was to screen all coding and splice regions of *GALNT12* (i.e. 10 exons and intronic boundaries) using Sanger sequencing on 479 germline DNA samples. After annotating and interpreting rare genetic variants, the second objective was to identify the frequency of these variants in population-matched NL controls (*N*=400). The final objective was to characterize the functional consequence of the rare variants identified in this study.

4.3 Author Contributions

The following is an original research article peer-reviewed and published in Human Mutation (Evans DR *et al.,* 2018). As first author of the study, I was responsible for identification and interpretation of *GALNT12* variants via sequencing of NFCCR cases and NL controls. I prepared the manuscript and interpreted key study findings. The Sanger DNA sequencing was a large-scale endeavor that required clean sequence traces of the entire length of the *GALNT12* coding region and splice junctions. Therefore, among cases and controls, the effort encompassed sequencing across all

coding bases and splice boundaries of *N*=5500 exons. Amanda T. Dohey and Erica C. Clarke assisted in primary data collection and sequenced approximately one third of exons in the NFCCR cases. Amy E. Powell and Julia J. Pennell assisted in primary data collection of 150 control samples (*N*=150 exons) in the G-C rich exon 1. JSG assisted in manuscript review and provided clinical data for *GALNT12* carriers. MOW conceived of the study and assisted in all aspects of genetic, molecular and clinical interpretation during the study. MOW supervised and revised the manuscript, and is the corresponding author.

Functional characterization of *GALNT12* candidate variants (i.e. enzyme assay) was achieved through collaboration with investigators at the Case Western Reserve University Comprehensive Cancer Center (Ohio, United States). In this regard, Srividya Venkitachalam (SV) was a key contributor and co-first author of the study. SV, Lakshmeswari Ravi (LaR) and Erina Quinn (EQ) performed cloning, mutagenesis and recombinant protein purification. SV and Leslie Reveredo (LeR) performed functional assays. SV assisted in the methods section and later critically reviewed the manuscript. Thomas A. Gerkin (TAG) and Kishore Guda (KG) supervised all enzyme assays. KG supervised the collection of molecular and genetic data, contributed to study design and assisted in preparation and review of the manuscript. All authors read and approved the final manuscript.

Co-authors had the following academic affiliations during the work described in this chapter. MOW, JSG, JJP and ATD were affiliated with the Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL A1B 3V6

Canada. SV, KG, EQ and LaR were affiliated with Division of General Medical Sciences-Oncology, Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland OH 44106, U.S.A. LeR was affiliated with Department of Chemistry, Case Western Reserve University, Cleveland OH 44106, U.S.A. TAG was affiliated with Department of Chemistry, Case Western Reserve University, Cleveland OH 44106, U.S.A as well as the Departments of Pediatrics & Biochemistry, Case Western Reserve University, School of Medicine, Cleveland OH 44106, U.S.A.

4.4 Abstract

Characterizing moderate penetrance susceptibility genes is an emerging frontier in CRC research. *GALNT12* is a strong candidate CRC-susceptibility gene given previous linkage and association studies, and inactivating somatic and germline alleles in CRC patients. Previously, we found rare segregating germline *GALNT12* variants in a clinic-based cohort (*N*=118) with predisposition for CRC. Here, we screened a new population-based cohort of incident CRC cases (*N*=479) for rare (MAF ≤1%) deleterious germline *GALNT12* variants. *GALNT12* screening revealed 8 rare variants. Two variants were previously described (p.Asp303Asn, p.Arg297Trp), and additionally, we found 6 other rare variants: five missense (p.His101Gln, p.Ile142Thr, p.Glu239Gln, p.Thr286Met, p.Val290Phe) and one putative splice-altering variant (c.732-8 G>T). Sequencing of population-matched controls (*N*=400) revealed higher burden of these variants in CRC cases compared to healthy controls (P=0.0381). We then functionally characterized the impact of these substitutions on

GALNT12 enzyme activity using *in vitro*-derived peptide substrates. Three of the newly identified *GALNT12* missense variants (p.His101Gln, p.Ile142Thr, p.Val290Phe) demonstrated a marked loss (>2-fold reduction) of enzymatic activity compared to wild-type (P≤0.05), while p.Glu239Gln exhibited a ~2-fold reduction in activity (P=0.077). These findings provide strong, independent evidence for the association of *GALNT12* defects with CRC-susceptibility; underscoring implications for glycosylation pathway defects in CRC.

4.5 Introduction

Unraveling the genetic predisposition to CRC is a subject of global significance (Rafiemanesh *et al.,* 2016; Valle, 2014). Since the elucidation of MMR genes in hereditary CRC, the search for new and highly penetrant CRC-susceptibility genes has yielded diminishing returns. Indeed, recent studies now postulate that the existence of any major high penetrance susceptibility genes, of similar caliber as the MMR genes and *APC*, is unlikely (Chubb *et al.,* 2016). This emphasizes the increasing relevance of CRC-susceptibility models that recognize the impact of rare alleles with modest effect sizes; which are difficult to detect by classic approaches such as genome-wide association and family-based linkage studies, yet could explain a sizeable fraction of the missing heritability in CRC (Hahn *et al.,* 2016; Manolio *et al.,* 2009). The genetic investigation of large and well-characterized patient cohorts is a useful way to identify moderate penetrance variants. While high penetrance variants are often the most clinically actionable, CRC screening guidelines now recognize moderate penetrance alleles in *APC*, *CHEK2* and *MUTYH* (Tung *et al.,* 2016). With the

genetic architecture of CRC becoming increasingly apparent, and the role of moderate penetrance genes becoming a new focus, this study explores *GALNT12* (MIM #610290, RefSeq: NM_024642.4) as a moderate penetrance gene for CRC susceptibility.

The initiation of mucin-type *O*-linked glycosylation is governed by the GALNT family of enzymes (Brockhausen *et al.,* 2009), and glycosylation plays a vital role in regulating key cellular functions including adhesion, migration and immune surveillance (Ohtsubo and Marth, 2006). PpGalNAc-T12 (encoded by *GALNT12*), is highly expressed in normal colon tissue and downregulated in colonic cancers (J.M. Guo, Chen, Wang, Zhang, & Narimatsu, 2004; J.M. Guo *et al.,* 2002). Aberrant glycosylation can impair the normal functioning of lipids and proteins by affecting cellular glycan structure (Pinho and Reis, 2015; Stowell *et al.,* 2015) and is a hallmark of several different cancers; including CRC (Bergstrom & Xia, 2013; Brockhausen, 2006). It is unsurprising, therefore, that genetic and epigenetic defects underlying the *O*-linked pathway are proposed to contribute to CRC pathogenesis (Brockhausen, 2006; Brockhausen *et al.,* 2009; Tran and Ten Hagen, 2013).

Genetic evidence demonstrates that *GALNT12* is a candidate CRC-susceptibility gene; as we and others have identified a CRC-susceptibility locus spanning 9q22.32-31.1 (Gray-McGuire *et al.,* 2010; Skoglund *et al.,* 2006; Wiesner *et al.,* 2003) and *GALNT12* resides in the 6.5cM critical interval (markers D9S1851-D9S277) within this region (Kemp *et al.,* 2006). Bolstered by functional evidence showing that inactivating *GALNT12* alleles are over-represented in CRC cell lines

(Guda *et al.,* 2009), we previously characterized *GALNT12* in a clinic-based discovery cohort (*N*=118) with predilection for CRC, and identified rare and segregating *GALNT12* alleles in high-risk CRC families (Clarke *et al.,* 2012).

Here, our aim was to expand the investigation of *GALNT12* by screening for rare and deleterious alleles in a larger and well-characterized population-based cohort of incident CRC cases (*N*=479) from the province of Newfoundland & Labrador (NL). This island of Atlantic Canada harbors unique founder populations of predominantly English and Irish descent, and is well-suited for genetic research; given geographic and historical isolation (Mannion, 1977), and willingness to participate in research studies. With one of the highest reported incidence rates of familial CRC in the world (Green *et al.,* 2007), and the highest incidence rate of CRC reported in Canada (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2015), the NL population offers an invaluable opportunity to study the genetic etiology of CRC (Rahman *et al.,* 2003; Zhai *et al.,* 2016). This was most significantly demonstrated by 'Family C', an NL kindred that helped elucidate the role of *MSH2* in hereditary CRC (Leach *et al.,* 1993).

To further our investigation into the putative association of *GALNT12* defects and CRC susceptibility, and leveraging the unique attributes of the NL population, we examined the germline DNAs of 479 incident CRC cases from the Newfoundland Colorectal Cancer Registry (NFCCR) (Green *et al.,* 2007). Herein, we describe novel findings of rare and deleterious *GALNT12* alleles in CRC patients, which cluster around the glycosyl-transferase domain, and functionally impair this enzyme.

4.6 Methods

4.6.1 Newfoundland Colorectal Cancer Registry and Population-matched Controls

During the prior NFCCR study, all pathologically confirmed incident CRC cases that occurred over a 5-year period (January 1st, 1999- December 31st, 2003), with diagnosis under the age of 75, were identified through provincial tumor registries (Green *et al.,* 2007). After providing written and informed consent, study participants completed family history and risk factor questionnaires and provided access to blood samples, tumor tissues, medical records and gave permission for kin contact. As well, a cohort of CRC-free population controls was ascertained through random digit dialing (Raptis *et al.,* 2007; Wang *et al.,* 2009). After providing written and informed consent, population control participants provided family history, personal history and diet questionnaires, as well as blood samples for DNA extraction.

Patients of the NFCCR cohort have previously been screened for MMR gene variants and MSI status (Woods *et al.,* 2010). MMR gene variants were assessed for pathogenicity using the InSiGHT variant database (Thompson *et al.,* 2014). We assigned familial CRC risk criteria based on relationship to the proband of each family. Risk criteria included in this study were the Amsterdam Criteria-1 (AC-1) (Vasen *et al.,* 1991), the Age and Cancer Modified Amsterdam Criteria (ACMAC) (Woods *et al.,* 2005), the Revised Bethesda Criteria (Umar *et al.,* 2004), and criteria for FCCTX (Lindor, 2009).

During the current study, we chose germline DNAs of 479 NFCCR cases and 400 population-matched CRC-free controls, which then underwent *GALNT12* screening. We were blind to both MMR variant and microsatellite status during the study. The Health Research Ethics Authority of Newfoundland & Labrador granted approval for this study (HREB # 2010.035).

4.6.2 Sanger DNA Sequencing and Variant Interpretation

CRC cases and population controls were sequenced using Sanger sequencing. Primer design was followed as per previous studies (Guda *et al.,* 2009) and PCR amplification was performed for all 10 exons and intronic boundaries of *GALNT12* (conditions available upon request). The *GALNT12* RefSeq (NM_024642.4) and the canonical transcript (ENST00000375011.3) were selected using Ensembl genome browser, with the GRCh38.p10 reference genome assembly (Zerbino *et al.,* 2018). Standard sequencing protocols for automated sequencing on an ABI 3130XL Genetic Analyzer (Applied Biosystems by Life Technologies, CA, USA) were followed. Sequencing data was then analyzed and interpreted using Sequencing Analysis 5.2 (Applied Biosystems by Life Technologies, CA, USA) and Sequencher 4.9 (Gene Codes Corporation, Michigan, USA).

Minor allele frequency (MAF) was assessed using publically available databases including: dbSNP (Sherry *et al.,* 2001), the NHLBI exome variant server (NHLBI GO Exome Sequencing Project, Seattle, WA; URL http://evs.gs.washington.edu/EVS), the ExAC browser and the GnomAD browser

(Lek *et al.,* 2016). We calculated the maximum credible allele frequency of a CRC variant, and variant penetrance, using GnomAD frequencies (Whiffin *et al.,* 2017) (Supp. Methods). A *GALNT12* variant MAF threshold of ≤1% in more than one population database was used to identify high impact variants.

For each variant, bioinformatics tools including SIFT (Sim *et al.,* 2012), Polyphen-2 (Adzhubei *et al.,* 2013), CADD (Kircher *et al.,* 2014) and REVEL (Ioannidis *et al.,* 2016) were used to predict functional consequences of rare variants. The putative splice-altering variant was assessed using three programs including Spliceman (http://fairbrother.biomed.brown.edu/spliceman/index.cgi), Human Splicing Finder (http://www.umd.be/HSF/), and SpliceView (http://bioinfo.itb.cnr.it/~webgene/wwwspliceview.html). Further details about these tools, including program inputs and settings are provided in Supplementary Methods section. We submitted rare *GALNT12* variants to the Leiden Open Variation Database (LOVD) (www.lovd.nl/GALNT12).

4.6.3 Generation of Secreted pIHV Constructs

A *GALNT12* cDNA fragment excluding the N-terminal transmembrane domain, corresponding to amino acids 38-581, was amplified by RT-PCR from RNA of a normal control as described previously (Guda *et al.,* 2009). The resulting PCR fragments were cloned into *pIHV*, a modified SV40 promoter-driven pZeoSV2 vector (Life Technologies, Carlsbad, CA, USA) that contains an insulin secretion signal to direct the secretion of the recombinant protein into the cell culture medium and an N-terminal His6 and V5 epitope tags to facilitate purification and detection of the

recombinant protein. cDNA fragments encoding mutant GALNT12 (p.His101Gln, p.Ile142Thr, p.Glu239Gln, p.Thr286Met, p.Val290Phe, p.Arg297Trp, p.Asp303Asn and p.Tyr396Cys) were generated by site directed mutagenesis (QuikChange Lightning, Agilent Technologies, Santa Clara, CA, USA).

4.6.4 Cell lines and DNA Transfection

SW480 cells were obtained from American Type Culture Collection (ATCC) in 2012, tested for authenticity using short tandem repeat genotyping at least 3 months prior to experimental assessments, and screened periodically for mycoplasma contamination using the MycoAlert mycoplasma detection kit (Lonza, Basel, Switzerland). Transfection was performed using Lipofectamine 2000 (Life Technologies, Carlsbad, CA, USA) according to standard protocol. Briefly, 10^6 SW480 cells were plated per 100-mm dish 24 hours before transfection, and incubated in 5% CO_2 at 37 °C overnight. 4 μg of plasmid DNA was used per 100-mm dish.

4.6.5 Recombinant Protein Purification

The cell monolayers were washed twice with ice-cold PBS and incubated with lysis buffer (50mMTris, pH 7.5/150mMNaCl/1mMCaCl2/1mM MnCl2/EDTA-free protease inhibitor pellets/ 0.3% CHAPS) for 15 min on ice. After scraping, the lysates were clarified by centrifugation for 15 min at maximal speed. The recombinant protein was immunoprecipitated from the lysates using anti-V5 agarose beads (Sigma-Aldrich, St Louis, MO, USA), and subsequently washed with wash buffer (50mMTris, pH 7.5/150mMNaCl/1mMCaCl2/1mM MnCl2/EDTA-free protease

inhibitor pellets).

4.6.6 Western Blot Analysis

After immunoprecipitation, 1/10 fraction of the recombinant protein was mixed with equal volume of Laemmli sample buffer (Bio-Rad, Hercules, CA, USA) at 95 °C for 5 min, and loaded onto a Bis-Tris SDS/4–12% polyacrylamide gel (Life Technologies, Carlsbad, CA, USA). After SDS/PAGE, proteins were transferred onto Immobilon-P PVDF membranes (EMD Millipore, Billerica, MA). Membranes were blocked for 1 hr with 5% nonfat milk, and incubated with appropriate dilution of mouse anti-V5 antibody conjugated to horseradish peroxidase (Life Technologies, Carlsbad, CA, USA) to detect the V5-tagged proteins. Enhanced Chemiluminescence Plus (GE Healthcare-BioSciences, Pittsburg, PA, USA) and ImageJ software (NIH, Bethesda, MA, USA) were used to detect and quantitate respective protein bands.

4.6.7 Enzymatic Assay for GALNT12 Variants

Negative control (empty vector), positive control (wild-type GALNT12), and missense mutant versions of GALNT12-bound beads were added to transferase-specific reaction mixtures in 1.5mL Eppendorf microcentrifuge tubes. Transferase-specific reaction mixtures contained 40 mM sodium cacodylate pH 6.8, 0.32 mM 2-mercaptoethanol, 0.03% Triton-X 100, 10 mM $MnCl_2$, 1 mM UDP-GalNAc containing 0.1 *u*Ci of [^{3}H] UDP-GalNAc (American Radiolabeled Chemicals Inc., St. Louis, MO, USA), protease inhibitors A and B (Sigma-Aldrich, St Louis, MO, USA) and 1 mM (0.5 mg/mL) OPT-T12 substrate GAGAYYITPRPGAGA (RS Synthesis, Louisville, KY). Of

note, OPT-T12 is an optimal peptide substrate specifically determined for ppGalNAc T12 by the Gerken Lab (Gerken *et al.,* 2011) and custom synthesized by RS Synthesis. A 10 mmol stock solution of OPT-T12 substrate was prepared by lyophilizing from water several times and adjusting to pH 7.4 with dilute NaOH/HCL. Reagent reaction mixtures were combined with 100 µL transferase-bound beads to a final reaction volume of 250 µL. The pH of the total reaction was adjusted to 6.8 and the mixture was agitated at 37°C in a TAITEC shaking Microincubator M-36 to maintain the beads in suspension. Following an overnight incubation, reaction mixtures were quenched with an equal volume of 250 mM EDTA and frozen for later processing. UDP and non-hydrolyzed UDP-GalNAc were removed by passing the sample through a column of ~3mL of Dowex 1x8 anion exchange resin (Acros Organics, Thermo Fisher Scientific, Pittsburg, PA). Subsequently, eluents were lyophilized, reconstituted with water to a final volume of 1ml and passed through G10 columns. [^{3}H]-GalNAc incorporation onto the OPT-T12 substrate was determined by scintillation counting on a Beckman LS5801 scintillation counter while absorbance at 220 nm and 280 nm was measured using a spectrophotometer. Transferase activity at each time point was calculated as a ratio of the post-G10 counts to the absorbance at 220 nm. Mutant transferase-specific activities were then normalized to the WT transferase to yield an approximate relative specific activity of each mutant. The specific activities of the wild-type, and each of the GALNT12 mutants, were assayed and expressed as the average activity derived from 3 independent biological replicates per experimental arm.

4.6.8 Statistical Analyses

Transferase activities of wild-type and mutant GALNT12 proteins were compared using a Student's t-test, with a P value ≤0.05 being considered statistically significant. A one-sided Fisher's exact test was used to compare the frequency of *GALNT12* variants in CRC cases and population-matched controls, with a P value of ≤0.05 considered statistically significant.

4.7 Results

Germline DNA sequencing of 479 NFCCR cases revealed 21 polymorphisms in *GALNT12* (Supp. Table S1). Analysis of rare (MAF ≤1%) *GALNT12* variants identified 8 candidates for further study (Table 4.1). Among these were 7 missense variants (c.303C>G, p.His101Gln; c.425T>C, p. Ile142Thr; c.715G>C, p.Glu239Gln; c.857C>T, p.Thr286Met; c.868G>T, p.Val290Phe; c.889C>T, p.Arg297Trp; c.907G>A, p.Asp303Asn) and 1 putative splice-altering variant (c.732-8G>T) (Figure 4.8) GnomAD MAFs ranged from 0.1244% (p.Asp303Asn) (344 alleles) to 0.0004061% (p.Glu239Gln) (1 allele). Maximum credible allele frequency was calculated under several penetrance models to explore potential disease allele frequency thresholds (Appendix Supp Table 6). The 8 candidate *GALNT12* variants were sequenced in a population-matched CRC-free control cohort (*N*=400). Control sequencing revealed these 8 variants were over-represented in CRC cases compared to population-matched CRC-free controls (P = 0.03814), with the p.Asp303Asn and p.Arg297Trp variants each found in a single control, respectively (Table 4.1, Appendix

Table 4.2 Rare *GALNT12* variants identified in a population-based cohort of 479 incident CRC cases from NL. Reprinted from Evans *et al.,* (2018) with copyright permission.

GALNT12 Variant	Type	dbSNP rsID	NFCCR Cohort	GnomAD (MAF/AC/Total Alleles)	SIFT	Polyphen-2 HumVar	REVEL	CADD	NL Controls (CRC-free)
c.303C>G (p.His101Gln)	SNV	rs201926457	1/479	0.0001516/27/178098	0.01 (D)	0.979 (PRD)	0.3301	28.1	0/400
c.425T>C (p.Ile142Thr)	SNV	rs757214097	1/479	NA	0 (D)	0.998 (PRD)	0.76	27.5	0/400
c.715G>C (p.Glu239Gln)	SNV	rs777144221	1/479	0.00008728/17/194766	0.05 (D)	0.99 (PRD)	0.5701	29.6	0/400
g.98831764G>T (c.732-8 G>T)	intronic/splice variant	rs763682300	1/479	0.000004061/1/246220	NA	NA	NA	12.18	0/400
c.857C>T (p.Thr286Met)	SNV	rs548915885	1/479	0.00002844/7/246124	0.01 (D)	0.999 (PRD)	0.36	34	0/400
c.868G>T (p.Val290Phe)	SNV	rs371949942	1/479	0.00009383/26/277090	0 (D)	0.94 (PRD)	0.36	25	0/400
c.889C>T (p.Arg297Trp)	SNV	rs149726976	1/479	0.0002744/76/276924	0 (D)	0.908 (POD)	0.3000	27.3	1/400
c.907G>A (p.Asp303Asn)	SNV	rs145236923	3/479	0.001244/344/276526	0.47 (T)	0.999 (PRD)	0.4000	24.8	1/400

Legend Abbreviations: SNV (single nucleotide variant); nSNV (nonsynonymous single nucleotide variant); UTR (untranslated region); NFCCR (Newfoundland Colorectal Cancer Registry); NL (Newfoundland & Labrador); SIFT (sorting tolerant from intolerant); T (tolerated); D (damaging); B (benign); POD (possibly damaging); PRD (probably damaging); NA (not applicable); MAF (minor allele frequency); AC (allele count); dbSNP (the single nucleotide polymorphism database) (available at https://www.ncbi.nlm.nih.gov/SNP/) (Assembly GRCh38.p7); rsID (reference SNP cluster ID); GnomAD (Genome Aggregation Database) (available at http://gnomad.broadinstitute.org/) (version r2.0.2); CADD (Combined Annotation Dependent Depletion); REVEL (Rare Exome Variant Ensemble Learner). Note: RefSeqs used included NM_024642.4; NP_078918.3; NC_000009.12.

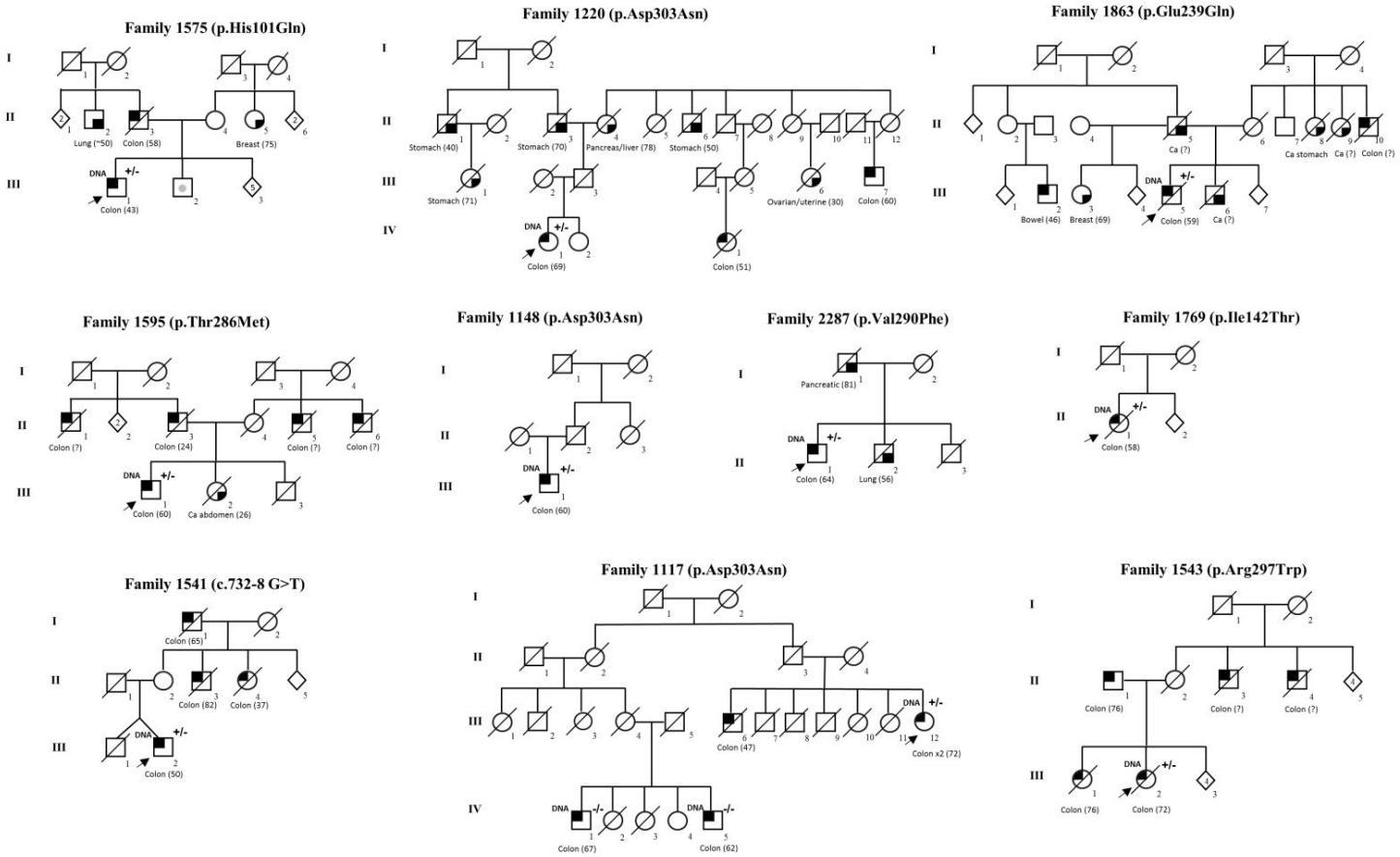


Figure 4.8 Families from the Newfoundland Colorectal Cancer Registry with rare GALNT12 variants. Reprinted from Evans *et al.,* (2018) with copyright permission.

Ten unrelated families harboring rare *GALNT12* alleles were identified in this study. Three families (1220, 1148, 1117) harbored the p.Asp303Asn allele, while the remaining variants were each present in probands of a single family. Eight families were MMR proficient and MSS, while two families (1575 and 1595) carried pathogenic MMR gene variants. Symbols: squares (males), circles (females), upper quadrant shaded (CRC), lower quadrant shaded (other cancer), grey shaded circle (polyps), diagonal line (deceased), arrow (proband), +/- (heterozygous variant). Age of cancer diagnoses are indicated in brackets. RefSeqs: NM_024642.4; NP_078918.3; NC_000009.12.

Supp Figure 2). Using multiple *in silico* tools (SIFT, Polyphen, CADD, REVEL), consequences of *GALNT12* variants were predicted (Table 4.1). Missense variants scored deleterious according to CADD, SIFT and Polyphen-2. CADD scores were highly deleterious, ranging from 24.8 (p.Asp303Asn) to 34.0 (p.Thr286Met).

GALNT12 variants clustered within or near the glycosyl-transferase domain encoded by *GALNT12* (Figure 4.9). Transferase activity of missense variants was characterized using an *in vitro*-derived peptide substrate assay (Figure 4.10). Three of the newly identified *GALNT12* missense variants (p.His101Gln, p.Ile142Thr, p.Val290Phe) demonstrated a marked loss (>2-fold reduction) of enzymatic activity compared to the wild-type protein ($P \leq 0.05$), while the p.Glu239Gln variant exhibited a ~2-fold reduction in enzymatic activity ($P = 0.077$). Additionally, we tested the p.Tyr396Cys allele identified in our previous study (Clarke *et al.,* 2012), which demonstrated a significant loss in GALNT12 activity ($P \leq 0.05$).

In silico tools (Spliceman, Human Splicing Finder, Spliceview) were used to predict consequences of the putative splice-altering variant (c.732-8G>T), as RNA was not accessible for this kindred living in a rural and remote area (Appendix Section 0 Supplementary Methods). Predictions were discordant, as Spliceman predicted high probability of disrupted splicing (L1 81%), whereas Human Splicing Finder predicted alteration of an exonic splice silencing site or alternatively no effect, while Spliceview predicted no influence on splicing.

The 8 rare *GALNT12* variants were derived from probands of 10 unrelated families, with p.Asp303Asn represented in 3 unrelated families (Figure 4.8). We

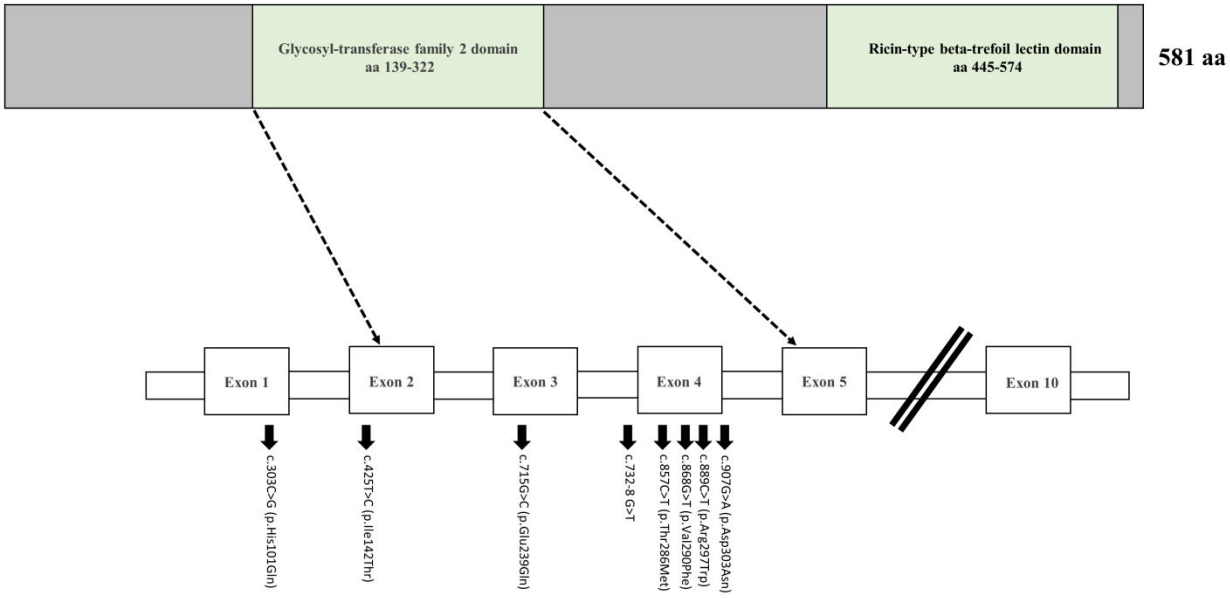


Figure 4.9 Rare germline variants of *GALNT12* identified in CRC patients cluster within and around the glycosyl-transferase domain of the protein. Reprinted from Evans *et al.,* (2018) with copyright permission.

Eight rare candidate *GALNT12* variants were identified in CRC cases in this study. Six out of the eight variants were situated within the glycosyl-transferase domain, while two (p.His101Gln, p.Ile142Thr) were located upstream. RefSeqs: NM_024642.4; NP_078918.3; NC_000009.12. Reprinted from Evans *et al.,* 2018 with Copyright permission.

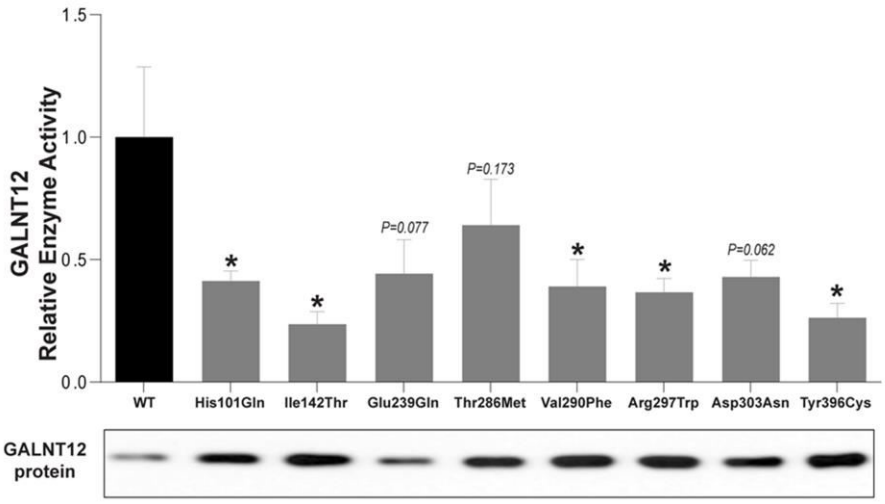


Figure 4.10 Biochemical characterization of wild-type and mutant GALNT12 proteins. Reprinted from Evans et al. (2018) with copyright permission.

Y-axis depicts the enzyme activities of mutant versions of GALNT12, relative to the wild-type (WT) protein (X-axis). Error bars represent standard error of the means derived from 3 independent replicate experiments. (*) indicates significant difference (P≤0.05) in enzyme activity of respective mutant GALNT12 as compared to the WT protein, estimated by one-tailed Student's t-test. Shown below the graph is a representative Western blot image demonstrating protein expression of wild-type and mutant versions of GALNT12 in SW480 cells transfected with respective cDNA constructs. The p.Tyr396Cys GALNT12 variant was identified in our previous clinical cohort-based study (Clarke *et al.,* 2012) and was included in this *in vitro* enzyme assay to test its enzymatic potential.

proceeded with familial co-segregation analysis to assess penetrance. DNA was not available for segregation analysis in 9 out of 10 families, and many relatives were deceased at the time of study. Segregation was tested in Family 1117 (Figure 4.8), as DNA was available for two distant CRC-affected relatives of the proband (first cousins once removed). Neither carried the p.Asp303Asn variant, and it is unclear whether the proband's risk for CRC derives from the paternal or maternal lineages. Our limited co-segregation analysis was therefore inconclusive, but not supportive of p.Asp303Asn segregation (Appendix Supp Figure 3). Given the limited DNA available for segregation analysis, we estimated the penetrance (Minikel *et al.,* 2016; Whiffin *et al.,* 2017) under different assumptions for the NL population (Appendix Section 0 Supplementary Methods). We observed evidence of moderate penetrance for these *GALNT12* variants (Appendix Section Supp Table 7).

Table 2 shows the clinical and molecular characteristics of CRC probands carrying the rare *GALNT12* variants. The age of proband CRC diagnosis ranged from 43-72 years. Patients of the NFCCR cohort were screened for MMR defects or MSI during prior studies (Woods *et al.,* 2010), and during this study, we had no *a priori* knowledge of patient MMR status. We found that the majority of probands with rare *GALNT12* variants (8 out of 10) were MSS and MMR proficient. Probands from two families (Figure 4.8: 1575, 1595) harbored pathogenic (InSiGHT class 5) MMR gene variants with high tumor MSI-H (Table 4.3). Finally, we assessed risk of genetic predisposition for CRC in all families using established risk criteria. Two families (Figure 1: 1595, 1541) fulfilled the AC-1 (Vasen *et al.,* 1991). Of these two AC-1

Table 4.3 Clinical and pathological characteristics of Newfoundland Colorectal Cancer Registry patients harboring rare candidate GALNT12 variants. Reprinted from Evans *et al.,* (2018) with copyright permission.

Patient	Family	Variant	Age of diagnosis	Tumor location	Microsatellite stability	MMR protein IHC	MMR variants	Risk criteria
III-1	1575	c.303C>G; p.His101Gln	43	Rectum	MSI-H	MSH6 deficient	*MSH6*: c.3514dupA; Arg1172Lysfs*5	Bethesda
II-1	1769	c.425T>C; p.Ile142Thr	58	Transverse	MSS	Intact	Negative	NA
III-5	1863	c.715G>C; p.Glu239Gln	59	Rectum	MSS	Intact	Negative	Bethesda
III-2	1541	g.98831764G>T; c.732-8 G>T	51	Transverse	MSS	Intact	Negative	AC-1,FCCTX, ACMAC, Bethesda
III-1	1595	c.857C>T; p.Thr286Met	60	Descending	MSI-H	MLH1 deficient	*MLH1*: c.793C>T; p.His264Leufs*2	AC-1, ACMAC, Bethesda
II-1	2287	c.868G>T; p.Val290Phe	64	Rectum	MSS	Intact	Negative	NA
III-2	1543	c.889C>T; Arg297Trp	72	Ascending	MSS	Intact	Negative	NA
III-1	1148	c.907G>A; p.Asp303Asn	72	Cecum	MSS	Intact	Negative	NA
IV-1	1220	c.907G>A; p.Asp303Asn	59	Descending	MSS	Intact	Negative	NA
III-12	1148	c.907G>A; p.Asp303Asn	60	Sigmoid	MSS	Intact	Negative	Bethesda

Legend: High microsatellite instability (MSI-H), microsatellite stable (MSS), immunohistochemistry (IHC), mismatch repair (MMR). Risk Criteria: Familial colorectal cancer type X (FCCTX), Revised Bethesda Criteria (Bethesda), Age and Cancer Modified Amsterdam Criteria (ACMAC). RefSeqs used include NM_024642.4; NP_078918.3; NC_000009.12.

families, the proband of Family 1595 was MSI-H with MLH1 deficient, whereas the proband of Family 1541 was MSS and MMR proficient; meeting criteria for FCCTX. A total of five families met the Revised Bethesda Criteria (Umar *et al.,* 2004). Two were the AC-1 families described above, and of the remaining three, one was MSI-H and MSH6 deficient, while two families were MSS and MMR intact.

4.8 Discussion

Here, we screened a population-based cohort of 479 incident CRC cases, identifying 8 rare *GALNT12* variants (Table 4.1). We observed a higher burden of these variants in CRC cases compared to population-matched CRC-free controls ($P \leq 0.05$). Multiple *in silico* tools (CADD, SIFT and Polyphen-2) predicted these missense variants are deleterious (Table 4.1), and variants clustered within or near the functionally relevant glycosyl-transferase domain (Figure 4.9). Transferase assays revealed GALNT12 mutants demonstrate a marked reduction of enzyme activity compared to wild-type (Figure 4.10).

To explore *GALNT12* variation in this study, we chose a rare variant MAF threshold of ≤1%. This allowed us to assess rare variants under reduced penetrance, and consider potential for NL founder effects. Recently, Whiffin et al proposed a strategy that leverages allele frequencies available in large population databases to assess 'maximum credible allele frequencies' (Whiffin *et al.,* 2017). Predicted maximum credible allele frequencies for CRC variants in NL are shown under different models of penetrance in Appendix Supp Table 6. For example, under 50%

penetrance, we estimated a MAF of no greater than 0.00997%, or 37 GnomAD alleles (See Appendix B, Supplementary Methods). This analysis however, does not consider the potential for genetic drift in our founder population. During the analysis, we also noted a limitation, as the most common known pathogenic CRC variant in NL, a founder variant in *MSH2* (p.Val265_Gln314del), greatly exceeds our calculated maximum credible allele frequencies, with a GnomAD MAF of 1.125% (330 alleles out of 29,332 total alleles; 18 homozygotes). This pathogenic variant is known to cause CRC in multiple families around the world, and explains up to 71% of CRC caused by *MSH2* in NL (Woods *et al.,* 2010). This finding likely arises from a very high number of false positive calls, due to low sequence complexity in this region; highlighting a confounding variable in using these large databases, which rely on next-generation DNA sequencing technology. We advocate for a cautious approach to relying on expected frequencies as a sole indicator of pathogenicity. Ultimately, we observed that a majority of *GALNT12* variants identified in this study were well below 0.02% global MAF.

Bioinformatic tools (SIFT, Polyphen-2) are well-established for predicting consequences of genetic variation (Dong *et al.,* 2015). SIFT assumes variation within evolutionarily conserved regions is more likely to impact protein function (Sim *et al.,* 2012), while Polyphen-2 considers sequence-based, phylogenetic and structural features to predict consequences (Adzhubei *et al.,* 2013). We observed agreement between SIFT and Polyphen-2 in predicting deleterious consequences for *GALNT12* missense variants (Table 4.1). The sole exception was p.Asp303Asn; which SIFT

predicted 'tolerated' while Polyphen-2 predicted 'probably damaging'. A limitation of these tools is that they cannot predict intronic variation. Newer ensemble prediction tools such as CADD and REVEL are gaining favor since they incorporate weighted estimates using multiple *in silico* tools, trained on machine-learning algorithms, to assist in variant interpretation (Ioannidis *et al.,* 2016; Kircher *et al.,* 2014). CADD scores have demonstrated clinical utility for interpreting MMR variants in a manually-curated dataset from InSiGHT (van der Velde *et al.,* 2015). We observed highly deleterious CADD scores (24.8-34.0) for all *GALNT12* missense variants (Table 4.2). REVEL scores predicted two *GALNT12* variants (p.Ile142Thr, p.Glu239Gln) are deleterious (score >0.5); thus CADD and REVEL were not concordant. Though both tools have merits, they are limited by circular errors that arise from benchmarking datasets (Grimm *et al.,* 2015). Overall, rare missense *GALNT12* variants were predicted deleterious according to three programs -- CADD, SIFT and Polyphen-2. Since these deleterious variants clustered around the glycosyl-transferase domain (Figure 4.9), we postulated that assaying enzyme activity might further delineate any consequences of these variants.

Accordingly, we characterized mutant enzyme activities using an *in-vitro* peptide derived assay (Figure 4.10). We observed that GALNT12 mutants showed a ≥ 2-fold reduction of enzyme activity in 7 of the variants (including the previously identified p.Tyr396Cys), with five mutants (p.His101Gln, p.Ile142Thr, p.Val290Phe, p.Arg297Trp, p.Tyr396Cys) reaching the threshold for statistical significance (P≤0.05). Moreover, GALNT12 expression was consistently higher in mutants

compared to wild-type, which demonstrated the lowest protein expression (Figure 4.10). This observation suggests our assay is likely to underestimate the reduction of enzyme activity in seen in mutants. The consistent differences in GALNT12 expression could result from a compensatory mechanism arising from altered protein stability. Taken together, our enzymatic assays provided corroborating evidence for the deleterious effects of *GALNT12* variants.

Due to our ascertainment method (incident cases), were not able to pursue familial co-segregation analysis for most *GALNT12* variants in this study; given that additional DNAs were not available for 9 out of 10 families. We tested segregation of p.Asp303Asn in Family 1117 (Figure 4.8). The results did not support segregation, though pedigree structure and limited DNA sample numbers rendered the analysis inconclusive (Appendix Supp Figure 3). Recently, Whiffin et al. proposed a strategy to estimate variant penetrance using allele frequencies in large population databases (Minikel *et al.,* 2016; Whiffin *et al.,* 2017). Using prevalence of CRC in NL and lifetime CRC risk in Canada (See Appendix B, Supplementary Methods), we estimated penetrance of *GALNT12* variants (Appendix Supp Table 7), and found *in silico* evidence for moderate penetrance. While these *in silico* estimates provide a broad estimate of penetrance (orders of magnitude), confounding factors include population stratification and sampling variance, especially for very low frequency alleles in the databases. While encouraging, a formal assessment of penetrance for *GALNT12* variants remains an important avenue for future work, and argues for pursuing these variants in CRC families from diverse populations.

The tumor characteristics and molecular profiles of probands harboring *GALNT12* variants are shown in Table 4.3. Without *a priori* knowledge of MMR or MSI status, we observed most probands with a rare *GALNT12* variant (8 out of 10) were MSS with MMR stable. Two probands (Figure 4.8: 1575, 1595) concomitantly carried one *GALNT12* variant (p.His101Gln or p.Thr286Met) with a pathogenic (InSiGHT class 5) MMR variant (*MSH6;* NM_000179.2: p.Arg1172Lysfs*5 or *MLH1* NM_000249.3: p.His264Leufs*2) respectively. GALNT12 p.His101Gln mutants demonstrated significant reduction in enzyme activity, while p.Thr286Met mutants did not (Figure 4.10). It is difficult to assess the role of these particular *GALNT12* variants, as the concomitant pathogenic MMR variants undoubtedly contributed to CRC predisposition in these LS families. One advantage to our population-based approach is that we did not exclude these patients with prior known causes of CRC, which is an exclusionary bias often present in studies exploring novel genes. Regardless, the concomitant MMR variants should not prompt exclusion of these *GALNT12* variants from further consideration, since this reasoning would overlook the influence of ascertainment bias and more importantly, population genetics. For example, there are previously documented cases of bilineal inheritance and trans-heterozygotes for autosomal dominant polycystic kidney disease segregating in the NL population (Pei *et al.,* 2001). Likewise, an NL family with autosomal dominant sensorineural hearing loss harbors a pathogenic *KCNQ4* deletion, with a likely second, undiscovered deafness gene in 10 other affected relatives (Abdelfatah *et al.,* 2013). Thus, two pathogenic variants for a monogenic disease can, and do co-occur

within NL families. Further study encompassing a broader array of populations will be useful to delineate the role of these *GALNT12* variants. Review of the CBioPortal database (www.cbioportal.org), an online database for somatic mutations in cancer tissues, did not identify these candidate variants in any colon tumors.

In total, three non-LS families met the Revised Bethesda Criteria, with one family meeting criteria for FCCTX (Table 4.3). Recently, germline DNA sequencing of *GALNT12* in a Spanish cohort of 103 FCCTX families failed to identify any rare candidate variants in this gene (Seguí *et al.,* 2014). While only 45.6% of families in the Seguí *et al.* (2014) study met AC-1 and were truly FCCTX, the differences between this study and ours underscore the importance in recognizing the influence of population genetics as well as ascertainment criteria. A negative screening study in one population does not refute a role for *GALNT12* in CRC.

Strong arguments exist for a role for *GALNT12* in CRC pathogenesis. Genetic evidence demonstrates that *GALNT12* resides within a CRC susceptibility locus, with both rare and functionally deleterious somatic and germline *GALNT12* variants found in CRC patients (Clarke *et al.,* 2012; Guda *et al.,* 2009). Moreover, GALNT12 is selectively expressed in intestinal epithelia and frequently downregulated in CRCs (Guo *et al.,* 2002, Guo *et al.,* 2004). Moreover, *GALNT12* has been identified as one of the key genes facilitating normal morphology and polarization in colon epithelial cells (Li *et al.,* 2017). GALNT12 is a key enzyme catalyzing the initiation of mucin type *O*-glycosylation (Bennett *et al.,* 2012); a pathway shown to play a critical role in CRC progression (Pinho and Reis, 2015; Stowell *et al.,* 2015). Mice deficient in

intestinal mucin-type *O*-glycans spontaneously develop colitis-associated colon cancer due to inflammation resulting from impaired mucin barrier function (Bergstrom *et al.,* 2016). Moreover, mice lacking Muc2 spontaneously develop adenomas and CRC (Velcich *et al.,* 2002). Finally, recent studies of human glycosylation-associated genes from CRC-cell lines have identified candidates that are significantly mutated in CRC (Venkitachalam *et al.,* 2016).

Taken together, our findings suggest a thorough investigation for any putative link between *GALNT12* and CRC progression is necessary. A broader approach for testing rare *GALNT12* alleles in diverse populations, and characterizing their biological function in pre-clinical animal models may help decipher a role of *GALNT12* in CRC pathogenesis. The discovery of moderate penetrance genes is an important avenue to explain the missing heritability in CRC. Although challenging, the characterization of these genes is poised to impart novel insights for interpreting cancer risk moving forward. Screening large and well-characterized cohorts, in unique populations, represents an important strategy moving forward.

4.9 References

Abdelfatah, N., McComiskey, D. A., Doucette, L., Griffin, A., Moore, S. J., Negrijn, C., … Young, T.-L. (2013). Identification of a novel in-frame deletion in KCNQ4 (DFNA2A) and evidence of multiple phenocopies of unknown origin in a family with ADSNHL. *European Journal of Human Genetics*, *21*(10), 1112–1119. https://doi.org/10.1038/ejhg.2013.5

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, *Chapter 7*, Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76

Ahnen, D. J. (2011). The American College of Gastroenterology Emily Couric Lecture—The Adenoma–Carcinoma Sequence Revisited: Has the Era of Genetic Tailoring Finally Arrived? *The American Journal of Gastroenterology*, *106*(2), 190–198. https://doi.org/10.1038/ajg.2010.423

Anaya, D. A., Chang, G. J., & Rodriguez-Bigas, M. A. (2008). Extracolonic Manifestations of Hereditary Colorectal Cancer Syndromes. *Clinics in Colon and Rectal Surgery*, *21*(4), 263–272. https://doi.org/10.1055/s-0028-1089941

Bansidhar, B. J. (2012). Extracolonic Manifestations of Lynch Syndrome. *Clinics in Colon and Rectal Surgery*, *25*(2), 103–110. https://doi.org/10.1055/s-0032-1313781

Benchabane, H., & Ahmed, Y. (2009). The Adenomatous Polyposis Coli Tumor Suppressor and Wnt signaling in the Regulation of Apoptosis. *Advances in Experimental Medicine and Biology*, *656*, 75–84.

Bennett, E. P., Mandel, U., Clausen, H., Gerken, T. A., Fritz, T. A., & Tabak, L. A. (2012). Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology*, *22*(6), 736–756. https://doi.org/10.1093/glycob/cwr182

Benson, A. B., Venook, A. P., Cederquist, L., Chan, E., Chen, Y.-J., Cooper, H. S., … Freedman-Cass, D. (2017). Colon Cancer, Version 1.2017, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network*, *15*(3), 370–398. https://doi.org/10.6004/jnccn.2017.0036

Bergstrom, K., Liu, X., Zhao, Y., Gao, N., Wu, Q., Song, K., … Xia, L. (2016). Defective Intestinal Mucin-Type O-Glycosylation Causes Spontaneous Colitis-Associated Cancer in Mice. *Gastroenterology*, *151*(1), 152-164.e11. https://doi.org/10.1053/j.gastro.2016.03.039

Bergstrom, K. S. B., & Xia, L. (2013). Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology*, *23*(9), 1026–1037. https://doi.org/10.1093/glycob/cwt045

Brandão, C., & Lage, J. (2015). Management of Patients with Hereditary Colorectal Cancer Syndromes. *GE Portuguese Journal of Gastroenterology*, *22*(5), 204–212. https://doi.org/10.1016/j.jpge.2015.06.003

Brockhausen, I. (2006). Mucin-type O-glycans in human colon and breast cancer: glycodynamics and functions. *EMBO Reports*, *7*(6), 599–604. https://doi.org/10.1038/sj.embor.7400705

Brockhausen, I., Schachter, H., & Stanley, P. (2009). O-GalNAc Glycans. In A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, … M. E. Etzler (Eds.), *Essentials of Glycobiology* (2nd ed.). Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1896/

Brosens, L. A. A., van Hattem, A., Hylind, L. M., Iacobuzio-Donahue, C., Romans, K. E., Axilbund, J., … Giardiello, F. M. (2007). Risk of colorectal cancer in juvenile polyposis. *Gut*, *56*(7), 965–967. https://doi.org/10.1136/gut.2006.116913

Canadian Cancer Society's Advisory Committee on Cancer Statistics. (2015). *Canadian Cancer Statistics 2015*. Toronto, ON: Canadian Cancer Society.

Canadian Task Force on Preventive Health Care. (2016). Recommendations on screening for colorectal cancer in primary care. *CMAJ : Canadian Medical Association Journal*, *188*(5), 340–348. https://doi.org/10.1503/cmaj.151125

Chubb, D., Broderick, P., Dobbins, S. E., Frampton, M., Kinnersley, B., Penegar, S., … Houlston, R. S. (2016). Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nature Communications*, *7*, 11883. https://doi.org/10.1038/ncomms11883

Church, J. M. (2014). Polymerase proofreading-associated polyposis: a new, dominantly inherited syndrome of hereditary colorectal cancer predisposition. *Diseases of the Colon and Rectum*, *57*(3), 396–397. https://doi.org/10.1097/DCR.0000000000000084

Clarke, E., Green, R. C., Green, J. S., Mahoney, K., Parfrey, P. S., Younghusband, H. B., & Woods, M. O. (2012). Inherited deleterious variants in GALNT12 are associated with CRC susceptibility. *Human Mutation*, *33*(7), 1056–1058. https://doi.org/10.1002/humu.22088

Colussi, D., Brandi, G., Bazzoli, F., & Ricciardiello, L. (2013). Molecular Pathways Involved in Colorectal Cancer: Implications for Disease Behavior and

Prevention. *International Journal of Molecular Sciences*, *14*(8), 16365–16385. https://doi.org/10.3390/ijms140816365

Dominguez-Valentin, M., Therkildsen, C., Silva, S. D., & Nilbert, M. (2015). Familial colorectal cancer type X: genetic profiles and phenotypic features. *Modern Pathology*, *28*(1), 30–36. https://doi.org/10.1038/modpathol.2014.49

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137. https://doi.org/10.1093/hmg/ddu733

Evans, D. R., Venkitachalam, S., Revoredo, L., Dohey, A. T., Clarke, E., Pennell, J. J., … Guda, K. (2018). Evidence for GALNT12 as a moderate penetrance gene for colorectal cancer. *Human Mutation*, *39*(8), 1092–1101. https://doi.org/10.1002/humu.23549

Gerken, T. A., Jamison, O., Perrine, C. L., Collette, J. C., Moinova, H., Ravi, L., … Tabak, L. A. (2011). Emerging paradigms for the initiation of mucin-type protein O-glycosylation by the polypeptide GalNAc transferase family of glycosyltransferases. *The Journal of Biological Chemistry*, *286*(16), 14493–14507. https://doi.org/10.1074/jbc.M111.218701

Gonzalez-Pons, M., & Cruz-Correa, M. (2015). Colorectal Cancer Biomarkers: Where Are We Now? [Research article]. https://doi.org/10.1155/2015/149014

Gray-McGuire, C., Guda, K., Adrianto, I., Lin, C. P., Natale, L., Potter, J. D., … Wiesner, G. L. (2010). Confirmation of linkage to and localization of familial colon cancer

risk haplotype on chromosome 9q22. *Cancer Research*, *70*(13), 5409–5418. https://doi.org/10.1158/0008-5472.CAN-10-0188

Green, R. C., Green, J. S., Buehler, S. K., Robb, J. D., Daftary, D., Gallinger, S., … Younghusband, H. B. (2007). Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Familial Cancer*, *6*(1), 53–62. https://doi.org/10.1007/s10689-006-9104-x

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., … Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, *36*(5), 513–523. https://doi.org/10.1002/humu.22768

Groen, E. J., Roos, A., Muntinghe, F. L., Enting, R. H., de Vries, J., Kleibeuker, J. H., … van Beek, A. P. (2008). Extra-Intestinal Manifestations of Familial Adenomatous Polyposis. *Annals of Surgical Oncology*, *15*(9), 2439–2450. https://doi.org/10.1245/s10434-008-9981-3

Guda, K., Moinova, H., He, J., Jamison, O., Ravi, L., Natale, L., … Markowitz, S. D. (2009). Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12921–12925. https://doi.org/10.1073/pnas.0901454106

Guo, J.-M., Chen, H.-L., Wang, G.-M., Zhang, Y.-K., & Narimatsu, H. (2004). Expression of UDP-GalNAc:Polypeptide N-Acetylgalactosaminyltransferase-12 in Gastric

and Colonic Cancer Cell Lines and in Human Colorectal Cancer. *Oncology*, *67*(3–4), 271–276. https://doi.org/10.1159/000081328

Guo, J.-M., Zhang, Y., Cheng, L., Iwasaki, H., Wang, H., Kubota, T., … Narimatsu, H. (2002). Molecular cloning and characterization of a novel member of the UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase family, pp-GalNAc-T12. *FEBS Letters*, *524*(1–3), 211–218.

Hahn, M. M., de Voer, R. M., Hoogerbrugge, N., Ligtenberg, M. J. L., Kuiper, R. P., & van Kessel, A. G. (2016). The genetic heterogeneity of colorectal cancer predisposition - guidelines for gene discovery. *Cellular Oncology (Dordrecht)*, *39*(6), 491–510. https://doi.org/10.1007/s13402-016-0284-6

Half, E., Bercovich, D., & Rozen, P. (2009). Familial adenomatous polyposis. *Orphanet Journal of Rare Diseases*, *4*, 22. https://doi.org/10.1186/1750-1172-4-22

Hassan, C., Repici, A., & Rex, D. K. (2016). Serrated polyposis syndrome: risk stratification or reduction? *Gut*, gutjnl-2015-311357. https://doi.org/10.1136/gutjnl-2015-311357

Hudson, S. V., Ferrante, J. M., Ohman-Strickland, P., Hahn, K. A., Shaw, E. K., Hemler, J., & Crabtree, B. F. (2012). Physician Recommendation and Patient Adherence for Colorectal Cancer Screening. *The Journal of the American Board of Family Medicine*, *25*(6), 782–791. https://doi.org/10.3122/jabfm.2012.06.110254

IJspeert, J. E. G., Rana, S. a. Q., Atkinson, N. S. S., van Herwaarden, Y. J., Bastiaansen, B. a. J., van Leerdam, M. E., … Dutch workgroup serrated polyps & polyposis (WASP). (2017). Clinical risk factors of colorectal cancer in patients with

serrated polyposis syndrome: a multicentre cohort analysis. *Gut*, *66*(2), 278–284. https://doi.org/10.1136/gutjnl-2015-310630

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., & Burt, R. W. (2010). Hereditary and familial colon cancer. *Gastroenterology*, *138*(6), 2044–2058. https://doi.org/10.1053/j.gastro.2010.01.054

Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B. J., … Hsu, L. (2014). Estimating the heritability of colorectal cancer. *Human Molecular Genetics*, *23*(14), 3898–3905. https://doi.org/10.1093/hmg/ddu087

Kantor, M., Sobrado, J., Patel, S., Eiseler, S., & Ochner, C. (2017). Hereditary Colorectal Tumors: A Literature Review on MUTYH-Associated Polyposis. *Gastroenterology Research and Practice*, *2017*. https://doi.org/10.1155/2017/8693182

Kemp, Z. E., Carvajal-Carmona, L. G., Barclay, E., Gorman, M., Martin, L., Wood, W., … Colorectal Tumour Gene Identification Study Consortium. (2006). Evidence of linkage to chromosome 9q22.33 in colorectal cancer kindreds from the United Kingdom. *Cancer Research*, *66*(10), 5003–5006. https://doi.org/10.1158/0008-5472.CAN-05-4074

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kohlmann, W., & Gruber, S. B. (1993). Lynch Syndrome. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, K. Stephens, & A. Amemiya (Eds.), *GeneReviews*®. Seattle (WA): University of Washington, Seattle. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1211/

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., … Watanabe, T. (2015). Colorectal cancer. *Nature Reviews Disease Primers*, *1*, 15065. https://doi.org/10.1038/nrdp.2015.65

Larsen Haidle, J., & Howe, J. R. (1993). Juvenile Polyposis Syndrome. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, K. Stephens, & A. Amemiya (Eds.), *GeneReviews*®. Seattle (WA): University of Washington, Seattle. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1469/

Leach, F. S., Nicolaides, N. C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., … Nyström-Lahti, M. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell*, *75*(6), 1215–1225.

Lech, G., Słotwiński, R., Słodkowski, M., & Krasnodębski, I. W. (2016). Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances. *World Journal of Gastroenterology*, *22*(5), 1745–1755. https://doi.org/10.3748/wjg.v22.i5.1745

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, C., Singh, B., Graves-Deal, R., Ma, H., Starchenko, A., Fry, W. H., … Coffey, R. J. (2017). Three-dimensional culture system identifies a new mode of cetuximab resistance and disease-relevant genes in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(14), E2852–E2861. https://doi.org/10.1073/pnas.1618297114

Lieberman, S., Walsh, T., Schechter, M., Adar, T., Goldin, E., Beeri, R., … Goldberg, Y. (2017). Features of Patients With Hereditary Mixed Polyposis Syndrome Caused by Duplication of GREM1 and Implications for Screening and Surveillance. *Gastroenterology*, *152*(8), 1876-1880.e1. https://doi.org/10.1053/j.gastro.2017.02.014

Lindor, N. M. (2009). Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. *Surgical Oncology Clinics of North America*, *18*(4), 637–645. https://doi.org/10.1016/j.soc.2009.07.003

Malloy, A., & Smith, R. R. (2015). Prophylactic surgery for gastrointestinal malignancies. *Translational Gastrointestinal Cancer*, *4*(5), 337–351.

Mannion, J. J. (1977). *The peopling of Newfoundland : essays in historical geography*. St. Johns, NL: Institute of Social and Economic Research, Memorial University

of Newfoundland. Retrieved from http://collections.mun.ca/cdm/ref/collection/cns/id/34124

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Minikel, E. V., Vallabh, S. M., Lek, M., Estrada, K., Samocha, K. E., Sathirapongsasuti, J. F., … MacArthur, D. G. (2016). Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine*, *8*(322), 322ra9. https://doi.org/10.1126/scitranslmed.aad5169

Nieminen, T. T., Abdel–Rahman, W. M., Ristimäki, A., Lappalainen, M., Lahermo, P., Mecklin, J.-P., … Peltomäki, P. (2011). BMPR1A Mutations in Hereditary Nonpolyposis Colorectal Cancer Without Mismatch Repair Deficiency. *Gastroenterology*, *141*(1), e23–e26. https://doi.org/10.1053/j.gastro.2011.03.063

Nieminen, T. T., O'Donohue, M.-F., Wu, Y., Lohi, H., Scherer, S. W., Paterson, A. D., … Peltomäki, P. (2014). Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. *Gastroenterology*, *147*(3), 595-598.e5. https://doi.org/10.1053/j.gastro.2014.06.009

Ohtsubo, K., & Marth, J. D. (2006). Glycosylation in cellular mechanisms of health and disease. *Cell*, *126*(5), 855–867. https://doi.org/10.1016/j.cell.2006.08.019

O'Riordan, J. M., O'Donoghue, D., Green, A., Keegan, D., Hawkes, L. A., Payne, S. J., … Winter, D. C. (2010). Hereditary mixed polyposis syndrome due to a BMPR1A mutation. *Colorectal Disease: The Official Journal of the Association of Coloproctology of Great Britain and Ireland*, *12*(6), 570–573. https://doi.org/10.1111/j.1463-1318.2009.01931.x

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, *45*(2), 136–144. https://doi.org/10.1038/ng.2503

Pan, J., Xin, L., Ma, Y.-F., Hu, L.-H., & Li, Z.-S. (2016). Colonoscopy Reduces Colorectal Cancer Incidence and Mortality in Patients With Non-Malignant Findings: A Meta-Analysis. *The American Journal of Gastroenterology*, *111*(3), 355–365. https://doi.org/10.1038/ajg.2015.418

Pei, Y., Paterson, A. D., Wang, K. R., He, N., Hefferton, D., Watnick, T., … St George-Hyslop, P. (2001). Bilineal disease and trans-heterozygotes in autosomal dominant polycystic kidney disease. *American Journal of Human Genetics*, *68*(2), 355–363.

Pinho, S. S., & Reis, C. A. (2015). Glycosylation in cancer: mechanisms and clinical implications. *Nature Reviews. Cancer*, *15*(9), 540–555. https://doi.org/10.1038/nrc3982

Poulsen, M. L. ., & Bisgaard, M. . (2008). MUTYH Associated Polyposis (MAP). *Current Genomics*, *9*(6), 420–435. https://doi.org/10.2174/138920208785699562

Rafiemanesh, H., Mohammadian-Hafshejani, A., Ghoncheh, M., Sepehri, Z., Shamlou, R., Salehiniya, H., … Makhsosi, B. R. (2016). Incidence and Mortality of Colorectal Cancer and Relationships with the Human Development Index across the World. *Asian Pacific Journal of Cancer Prevention: APJCP*, *17*(5), 2465–2473.

Rahman, P., Jones, A., Curtis, J., Bartlett, S., Peddle, L., Fernandez, B. A., & Freimer, N. B. (2003). The Newfoundland population: a unique resource for genetic investigation of complex diseases. *Human Molecular Genetics*, *12 Spec No 2*, R167-172. https://doi.org/10.1093/hmg/ddg257

Raptis, S., Mrkonjic, M., Green, R. C., Pethe, V. V., Monga, N., Chan, Y. M., … Bapat, B. (2007). MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. *Journal of the National Cancer Institute*, *99*(6), 463–474. https://doi.org/10.1093/jnci/djk095

Rivera, B., González, S., Sánchez-Tomé, E., Blanco, I., Mercadillo, F., Letón, R., … Urioste, M. (2011). Clinical and genetic characterization of classical forms of familial adenomatous polyposis: a Spanish population study. *Annals of Oncology*, *22*(4), 903–909. https://doi.org/10.1093/annonc/mdq465

Schulz, E., Klampfl, P., Holzapfel, S., Janecke, A. R., Ulz, P., Renner, W., … Sill, H. (2014). Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X. *Nature Communications*, *5*, 5191. https://doi.org/10.1038/ncomms6191

Seguí, N., Pineda, M., Navarro, M., Lázaro, C., Brunet, J., Infante, M., … Valle, L. (2014). GALNT12 is not a major contributor of familial colorectal cancer type X. *Human Mutation*, *35*(1), 50–52. https://doi.org/10.1002/humu.22454

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(Web Server issue), W452-457. https://doi.org/10.1093/nar/gks539

Skoglund, J., Djureinovic, T., Zhou, X.-L., Vandrovcova, J., Renkonen, E., Iselius, L., … Lindblom, A. (2006). Linkage analysis in a large Swedish family supports the presence of a susceptibility locus for adenoma and colorectal cancer on chromosome 9q22.32-31.1. *Journal of Medical Genetics*, *43*(2), e7. https://doi.org/10.1136/jmg.2005.033928

Snover, D., Ahnen, D., Burt, R., & Odze. (2010). Serrated polyps of the colon and rectum and serrated polyposis. In *WHO Classification of Tumours of the digestive system* (pp. 160–165). Lyon, France.

Stanesby, O., & Jenkins, M. (2017). Comparison of the efficiency of colorectal cancer screening programs based on age and genetic risk for reduction of colorectal cancer mortality. *European Journal of Human Genetics: EJHG*, *25*(7), 832–838. https://doi.org/10.1038/ejhg.2017.60

Stowell, S. R., Ju, T., & Cummings, R. D. (2015). Protein glycosylation in cancer. *Annual Review of Pathology*, *10*, 473–510. https://doi.org/10.1146/annurev-pathol-012414-040438

Thompson, B. A., Spurdle, A. B., Plazzer, J.-P., Greenblatt, M. S., Akagi, K., Al-Mulla, F., … On Behalf Of InSiGHT. (2014). Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nature Genetics*, *46*(2), 107–115. https://doi.org/10.1038/ng.2854

Tran, D. T., & Ten Hagen, K. G. (2013). Mucin-type O-glycosylation during development. *The Journal of Biological Chemistry*, *288*(10), 6921–6929. https://doi.org/10.1074/jbc.R112.418558

Tung, N., Domchek, S. M., Stadler, Z., Nathanson, K. L., Couch, F., Garber, J. E., … Robson, M. E. (2016). Counselling framework for moderate-penetrance cancer-susceptibility mutations. *Nature Reviews. Clinical Oncology*, *13*(9), 581–588. https://doi.org/10.1038/nrclinonc.2016.90

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Rüschoff, J., … Srivastava, S. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, *96*(4), 261–268.

Valle, L. (2014). Genetic predisposition to colorectal cancer: where we stand and future perspectives. *World Journal of Gastroenterology*, *20*(29), 9828–9849. https://doi.org/10.3748/wjg.v20.i29.9828

Valle, L. (2017). Recent Discoveries in the Genetics of Familial Colorectal Cancer and Polyposis. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, *15*(6), 809–819. https://doi.org/10.1016/j.cgh.2016.09.148

van der Velde, K. J., Kuiper, J., Thompson, B. A., Plazzer, J.-P., van Valkenhoef, G., de Haan, M., … InSiGHT Group. (2015). Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. *Human Mutation*, *36*(7), 712–719. https://doi.org/10.1002/humu.22798

Vasen, H. F., Mecklin, J. P., Khan, P. M., & Lynch, H. T. (1991). The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Diseases of the Colon and Rectum*, *34*(5), 424–425.

Velcich, A., Yang, W., Heyer, J., Fragale, A., Nicholas, C., Viani, S., … Augenlicht, L. (2002). Colorectal cancer in mice genetically deficient in the mucin Muc2. *Science (New York, N.Y.)*, *295*(5560), 1726–1729. https://doi.org/10.1126/science.1069094

Venkitachalam, S., Revoredo, L., Varadan, V., Fecteau, R. E., Ravi, L., Lutterbaugh, J., … Guda, K. (2016). Biochemical and functional characterization of glycosylation-associated mutational landscapes in colon cancer. *Scientific Reports*, *6*, 23642. https://doi.org/10.1038/srep23642

Wang, P. P., Dicks, E., Gong, X., Buehler, S., Zhao, J., Squires, J., … Parfrey, P. S. (2009). Validity of random-digit-dialing in recruiting controls in a case-control study. *American Journal of Health Behavior*, *33*(5), 513–520.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., … Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *19*(10), 1151–1158. https://doi.org/10.1038/gim.2017.26

Wiesner, G. L., Daley, D., Lewis, S., Ticknor, C., Platzer, P., Lutterbaugh, J., … Markowitz, S. D. (2003). A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 12961–12965. https://doi.org/10.1073/pnas.2132286100

Williams, J.-C. B., Hamilton, J. K., Shiller, M., Fischer, L., dePrisco, G., & Boland, C. R. (2012). Combined juvenile polyposis and hereditary hemorrhagic telangiectasia. *Proceedings (Baylor University. Medical Center)*, *25*(4), 360–364.

Woods, M. O., Younghusband, H. B., Parfrey, P. S., Gallinger, S., McLaughlin, J., Dicks, E., … Green, R. C. (2010). The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut*, *59*(10), 1369–1377. https://doi.org/10.1136/gut.2010.208462

Woods, Michael O., Hyde, A. J., Curtis, F. K., Stuckless, S., Green, J. S., Pollett, A. F., … Parfrey, P. S. (2005). High frequency of hereditary colorectal cancer in Newfoundland likely involves novel susceptibility genes. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *11*(19 Pt 1), 6853–6861. https://doi.org/10.1158/1078-0432.CCR-05-0726

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

Zhai, G., Zhou, J., Woods, M. O., Green, J. S., Parfrey, P., Rahman, P., & Green, R. C. (2016). Genetic structure of the Newfoundland and Labrador population: founder effects modulate variability. *European Journal of Human Genetics: EJHG*, *24*(7), 1063–1070. https://doi.org/10.1038/ejhg.2015.256

Chapter 5. Summary

Both collectively and individually, Mendelian disorders place substantial burden on human populations (Section 1.5). Genetic discovery of these disorders not only unlocks fundamental understanding into pathophysiology, but it imparts clinical knowledge that can have substantial impacts on patients, their families, their loved ones and even whole communities. The unique architecture of the NL population, a culturally and geographically diverse clustering of individuals (Sections 1.7), has played a pivotal role in the exploration and genetic discovery of many Mendelian disorders (Section 1.8).

In this work, I employed modern strategies to discover genetic variants for three Mendelian disorders in the genetically heterogeneous NL founder populations. These disorders included arRP (Chapter 2), WMS (Chapter 3) and CRC (Chapter 4). For arRP and WMS, family-based studies using whole exome sequencing supplemented by either linkage (arRP) or homozygosity mapping (WMS) were employed. Meanwhile, I employed a candidate gene study to investigate *GALNT12* in a cohort of CRC patients.

For the WMS and arRP studies, I hypothesized the existence of a highly penetrant monogenic cause underlying each disorder, given strong family histories and previous genetic knowledge for these conditions. The FORGE Canada Consortium provided whole exome sequencing infrastructure and bioinformatic support. In both studies, a novel and highly penetrant pathogenic variant not previously reported was identified in a causative gene known for each disorder.

With respect to the CRC study, I instead employed a candidate gene approach based on emerging evidence surrounding the genetic architecture of CRC (Section 4.1). Particularly of interest to me was exploring the role of moderate penetrance variants in candidate genes for CRC; given the decreasing likelihood that novel highly penetrant genes explain familial CRC inheritance, which was eloquently described by Chubb and colleagues (Chubb *et al.,* 2016). In addition, a separate investigation for novel highly penetrant genes in 30 NL FCCTX probands has not identified any obvious candidate CRC genes to date (Woods & Evans, unpublished data). The strength of the candidate gene approach in the *GALNT12* study rested on the unique attributes of the NFCCR cohort. The candidate gene approach allowed me to screen a relatively large series of cases (N=479) as compared to a whole exome approach, which would have limited the total number of samples sequenced due to financial constraints imposed by the technology.

While the *GALNT12* study differed from the arRP and WMS studies in methodology – they each faced similar challenges that highlight the shifting paradigm of modern genetic discovery (Section 1.3). Whereas historically, gene mapping and Sanger DNA sequencing were the central focus of discovery studies, the new paradigm focuses on management of vast quantities of data and interpreting variants (Section 1.4). Indeed, assessing variant pathogenicity from a plethora of candidates remains a central challenge moving forward.

With respect to the arRP study, multiple lines of evidence corroborate the pathogenicity of the *MERTK* deletion (Section 2.8). First, the deletion has never been

previously reported worldwide and is thus extremely rare. Second, the effect of the deletion is large: it is multi-exonic and removes 201 amino acids from the encoded protein, thereby having large impact on the 3D protein structure as compared to the effect of a missense variant. Third, prior animal and human studies have indicated a role for *MERTK* in photoreceptor viability. Finally, pedigree structure and segregation analysis corroborates pathogenicity in the family, as there was strong evidence of genetic linkage to this locus (LOD 4.89) and the deletion segregated as expected for autosomal recessive inheritance.

Likewise, for the WMS study, evidence for pathogenicity of the *ADAMTS17* variant is strong (Section 3.8). Although the variant is previously reported in population databases, the frequency is exceedingly rare (0.0081% in NHLBI and unreported in GnomAD). As compared to the *MERTK* deletion, the pathogenicity of p.C1023Y in *ADAMTS17* was initially more difficult to assess, as it is a missense variant. While multiple *in silico* tools corroborate a likely deleterious effect for the variant, further functional assays were necessary to demonstrate pathogenicity. Shown in Figure 3.13, these assays demonstrate a marked reduction of ADAMTS17 secreted to the ECM in cells with p.C1023Y compared to wildtype. This is functionally consistent with haploinsufficiency seen in previous studies, given that a genotype-phenotype correlation already exists between *ADAMTS17* and WMS. Taken together, in addition to the homozygosity mapping indicating the shared 2.5 Mb region in these patients, the evidence for pathogenicity for p.C1023Y is strong.

Meanwhile, variant interpretation in the CRC study faced unique challenges. Guidelines for interpreting moderate penetrance variants do not yet exist, as few studies have attempted to explore this area. Nevertheless, *GALNT12* variants were assessed with the same rigor as variants for high penetrance. For example, variants with very low MAF were pursued for investigation. Next, there was strong *a priori* evidence indicating *GALNT12* is implicated in CRC. Furthermore, we predicted deleterious consequences for the missense variants using similar tools for high penetrance variants, and we bolstered evidence for pathogenicity using functional *in vitro* assays. Familial segregation was not assessed in this study and is therefore a missing piece of evidence for assessment of pathogenicity, although segregation for a moderate penetrance variant would surely be more nuanced compared to segregation for a high penetrance variant. Overall, and in light of the limitations in our understanding with respect to moderate penetrance variants, the evidence regarding of the rare *GALNT12* variants is likely to be moderately supportive for pathogenicity as compared to the arRP and WMS studies. Despite this, the impacts of the *GALNT12* study, if further replicated globally, will be much greater moving forward.

The role of moderately penetrant variants is a largely unexplored area of genetics. In hereditary CRC research, and other areas, emphasis has always been on the highly penetrant variants -- given that these have the largest clinical impact and the associated risks are more easily interpreted for patients. In hereditary CRC, Chubb and colleagues demonstrated that although many of the highly penetrant

variants are now described (i.e. the "low-hanging fruit"), a large familial fraction of CRC remains that is unexplained by the highly penetrant variants. This emphasizes the importance of exploring moderate penetrance variants. Both polygenic and environmental factors likely contribute to the phenotypic expression of these variants. For example, polymorphisms in modifier genes as well as gene-environment interactions are already known to influence highly penetrant pathogenic mutations (Cooper *et al*., 2013).

Exploring moderate penetrance variants has been hindered by current dogma, which often requires strong familial inheritance patterns to suggest causality. It will likely be challenging to demonstrate convincing patterns of inheritance for moderate penetrance variants; a problem that will be exacerbated for late-onset disorders. Moderate penetrance variants will require higher degrees of investigation into molecular and functional mechanisms of pathogenicity, which can be resource intensive given the hundreds to thousands of candidate rare variants to be explored in a given family. Moreover, the role of genetic background on moderate penetrance variants could be a complicating factor. Finally, return of results to patients with these variants will remain a significant challenge to clinicians until such a time as they are better understood, particularly given the uncertainties surrounding prediction of risk for an individual patient and how that will influence their health care decisions. Ultimately, our advancing understanding of rare genetic variants is poised to dramatically alter the way we manage patients in the future, as further insights are gained.

5.1 Conclusion

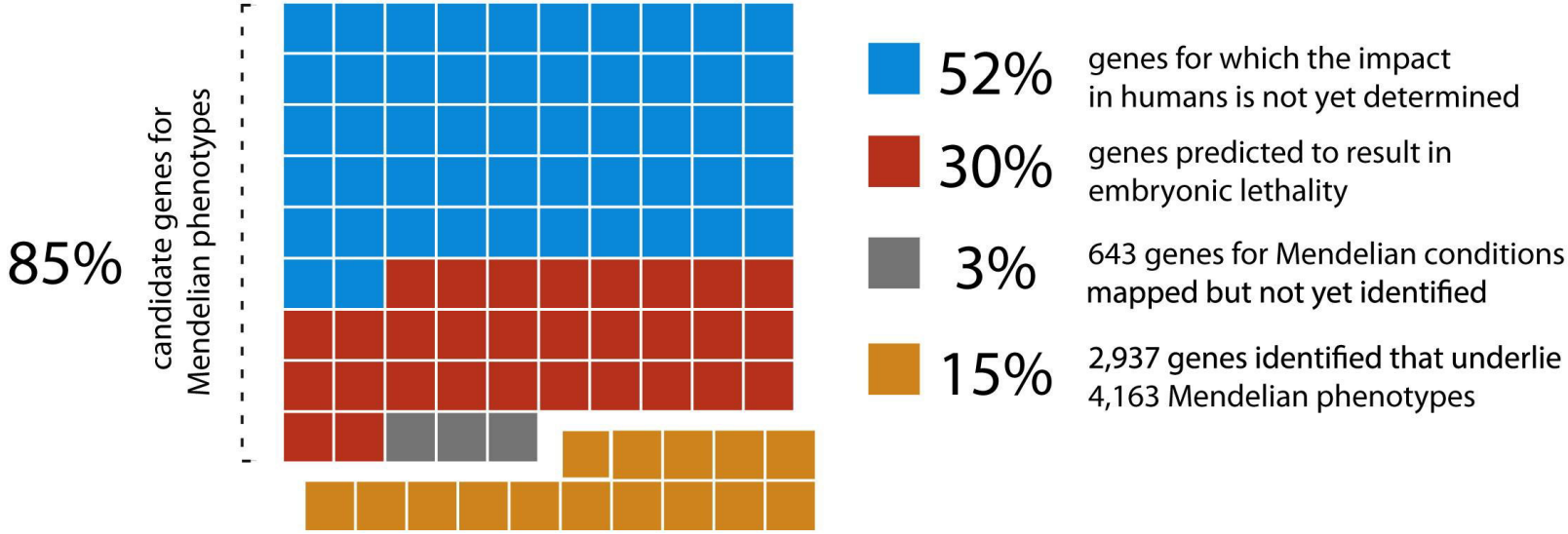
In conclusion, my investigation of three Mendelian disorders expands the collective understanding of Mendelian genetics in the NL founder population. From these studies, I conclude that the *MERTK* deletion is a new founder mutation, which explains RP in three families from the Northern Peninsula with a shared complex ancestry. The pathogenic *ADAMTS17* variant explains an exceedingly rare connective tissue disorder, and is not carried at high frequency among the healthy NL population, but demonstrates the aggregation of rare disorders in the province. The high incidence of rare *GALNT12* variants in the province could explain a fraction of CRC in NL. Further studies will be needed to bolster assessments of pathogenicity for *GALNT12* variants in other populations.

5.2 Future Directions

The widespread accessibily of whole exome sequencing has created new possibilities for exploring Mendelian disorders. There are likely to be many more discoveries on the horizon, given our limited understanding of the genome. Indeed, there are still many Mendelian disorders to explore by genetic investigation moving forward. Success rates of clinical diagnostic tests are one metric to shed light on our collective knowledge of Mendelian disorders, and to highlight how much there remains remains unsolved.

In 2014, a retrospective study found that traditional approaches (i.e. genetic testing) provided a diagnosis for only 46% of patients suspected of having a monogenic disorder, while another 53% had clinical findings consistent with a

Figure 5.1 Estimating the fraction of unexplored candidate genes implicated in Mendelian disorders. Reprinted from Chong *et al.,* (2015) with copyright permission.



genetic disorder, but no known pathogenic variant detected (Shashi *et al.,* 2014). Chong *et al.* (2015) summarize the work remaining to characterize Mendelian disorders with an intuitive analysis of the protein-coding genes in humans (Figure 5.1). Of their approximated 19,000 genes, at least 15% (2,937 genes in 2014) are known to underlie at least 4,163 Mendelian phenotypes. A further 3% (643) of genes map to disease regions, but have not yet been identified as causative in Mendelian disease. Studies of loss-of-function mutations in mice indicate that up to 30% of analogous genes produce embryonic lethality -- and are therefore candidates for lethal genetic disorders (Ayadi *et al.,* 2012). What remains is some 52% of genes with an unknown impact. This exemplifies our relatively poor understanding of the function of each of these coding genes and of their basic science, as well that Mendelian genetics has a strong role to play in this area. Furthermore, this example discusses only the genes in protein coding regions, but it is also likely that Mendelian disorders are influenced by non-coding and regulatory elements in much more complex, poorly understood ways. Therefore, it is likely that many genes remain to be studied, with an expanding pool of phenotypes to be characterized. Advances and increasing cost-effectiveness of other MPS technologies, such as whole genome sequencing will therefore also be integral to future genetic discoveries. Leveraging whole genome sequencing to its maximum potential will first require a broader understanding of the non-coding and regulatory elements of the genome.

5.3 References

Abouelhoda, M., Faquih, T., El-Kalioby, M., & Alkuraya, F. S. (2016). Revisiting the morbid genome of Mendelian disorders. *Genome Biology*, *17*(1), 235. https://doi.org/10.1186/s13059-016-1102-1

Ahmad, F., Li, D., Karibe, A., Gonzalez, O., Tapscott, T., Hill, R., … Roberts, R. (1998). Localization of a gene responsible for arrhythmogenic right ventricular dysplasia to chromosome 3p23. *Circulation*, *98*(25), 2791–2795.

Ahmad, S., Dahllund, L., Eriksson, A. B., Hellgren, D., Karlsson, U., Lund, P.-E., … Krupp, J. J. (2007). A stop codon mutation in SCN9A causes lack of pain sensation. *Human Molecular Genetics*, *16*(17), 2114–2121. https://doi.org/10.1093/hmg/ddm160

Allen, J. L. (1992). From Cabot to Cartier: The Early Exploration of Eastern North America, 1497-1543. *Annals of the Association of American Geographers*, *82*(3), 500–521.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(Database issue), D789-798. https://doi.org/10.1093/nar/gku1205

Amos, W., & Harwood, J. (1998). Factors affecting levels of genetic diversity in natural populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1366), 177–186.

Andermann, E., Jacob, J. C., Andermann, F., Carpenter, S., Wolfe, L., & Berkovic, S. F. (1988). The Newfoundland aggregate of neuronal ceroid-lipofuscinosis. *American Journal of Medical Genetics. Supplement*, *5*, 111–116.

Anjilvel, L. (1973). Familial hyperlipoproteinemia in an isolated part of Newfoundland. *Canadian Medical Association Journal*, *109*(9), 894-895 passim.

Ayadi, A., Birling, M.-C., Bottomley, J., Bussell, J., Fuchs, H., Fray, M., … Herault, Y. (2012). Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, *23*(9–10), 600–610. https://doi.org/10.1007/s00335-012-9418-y

Baird, P. A., Anderson, T. W., Newcombe, H. B., & Lowry, R. B. (1988). Genetic disorders in children and young adults: a population study. *American Journal of Human Genetics*, *42*(5), 677–693.

Baskin, B., Skinner, J. R., Sanatani, S., Terespolsky, D., Krahn, A. D., Ray, P. N., … Hamilton, R. M. (2013). TMEM43 mutations associated with arrhythmogenic right ventricular cardiomyopathy in non-Newfoundland populations. *Human Genetics*, *132*(11), 1245–1252. https://doi.org/10.1007/s00439-013-1323-2

Bassi, M. T., Schiaffino, M. V., Renieri, A., De Nigris, F., Galli, L., Bruttini, M., … Ballabio, A. (1995). Cloning of the gene for ocular albinism type 1 from the distal short

arm of the X chromosome. *Nature Genetics*, *10*(1), 13–19. https://doi.org/10.1038/ng0595-13

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., … Boycott, K. M. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, *94*(6), 809–817. https://doi.org/10.1016/j.ajhg.2014.05.003

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., … Kingsmore, S. F. (2011). Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. *Science Translational Medicine*, *3*(65), 65ra4. https://doi.org/10.1126/scitranslmed.3001756

Bosanquet, R. C., & Johnson, G. J. (1981). Peninsula Pupil: Anomaly Unique to Newfoundland and Labrador? *Archives of Ophthalmology*, *99*(10), 1824–1826. https://doi.org/10.1001/archopht.1981.03930020698015

Bourassa, C. V., Meijer, I. A., Merner, N. D., Grewal, K. K., Stefanelli, M. G., Hodgkinson, K., … Rouleau, G. A. (2012). VAMP1 mutation causes dominant hereditary spastic ataxia in Newfoundland families. *American Journal of Human Genetics*, *91*(3), 548–552. https://doi.org/10.1016/j.ajhg.2012.07.018

Canadian Cancer Society's Advisory Committee on Cancer Statistics. (2015). *Canadian Cancer Statistics 2015*. Toronto, ON: Canadian Cancer Society.

Charles, S. J., Green, J. S., Moore, A. T., Barton, D. E., & Yates, J. R. W. (1993). Genetic Mapping of X-Linked Ocular Albinism: Linkage Analysis in a Large

Newfoundland Kindred. *Genomics*, *16*(1), 259–261. https://doi.org/10.1006/geno.1993.1171

Check Hayden, E. (2014). Technology: The $1,000 genome. *Nature News*, *507*(7492), 294. https://doi.org/10.1038/507294a

Chiò, A., Borghero, G., Pugliatti, M., Ticca, A., Calvo, A., Moglia, C., … Italian Amyotrophic Lateral Sclerosis Genetic (ITALSGEN) Consortium. (2011). Large proportion of amyotrophic lateral sclerosis cases in Sardinia due to a single founder mutation of the TARDBP gene. *Archives of Neurology*, *68*(5), 594–598. https://doi.org/10.1001/archneurol.2010.352

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., … Bamshad, M. J. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *American Journal of Human Genetics*, *97*(2), 199–215. https://doi.org/10.1016/j.ajhg.2015.06.009

Christensen, A. H., Andersen, C. B., Tybjaerg-Hansen, A., Haunso, S., & Svendsen, J. H. (2011). Mutation analysis and evaluation of the cardiac localization of TMEM43 in arrhythmogenic right ventricular cardiomyopathy. *Clinical Genetics*, *80*(3), 256–264. https://doi.org/10.1111/j.1399-0004.2011.01623.x

Chubb, D., Broderick, P., Dobbins, S. E., Frampton, M., Kinnersley, B., Penegar, S., … Houlston, R. S. (2016). Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nature Communications*, *7*, 11883. https://doi.org/10.1038/ncomms11883

Churchill, D. N., Bear, J. C., Morgan, J., Payne, R. H., McManamon, P. J., & Gault, M. H. (1984). Prognosis of adult onset polycystic kidney disease re-evaluated. *Kidney International*, *26*(2), 190–193. https://doi.org/10.1038/ki.1984.154

Cooper D. N, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013). *Hum Genet*;132(10)-1077-1130.

Dewey, F. E., Pan, S., Wheeler, M. T., Quake, S. R., & Ashley, E. A. (2012). DNA sequencing: clinical applications of new DNA sequencing technologies. *Circulation*, *125*(7), 931–944. https://doi.org/10.1161/CIRCULATIONAHA.110.972828

Dipple, K. M., & McCabe, E. R. (2000). Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. *American Journal of Human Genetics*, *66*(6), 1729–1735. https://doi.org/10.1086/302938

Doucette, L., Green, J., Black, C., Schwartzentruber, J., Johnson, G. J., Galutira, D., & Young, T.-L. (2013). Molecular genetics of achromatopsia in Newfoundland reveal genetic heterogeneity, founder effects and the first cases of Jalili syndrome in North America. *Ophthalmic Genetics*, *34*(3), 119–129. https://doi.org/10.3109/13816810.2013.763993

Doucette, L., Green, J., Fernandez, B., Johnson, G. J., Parfrey, P., & Young, T.-L. (2011). A novel, non-stop mutation in FOXE3 causes an autosomal dominant form of variable anterior segment dysgenesis including Peters anomaly. *European*

Journal of Human Genetics, *19*(3), 293–299. https://doi.org/10.1038/ejhg.2010.210

Duggan, A. T., Harris, A. J. T., Marciniak, S., Marshall, I., Kuch, M., Kitchen, A., … Poinar, H. (2017). Genetic Discontinuity between the Maritime Archaic and Beothuk Populations in Newfoundland, Canada. *Current Biology*, *27*(20), 3149-3156.e11. https://doi.org/10.1016/j.cub.2017.08.053

Eichers, E. R., Green, J. S., Stockton, D. W., Jackman, C. S., Whelan, J., McNamara, J. A., … Katsanis, N. (2002). Newfoundland Rod-Cone Dystrophy, an Early-Onset Retinal Dystrophy, Is Caused by Splice-Junction Mutations in RLBP1. *American Journal of Human Genetics*, *70*(4), 955–964.

Eng, B., Greenlay, B., & Waye, J. S. (2009). Characterisation of the British α0-thalassaemia deletion: evidence of a founder effect in Newfoundland, Canada. *British Journal of Haematology*, *147*(1), 150–152. https://doi.org/10.1111/j.1365-2141.2009.07825.x

Fan, Y., Esmail, M. A., Ansley, S. J., Blacque, O. E., Boroevich, K., Ross, A. J., … Leroux, M. R. (2004). Mutations in a member of the Ras superfamily of small GTP-binding proteins causes Bardet-Biedl syndrome. *Nature Genetics*, *36*(9), 989–993. https://doi.org/10.1038/ng1414

Fernandez, B. A., Fox, G., Bhatia, R., Sala, E., Noble, B., Denic, N., … Woods, M. O. (2012). A Newfoundland cohort of familial and sporadic idiopathic pulmonary fibrosis patients: clinical and genetic features. *Respiratory Research*, *13*(1), 64. https://doi.org/10.1186/1465-9921-13-64

Froggatt, N., Green, J., Brassett, C., Evans, D., Bishop, D., Kolodner, R., & Maher, E. (1999). A common MSH2 mutation in English and North American HNPCC families: origin, phenotypic expression, and sex specific differences in colorectal cancer. *Journal of Medical Genetics*, *36*(2), 97–102.

Garcia, C. K., & Raghu, G. (2004). Inherited interstitial lung disease. *Clinics in Chest Medicine*, *25*(3), 421–433, v. https://doi.org/10.1016/j.ccm.2004.05.001

Garrod, A. (1928). The Lessons OF RARE MALADIES.: The Annual Oration delivered before the Medical Society of London on May 21st, 1928,. *The Lancet*, *211*(5465), 1055–1060. https://doi.org/10.1016/S0140-6736(00)99941-0

Goh G, Choi M (2012). Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics Inform*, 10(4):214-219.

Green, J. S., Bear, J. C., & Johnson, G. J. (1986). The burden of genetically determined eye disease. *The British Journal of Ophthalmology*, *70*(9), 696–699.

Green, J. S., Bowmer, M. I., & Johnson, G. J. (1986). Von Hippel-Lindau disease in a Newfoundland kindred. *CMAJ*, *134*(2), 133–138.

Green, J. S., & Johnson, G. J. (1983). Hereditary diseases as causes of blindness in Newfoundland: preliminary report. *Canadian Journal of Ophthalmology. Journal Canadien D'ophtalmologie*, *18*(6), 281–284.

Green, J. S., & Johnson, G. J. (1986). Congenital cataract with microcornea and Peters' anomaly as expressions of one autosomal dominant gene. *Ophthalmic Paediatrics and Genetics*, *7*(3), 187–194.

Green, J., Sheaves, A., Galutira, D., Whelan, J., Bautista, D., Younghusband, B., … Young, T. (2010). Molecular Etiology of Stargardt Disease in Newfoundland and Labrador. *Investigative Ophthalmology & Visual Science*, *51*(13), 2591–2591.

Green, Jane S., Parfrey, P. S., Harnett, J. D., Farid, N. R., Cramer, B. C., Johnson, G., … Pryse-Phillips, W. (1989). The Cardinal Manifestations of Bardet–Biedl Syndrome, a Form of Laurence–Moon–Biedl Syndrome. *New England Journal of Medicine*, *321*(15), 1002–1009. https://doi.org/10.1056/NEJM198910123211503

Green, R. C., Green, J. S., Buehler, S. K., Robb, J. D., Daftary, D., Gallinger, S., … Younghusband, H. B. (2007). Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Familial Cancer*, *6*(1), 53–62. https://doi.org/10.1007/s10689-006-9104-x

Grewal, K. K., Stefanelli, M. G., Meijer, I. A., Hand, C. K., Rouleau, G. A., & Ives, E. J. (2004). A founder effect in three large Newfoundland families with a novel clinically variable spastic ataxia and supranuclear gaze palsy. *American Journal of Medical Genetics. Part A*, *131*(3), 249–254. https://doi.org/10.1002/ajmg.a.30397

Grier, J., Hirano, M., Karaa, A., Shepard, E., & Thompson, J. L. P. (2018). Diagnostic odyssey of patients with mitochondrial disease. *Neurology: Genetics*, *4*(2). https://doi.org/10.1212/NXG.0000000000000230

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., … Martin, J. B. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, *306*(5940), 234–238. https://doi.org/10.1038/306234a0

Hanna, M. (1970). Congenital Deficiency of Factor Xiii: Report of a Family from Newfoundland with Associated Mild Deficiency of Factor Xii. *Pediatrics*, *46*(4), 611–619.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hildebrandt, F. (2010). Genetic kidney diseases. *Lancet (London, England)*, *375*(9722), 1287–1295. https://doi.org/10.1016/S0140-6736(10)60236-X

Hodgkinson, K. A., Connors, S. P., Merner, N., Haywood, A., Young, T.-L., McKenna, W. J., … Parfrey, P. S. (2013). The natural history of a genetic subtype of arrhythmogenic right ventricular cardiomyopathy caused by a p.S358L mutation in TMEM43. *Clinical Genetics*, *83*(4), 321–331. https://doi.org/10.1111/j.1399-0004.2012.01919.x

Hodgkinson, Kathleen A., Howes, A. J., Boland, P., Shen, X. S., Stuckless, S., Young, T.-L., … Connors, S. P. (2016). Long-Term Clinical Outcome of Arrhythmogenic Right Ventricular Cardiomyopathy in Individuals With a p.S358L Mutation in TMEM43 Following Implantable Cardioverter Defibrillator Therapy.

Circulation. Arrhythmia and Electrophysiology, *9*(3). https://doi.org/10.1161/CIRCEP.115.003589

Hodgkinson, Kathy A., Parfrey, P. S., Bassett, A. S., Kupprion, C., Drenckhahn, J., Norman, M. W., … Connors, S. P. (2005). The impact of implantable cardioverter-defibrillator therapy on survival in autosomal-dominant arrhythmogenic right ventricular cardiomyopathy (ARVD5). *Journal of the American College of Cardiology*, *45*(3), 400–408. https://doi.org/10.1016/j.jacc.2004.08.068

Huang, Y., Yu, S., Wu, Z., & Tang, B. (2014). Genetics of hereditary neurological disorders in children. *Translational Pediatrics*, *3*(2), 108–119. https://doi.org/10.3978/j.issn.2224-4336.2014.03.04

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Johnson, G. J., & Bosanquet, R. C. (1983). Spherophakia in a Newfoundland family: 8 years' experience. *Canadian Journal of Ophthalmology. Journal Canadien D'ophtalmologie*, *18*(4), 159–164.

Johnson, G. J., Gillan, J. G., & Pearce, W. G. (1971). Ocular albinism in Newfoundland. *Canadian Journal of Ophthalmology. Journal Canadien D'ophtalmologie*, *6*(4), 237–248.

Katsanis, N., Lewis, R. A., Stockton, D. W., Mai, P. M. T., Baird, L., Beales, P. L., … Lupski, J. R. (1999). Delineation of the Critical Interval of Bardet-Biedl

Syndrome 1 (BBS1) to a Small Region of 11q13, through Linkage and Haplotype Analysis of 91 Pedigrees. *The American Journal of Human Genetics*, *65*(6), 1672–1679. https://doi.org/10.1086/302684

Kaurah, P., MacMillan, A., Boyd, N., Senz, J., Luca, A. D., Chun, N., … Huntsman, D. (2007). Founder and Recurrent CDH1 Mutations in Families With Hereditary Diffuse Gastric Cancer. *JAMA*, *297*(21), 2360–2372. https://doi.org/10.1001/jama.297.21.2360

Kristiansson, K., Naukkarinen, J., & Peltonen, L. (2008). Isolated populations and complex disease gene identification. *Genome Biology*, *9*(8), 109. https://doi.org/10.1186/gb-2008-9-8-109

Lafreniere, R. G., MacDonald, M. L. E., Dube, M.-P., MacFarlane, J., O'Driscoll, M., Brais, B., … Samuels, M. E. (2004). Identification of a novel gene (HSN2) causing hereditary sensory and autonomic neuropathy type II through the Study of Canadian Genetic Isolates. *American Journal of Human Genetics*, *74*(5), 1064–1073. https://doi.org/10.1086/420795

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862-868. https://doi.org/10.1093/nar/gkv1222

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., … Venter, J. C. (2007). The Diploid Genome Sequence of an Individual Human. *PLOS Biology*, *5*(10), e254. https://doi.org/10.1371/journal.pbio.0050254

Lynch, H. T., Coronel, S. M., Okimoto, R., Hampel, H., Sweet, K., Lynch, J. F., … Chapelle, A. de la. (2004). A Founder Mutation of the MSH2 Gene and Hereditary Nonpolyposis Colorectal Cancer in the United States. *JAMA*, *291*(6), 718–724. https://doi.org/10.1001/jama.291.6.718

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., … Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, *508*(7497), 469–476. https://doi.org/10.1038/nature13127

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., … Harper, P. S. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, *72*(6), 971–983. https://doi.org/10.1016/0092-8674(93)90585-E

MacLeod, E. L., & Ney, D. M. (2010). Nutritional Management of Phenylketonuria. *Annales Nestlé*, *68*(2), 58–69. https://doi.org/10.1159/000312813

Mahoney, K., Buckley, D., Alam, M., Penney, S., Young, T.-L., Parfrey, P., & Moore, S. J. (2012). High incidence of pediatric idiopathic epilepsy is associated with familial and autosomal dominant disease in Eastern Newfoundland. *Epilepsy Research*, *98*(2–3), 140–147. https://doi.org/10.1016/j.eplepsyres.2011.09.003

Mahoney, K., Moore, S. J., Buckley, D., Alam, M., Parfrey, P., Penney, S., … Young, T.-L. (2009). Variable neurologic phenotype in a GEFS+ family with a novel

mutation in SCN1A. *Seizure*, *18*(7), 492–497. https://doi.org/10.1016/j.seizure.2009.04.009

Mannion, J. J. (1977). *The peopling of Newfoundland : essays in historical geography*. St. Johns, NL: Institute of Social and Economic Research, Memorial University of Newfoundland. Retrieved from http://collections.mun.ca/cdm/ref/collection/cns/id/34124

Marshall, I. (1998). *A History and Ethnography of the Beothuk*. McGill-Queen's Press - MQUP.

Marshall, W., Furey, M., Larsen, B., Rose, J., Sharratt, G., Sussex, B., & Virmani, V. (1988). Right Ventricular Cardiomyopathy and Sudden Death in Young People. *New England Journal of Medicine*, *319*(3), 174–176. https://doi.org/10.1056/NEJM198807213190312

Matute, D. R. (2013). The role of founder effects on the evolution of reproductive isolation. *Journal of Evolutionary Biology*, *26*(11), 2299–2311. https://doi.org/10.1111/jeb.12246

McBride, K. L., & Garg, V. (2010). Impact of Mendelian inheritance in cardiovascular disease. *Annals of the New York Academy of Sciences*, *1214*, 122–137. https://doi.org/10.1111/j.1749-6632.2010.05791.x

McKnight, A. J., Currie, D., & Maxwell, A. P. (2010). Unravelling the genetic basis of renal diseases; from single gene to multifactorial disorders. *The Journal of Pathology*, *220*(2), 198–216. https://doi.org/10.1002/path.2639

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics*, *80*(4), 588–604. https://doi.org/10.1086/514346

Meijer, I. A., Hand, C. K., Grewal, K. K., Stefanelli, M. G., Ives, E. J., & Rouleau, G. A. (2002). A Locus for Autosomal Dominant Hereditary Spastic Ataxia, SAX1, Maps to Chromosome 12p13. *American Journal of Human Genetics*, *70*(3), 763–769.

Merner, N. D., Hodgkinson, K. A., Haywood, A. F. M., Connors, S., French, V. M., Drenckhahn, J.-D., … Young, T.-L. (2008). Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 Is a Fully Penetrant, Lethal Arrhythmic Disorder Caused by a Missense Mutation in the TMEM43 Gene. *American Journal of Human Genetics*, *82*(4), 809–821. https://doi.org/10.1016/j.ajhg.2008.01.010

Milting, H., Klauke, B., Christensen, A. H., Müsebeck, J., Walhorn, V., Grannemann, S., … Anselmetti, D. (2015). The TMEM43 Newfoundland mutation p.S358L causing ARVC-5 was imported from Europe and increases the stiffness of the cell nucleus. *European Heart Journal*, *36*(14), 872–881. https://doi.org/10.1093/eurheartj/ehu077

Moore, S. J., Buckley, D. J., MacMillan, A., Marshall, H. D., Steele, L., Ray, P. N., … Parfrey, P. S. (2008). The clinical and genetic epidemiology of neuronal ceroid lipofuscinosis in Newfoundland. *Clinical Genetics*, *74*(3), 213–222. https://doi.org/10.1111/j.1399-0004.2008.01054.x

Moore, Susan J., Green, J. S., Fan, Y., Bhogal, A. K., Dicks, E., Fernandez, B. A., … Parfrey, P. S. (2005). Clinical and genetic epidemiology of Bardet-Biedl syndrome in Newfoundland: a 22-year prospective, population-based, cohort study. *American Journal of Medical Genetics. Part A*, *132A*(4), 352–360. https://doi.org/10.1002/ajmg.a.30406

Nagy, R., Sweet, K., & Eng, C. (2004). Highly penetrant hereditary cancer syndromes. *Oncogene*, *23*(38), 6445–6470. https://doi.org/10.1038/sj.onc.1207714

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., … Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, *42*(1), 30–35. https://doi.org/10.1038/ng.499

Norio, R., Nevanlinna, H. R., & Perheentupa, J. (1973). Hereditary diseases in Finland; rare flora in rare soul. *Annals of Clinical Research*, *5*(3), 109–141.

O'Dea, D., Murphy, S., Hefferton, D., & Parfrey, P. (1998). Higher risk for renal failure in first-degree relatives of white patients with end-stage renal disease: A population-based study. *American Journal of Kidney Diseases*, *32*(5), 794–801.

Parfrey, P., Davidson, W., & Green, J. (2002). Clinical and genetic epidemiology of inherited renal disease in Newfoundland., *61*(6), 1925–1934.

Parfrey, P. S., Bear, J. C., Morgan, J., Cramer, B. C., McManamon, P. J., Gault, M. H., … Reeders, S. T. (1990). The Diagnosis and Prognosis of Autosomal Dominant Polycystic Kidney Disease. *New England Journal of Medicine*, *323*(16), 1085–1090. https://doi.org/10.1056/NEJM199010183231601

Pearce, W. G., Johnson, G. J., & Sanger, R. (1971). Ocular albinism and Xg. *Lancet (London, England)*, *1*(7708), 1072.

Pei, Y., Paterson, A. D., Wang, K. R., He, N., Hefferton, D., Watnick, T., … St George-Hyslop, P. (2001). Bilineal disease and trans-heterozygotes in autosomal dominant polycystic kidney disease. *American Journal of Human Genetics*, *68*(2), 355–363.

Pfeiffer, C. J., Fodor, J. G., & Canning, E. (1973). An epidemiologic analysis of mortality and gastric cancer in Newfoundland. *Canadian Medical Association Journal*, *108*(11), 1374–1380.

Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., & Inoue, I. (2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *Journal of Human Genetics*, *57*(10), 621–632. https://doi.org/10.1038/jhg.2012.91

Rahman, P., Jones, A., Curtis, J., Bartlett, S., Peddle, L., Fernandez, B. A., & Freimer, N. B. (2003). The Newfoundland population: a unique resource for genetic investigation of complex diseases. *Human Molecular Genetics*, *12 Spec No 2*, R167-172. https://doi.org/10.1093/hmg/ddg257

Rattner, A., Sun, H., & Nathans, J. (1999). Molecular genetics of human retinal disease. *Annual Review of Genetics*, *33*, 89–131. https://doi.org/10.1146/annurev.genet.33.1.89

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the

interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., … Chou, J. L. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science (New York, N.Y.)*, *245*(4922), 1066–1073.

Royer-Pokora, B., Kunkel, L. M., Monaco, A. P., Goff, S. C., Newburger, P. E., Baehner, R. L., … Orkin, S. H. (1986). Cloning the gene for an inherited human disorder--chronic granulomatous disease--on the basis of its chromosomal location. *Nature*, *322*(6074), 32–38. https://doi.org/10.1038/322032a0

Russell, J. F., Fu, Y.-H., & Ptáček, L. J. (2013). Episodic neurologic disorders: syndromes, genes, and mechanisms. *Annual Review of Neuroscience*, *36*, 25–50. https://doi.org/10.1146/annurev-neuro-062012-170300

Schwarze, K., Buchanan, J., Taylor, J. C., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine*. https://doi.org/10.1038/gim.2017.247

Scriver, C. R. (2001). Human Genetics: Lessons from Quebec Populations. *Annual Review of Genomics and Human Genetics*, *2*(1), 69–101. https://doi.org/10.1146/annurev.genom.2.1.69

Shashi, V., McConkie-Rosell, A., Rosell, B., Schoch, K., Vellore, K., McDonald, M., … Goldstein, D. G. (2014). The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *16*(2), 176–182. https://doi.org/10.1038/gim.2013.99

Sonati, M. de F., & Costa, F. F. (2008). The genetics of blood disorders: hereditary hemoglobinopathies. *Jornal De Pediatria*, *84*(4 Suppl), S40-51. https://doi.org/doi:10.2223/JPED.1802

Spirio, L., Green, J., Robertson, J., Robertson, M., Otterud, B., Sheldon, J., … Leppert, M. (1999). The identical 5' splice-site acceptor mutation in five attenuated APC families from Newfoundland demonstrates a founder effect. *Human Genetics*, *105*(5), 388–398. https://doi.org/10.1007/s004399900153

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., … Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 665–677. https://doi.org/10.1007/s00439-017-1779-6

Stuckless, S., Parfrey, P. S., Woods, M. O., Cox, J., Fitzgerald, G. W., Green, J. S., & Green, R. C. (2007). The phenotypic expression of three MSH2 mutations in large Newfoundland families with Lynch syndrome. *Familial Cancer*, *6*(1), 1–12. https://doi.org/10.1007/s10689-006-0014-8

Thompson, B. A., Spurdle, A. B., Plazzer, J.-P., Greenblatt, M. S., Akagi, K., Al-Mulla, F., … On Behalf Of InSiGHT. (2014). Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nature Genetics*, *46*(2), 107–115. https://doi.org/10.1038/ng.2854

Tilstra, D. J., & Byers, P. H. (1994). Molecular basis of hereditary disorders of connective tissue. *Annual Review of Medicine*, *45*, 149–163. https://doi.org/10.1146/annurev.med.45.1.149

Verma, I. C., & Puri, R. D. (2015). Global burden of genetic disease and the role of genetic screening. *Seminars in Fetal and Neonatal Medicine*, *20*(5), 354–363. https://doi.org/10.1016/j.siny.2015.07.002

Voelkerding, K. V., Dames, S. A., & Durtschi, J. D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, *55*(4), 641–658. https://doi.org/10.1373/clinchem.2008.112789

Wang, J., & Shen, Y. (2014). When a “Disease-Causing Mutation” Is Not a Pathogenic Variant. *Clinical Chemistry*, *60*(5), 711–713. https://doi.org/10.1373/clinchem.2013.215947

Warden, G., Harnett, D., Green, J., Wish, T., Woods, M. O., Green, R., … Parfrey, P. (2013). A population-based study of hereditary non-polyposis colorectal cancer: evidence of pathologic and genetic heterogeneity. *Clinical Genetics*, *84*(6), 522–530. https://doi.org/10.1111/cge.12080

Webb, D., Mathews, A., Harris, M., Muir, I., Hostetter, J., Marshall, W., … Johnson, G. (1978). Myotonia dystrophica: unusual features in a Labrador family. *Canadian Medical Association Journal*, *118*(5), 497–552.

Wertheim-Tysarowska, K., Gos, M., Sykut-Cegielska, J., & Bal, J. (2015). Genetic analysis in inherited metabolic disorders--from diagnosis to treatment. Own experience, current state of knowledge and perspectives. *Developmental Period Medicine*, *19*(4), 413–431.

Woods, M. O., Young, T.-L., Parfrey, P. S., Hefferton, D., Green, J. S., & Davidson, W. S. (1999). Genetic Heterogeneity of Bardet–Biedl Syndrome in a Distinct Canadian Population: Evidence for a Fifth Locus. *Genomics*, *55*(1), 2–9. https://doi.org/10.1006/geno.1998.5626

Xie, Y. G., Zheng, H., Leggo, J., Scully, M. F., & Lillicrap, D. (2002). A founder factor VIII mutation, valine 2016 to alanine, in a population with an extraordinarily high prevalence of mild hemophilia A. *Thrombosis and Haemostasis*, *87*(1), 178–179.

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., … Tyler-Smith, C. (2012). Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale

Resequencing. *American Journal of Human Genetics*, *91*(6), 1022–1032. https://doi.org/10.1016/j.ajhg.2012.10.015

Young, T. L. (2003). Ophthalmic genetics/inherited eye disease. *Current Opinion in Ophthalmology*, *14*(5), 296–303.

Young, T.-L., Woods, M. O., Parfrey, P. S., Green, J. S., Hefferton, D., & Davidson, W. S. (1999). A Founder Effect in the Newfoundland Population Reduces the Bardet-Biedl Syndrome I (BBS1) Interval to 1 cM. *American Journal of Human Genetics*, *65*(6), 1680–1687.

Zeegers, M. P. A., van Poppel, F., Vlietinck, R., Spruijt, L., & Ostrer, H. (2004). Founder mutations among the Dutch. *European Journal of Human Genetics: EJHG*, *12*(7), 591–600. https://doi.org/10.1038/sj.ejhg.5201151

Zhai, G., Zhou, J., Woods, M. O., Green, J. S., Parfrey, P., Rahman, P., & Green, R. C. (2016). Genetic structure of the Newfoundland and Labrador population: founder effects modulate variability. *European Journal of Human Genetics: EJHG*, *24*(7), 1063–1070. https://doi.org/10.1038/ejhg.2015.256

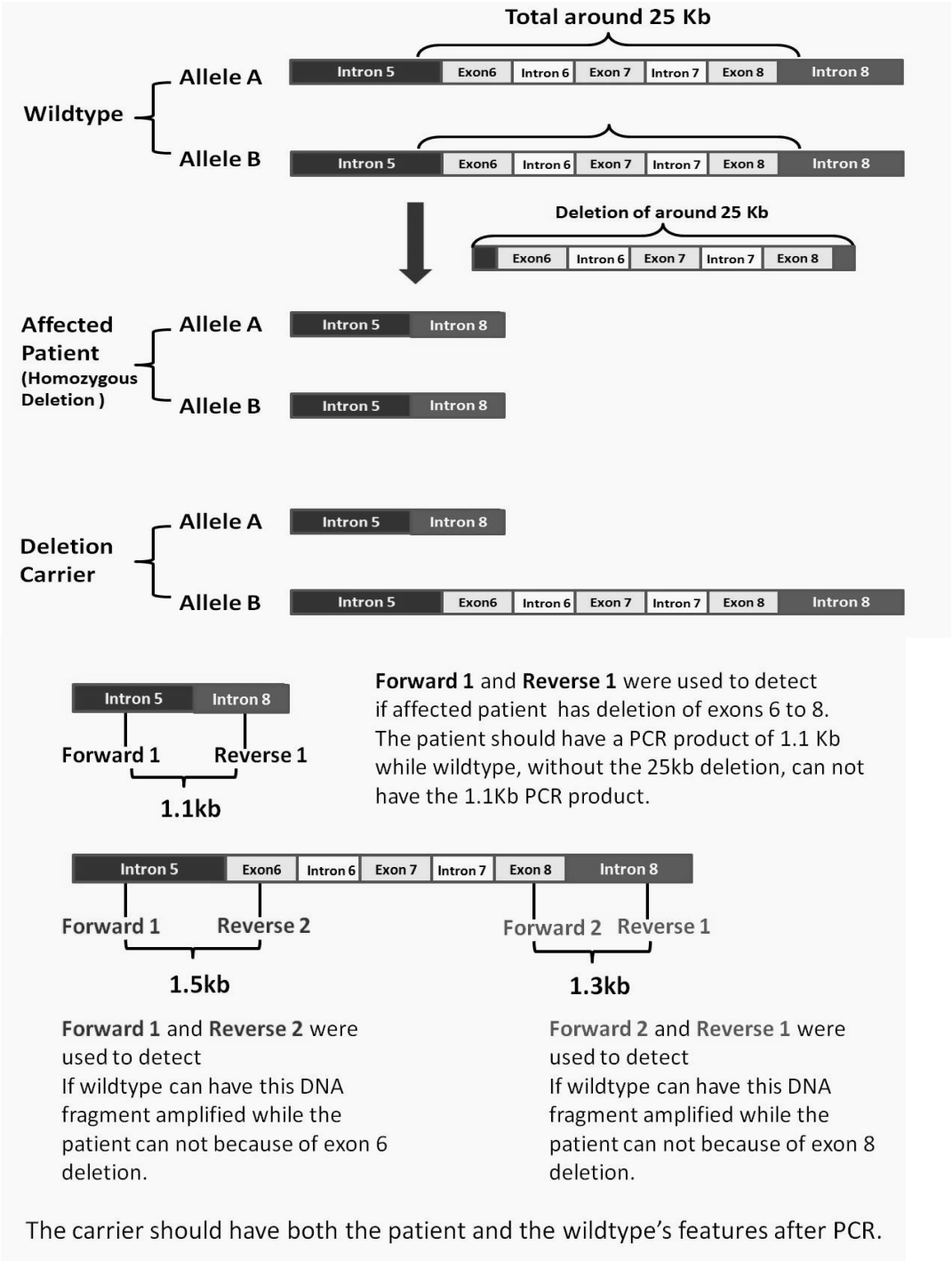
Appendix

Appendix A: Supporting Materials for arRP Study

Supplementary Methods

The detect this deletion, a custom PCR protocol was optimized using the QIAGEN HotStarTaq *Plus* DNA polymerase assay kit (cat. no. 203605) and dNTPs from the QIAGEN PCR grade dNTP set (cat. no. 201913). dNTPs were diluted to a working concentration of 2 mM. Primers were diluted to a working concentration of 5 µM. The 25 µl PCR reaction consisted of 5 µl of Q solution, 2.5 µl of 10x buffer, 2.5 µl of dNTPs, 2.5 µl of forward primer, 2.5 µl of reverse primer, 0.15 µl of HotStarTaq, 8.85 µl of dH_2O and 1 µl of template DNA at 100 ng/µl. The thermocycler protocol began with a denaturation stage of 95 °C for 5 minutes, and was followed by 40 cycles of 95 °C for 5 minutes, 60 °C for 30 seconds and 72 °C for 2.5 minutes, with a final elongation phase at 72 °C for 10 minutes. PCR products were purified using an Exo-Sap protocol following manufacturer's specifications (Affimetrix, Santa Clara, CA, USA) with exonuclease I (cat. no. 70073Z) and shrimp alkaline phosphatase (cat. no. 78390). Sanger sequencing was performed using the BigDye Terminator v.3.1 sequencing kit (cat. no. 4337454) under standard protocols for ABI automated direct sequencing on an ABI 3130xl genetic analyzer. Raw sequence data was processed using Sequencing Analysis version 5.2 (Applied Biosystems Inc., Foster City, CA, USA) and sequence chromatograms were visualized using Sequencher version 5.0 (Gene Codes Corporation, Ann Arbor, MI, USA.

Supplementary Tables & Figures



Supp Figure 2 Primer design for PCR experiment used to detect wildtype individuals from carriers and non-carriers of the *MERTK* deletion. Reprinted from Evans *et al.,* (2017) with copyright permission.

Supp Table 1 Summary of clinical features in an arRP family, segregating a large *MERTK* deletion.Reprinted from Evans *et al.,* (2017) with copyright permission.

Pedigree ID	Earliest VA (age)	Earliest VF (age)	First VA decline (age)	First VF decline (age)	Final VA (age)	Final VF (age)	Nyctalopia (age of onset)	Bone spicules	Optic disc pallor	BV thinning	Color defects (age)	Cataracts (age)	Fundus (age)
V-11	20/120 OU (28)	Central field loss (28)	HM OU (30)	NA	HM OU (30)	NA	NA	Y	N	N	NA	posterior capsular OS (35)	Bronze beaten with granular grey-white flecks in midperiphery (35).
VI-10	20/20 OU (11)	temporal constriction OS, full field OD (11)	20/25 OU (12)	Large central scotoma with peripheral islands of vision nasally (19)	HM OU (37)	Central scotomas to 30-40 degrees OU (37)	Y (always)	N	NA	NA	10/20 Ishihara (11), 19/20 ishihara (16)	N	Albinoid with marked decrease in chorioretinal pigmentation (11), salt and pepper appearance (12), extensive chorioretinal degeneration (37)
VI-3	NA (7)	central field constriction (7)	NA (7)	legally blind (12)	LP OU (35)	LP OU (35)	Y(7)	Y	Y	Y	Y(12)	N	Dense bone spicules at equator, atrophic area in mid-periphery with greyish spots. Macular degeneration with small atrophic area of choroidal sclerosis.
VI-9	20/20 OD (14)	full fields OU (14)	20/80 OD, 20/40 OS (17)	Significant central loss OU, peripheral islands nasally (20)	CF OU (30)	CF OU (30)	Y (always)	Y	Y	Y	N	N	Decreased chorioretinal pigmentation and stippling (14)

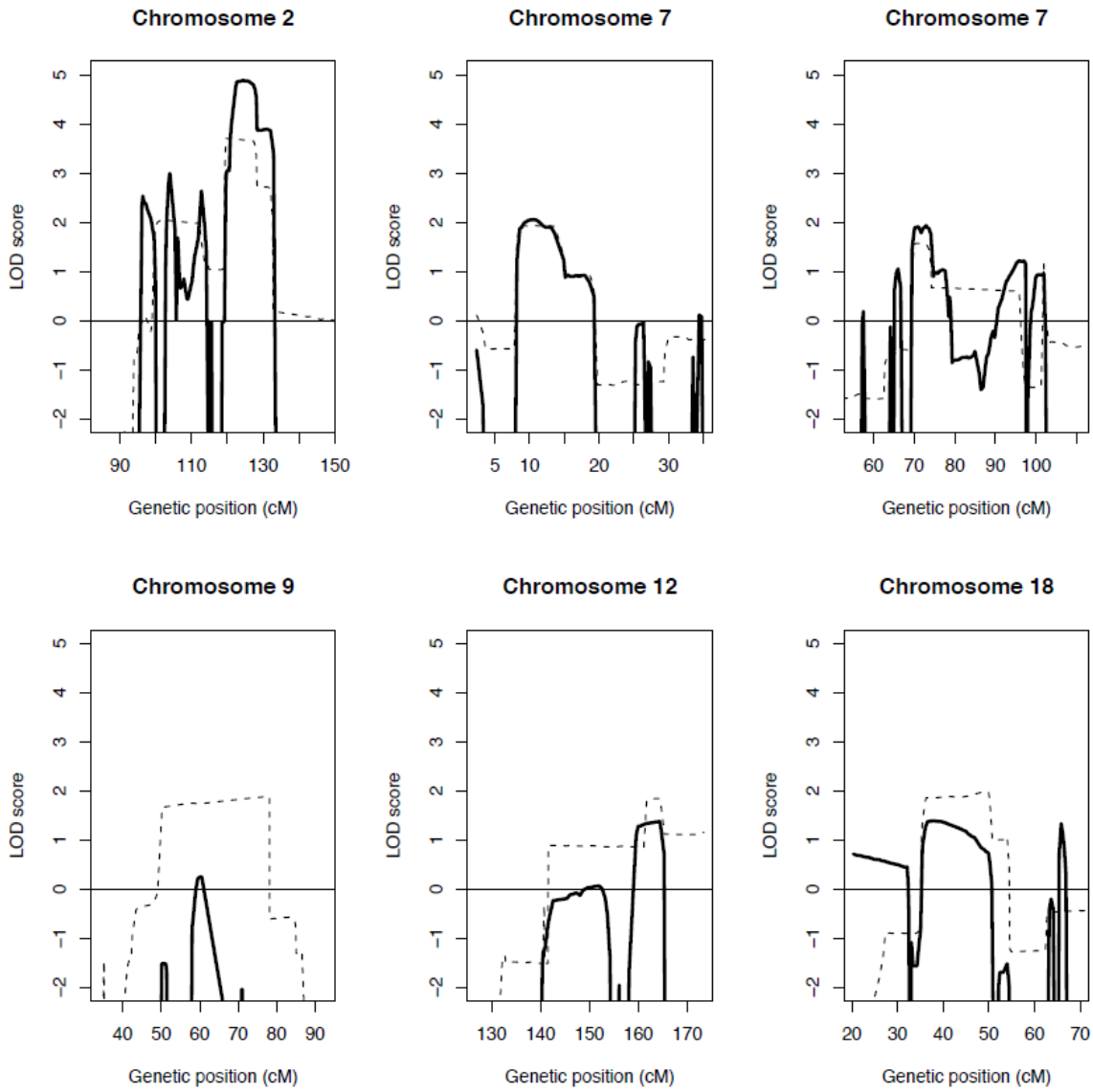
Legend: Abbreviations used in this table indicate the following: visual acuity (VA), visual fields (VF), blood vessel (BV), not available (NA), features not present (N), features present (Y), left eye (OS), right eye (OD), both eyes (OU), counting fingers (CF), hand movements only (HM), light perception only (LP).

Supp Table 2 Exact linkage analysis of sub-pedigrees, displaying regions with LOD scores > 1.5 under recessive model of inheritance with 99% penetrance and disease allele frequency of 0.1. Reprinted from Evans *et al.,* (2017) with copyright permission.

Chr	Max LOD	Start			End		
		SNP	cM	bp	SNP	cM	bp
2	3.71	rs7584136	119.23	105833093	rs10177102	131.20	121145352
7	1.93	rs2341587	8.17	3638547	rs6967307	17.99	8900786
7	1.58	rs874321	69.45	47119985	rs757134	95.83	82286106
9	1.89	rs1571649	50.14	27332892	rs1328430	78.13	82646539
12	1.84	rs1169303	141.59	121436376	rs11147178*	173.40	133531127
18	1.98	rs1918672	35.44	10808723	rs12104046	53.99	26971305

Supp Table 3 Refinement of six linked regions using approximate linkage analysis of the full family pedigree. Reprinted from Evans *et al.*, (2017) with copyright permission.

Chr	Max LOD	Start			End		
		SNP	cM	bp	SNP	cM	bp
2	4.89	rs4676093	120.79	107493123	rs4525709	128.32	117426163
7	2.05	rs4544975	7.80	3385384	rs10249817	15.13	7851307
7	1.94	rs874321	69.45	47119985	rs10238949	74.74	52245441
9	0.25	rs1408466	57.61	35992302	rs449851	65.91	70984372
12	1.38	rs3851661	158.13	127724123	rs10847680	165.51	129231326
18	1.39	rs676809	35.04	10687717	rs9954300	50.65	23307761



Supp Figure 3 The five chromosomes displaying genetic linkage in an arRP family. Dashed lines indicate exact calculation using two sub-pedigrees while solid lines indicate subsequent approximate linkage calculation using the full pedigree. Reprinted from Evans *et al.,* (2017) with copyright permission.

Supp Table 4 Candidate CNVs from whole exome data of patient VI-9, identified using FishingCNV. Reprinted from (Evans *et al.,* (2017) with copyright permission.

Chromosome	Start (bp)	End (bp)	Log2 fold change	p (Holm-Bonferroni)	Gene
16	72821063	72822747	4.0528	4.67E-49	*ZFHX3*
2	1.13E+08	1.13E+08	-5.1555	8.67E-29	*MERTK*
2	69723117	69723212	0.6295	0.006970274	*AAK1*
2	2.17E+08	2.17E+08	0.3469	0.016480077	*SMARCAL1*
10	32807344	32832313	0.5602	0.024287399	*CCDC7*
12	1.1E+08	1.1E+08	0.7673	0.024453384	*GIT2*
11	89914788	89914869	0.5903	0.02890554	*NAALAD2*
9	40784477	40784603	0.7005	0.029094809	*ZNF658*
1	60106952	60107023	-0.9596	0.036795731	*FGGY*

Appendix B: Supporting Materials for CRC study

Supplementary Methods

Splice Site Prediction

Three *in silico* tools were used (Spliceman, Human Splicing Finder and SpliceView) to predict consequences for the intronic *GALNT12* variant (NM_024642.4: c.732-8 G>T). Spliceman predicts the likelihood that distant mutations around splice sites disrupt splicing. Spliceman was accessed online (http://fairbrother.biomed.brown.edu/spliceman/index.cgi) on March 22nd 2018. The 'species' chosen from the input menu was Human (hg19). Additional options (reverse complement, masked splice site positions) were unselected (i.e. per default settings). The *GALNT12* c.732-8 G>T variant input sequence (tgggt(g/t)ctttc) was entered in FASTA format. Spliceman calculates an L1 percentile rank, which predicts higher likelihood that a variant disrupts splicing. Using these parameters, Spliceman predicts an 83% percentile for the c.732-8 G>T variant disrupting splicing.

Next, we calculated splice consequences of c.732-8 G>T using Human Splicing Finder (version 3.1, using Ensembl database release 75 of February 2014). We accessed the online application (http://www.umd.be/HSF/) on March 22nd 2018.The 'automatically select the longest transcript' option was selected, with the 'search for SNPs related to the analyzed sequence' option unselected. Under 'more database options', the 'Display all' and 'all' options were selected (as per default) for ESE/ESS, gene status and transcript status. For the number of nucleotides surrounding the exon (both 5' and 3' introns), a value of 100 nucleotides was

chosen, with no specifications for 5' intron or 3' introns (default settings). For the analysis type, we selected 'splice site analysis' and chose to identify sequence by Ensembl Gene ID. The *GALNT12* gene ID was used (ENSG00000119514). From the dropdown menu, 'Exon 4 (186bp) was selected for *GALNT12* with 100 nucleotides surrounding the exon. A test nucleotide sequence of 386 nucleotides; corresponding to exon 4 of *GALNT12* with 200 intronic nucleotides on 5' and 3' ends, was automatically generated. The position of the intronic variant was nucleotide 308, with a substitution to 'T' selected. Under these conditions, Human Splicing Finder predicts two possibilities: alteration of an intronic splice silencer site or, alternatively, no impact on splicing.

SpliceView was then used to assess the intronic variant. The online web application (http://bioinfo.itb.cnr.it/~webgene/wwwspliceview.html) was accessed on March 22nd 2018. 'Homo sapiens' was selected as the organism, with the direct strand option selected (default). The input sequence for the *GALNT12* variant used was:

ttattggaaggaagacaagaattcacccttgaatttcccaattgtcttcctgctgcccgtctgcataggagagacggatgga
tgtcttgggt(g/t)ctttcagGATCCATGAAGAGGAGTCGGCAGTGGTGTGCCCGGTGATTGATG
TGATCGACTGGAACACCTTCGAATACCTGGGGAACTCCGGGGAGCCCCAGATCGGCGGTT
TCGACTGGAGGCTGGTGTTCACGTGGCACACAGTTCCTGAGAGGGAGAGGATACGGATGC
AATCCCCCGTCGATGTCATCAG. In this sequence, the variant is the 93rd nucleotide. No donor splice sites were reported under these conditions. Two splice acceptor

sites were predicted for position 69 (score 82) and position 102 (score 83). Thus, the variant was not predicted to influence splicing using SpliceView.

In Silico Variant Prediction

SIFT (v1.03) predictions we assessed using the online tool (http://sift.jcvi.org/) on March 22nd 2018. The 'SIFT Human SNPs' option was selected, with Homo Sapiens GRC37 Ensembl 63 version. Genomic coordinates (GRCh37.p13, annotation release 105) were entered into the tool for prediction. Polyphen-2 (v2.2.2r398) predictions (http://genetics.bwh.harvard.edu/pph2/dokuwiki/start) were performed on March 22nd 2018. An individual query was ran for each variant, using the Ensembl protein identifier (ENST00000375011), with amino acid position and substitution inputs. Advanced query options were as per default parameters (hits sorted by identity, no mapping to mismatch, structural parameters calculated for first hit only, contacts calculated for first hit only, minimum alignment length of 100, minimal identity length of 0.5, maximal gap length in alignment of 20, 6 Å threshold for contacts). CADD scores (v1.3) were calculated online on March 22nd 2018 (http://cadd.gs.washington.edu/score). *GALNT12* variants were uploaded to the tool in VCF format according to dbSNP genomic positions using GRCh37.p13 annotation release 105. REVEL scores were obtained from (https://sites.google.com/site/revelgenomics/) on March 22nd 2018. The hg19 GRCh37 human genome build is used for this dataset. The pre-calculated REVEL scores were downloaded from (https://rothsj06.u.hpc.mssm.edu/revel/revel_segments/) and the region encoding

the rare *GALNT12* variants was selected (file name: revel_chrom_09_097062310-115422054.csv.zip).

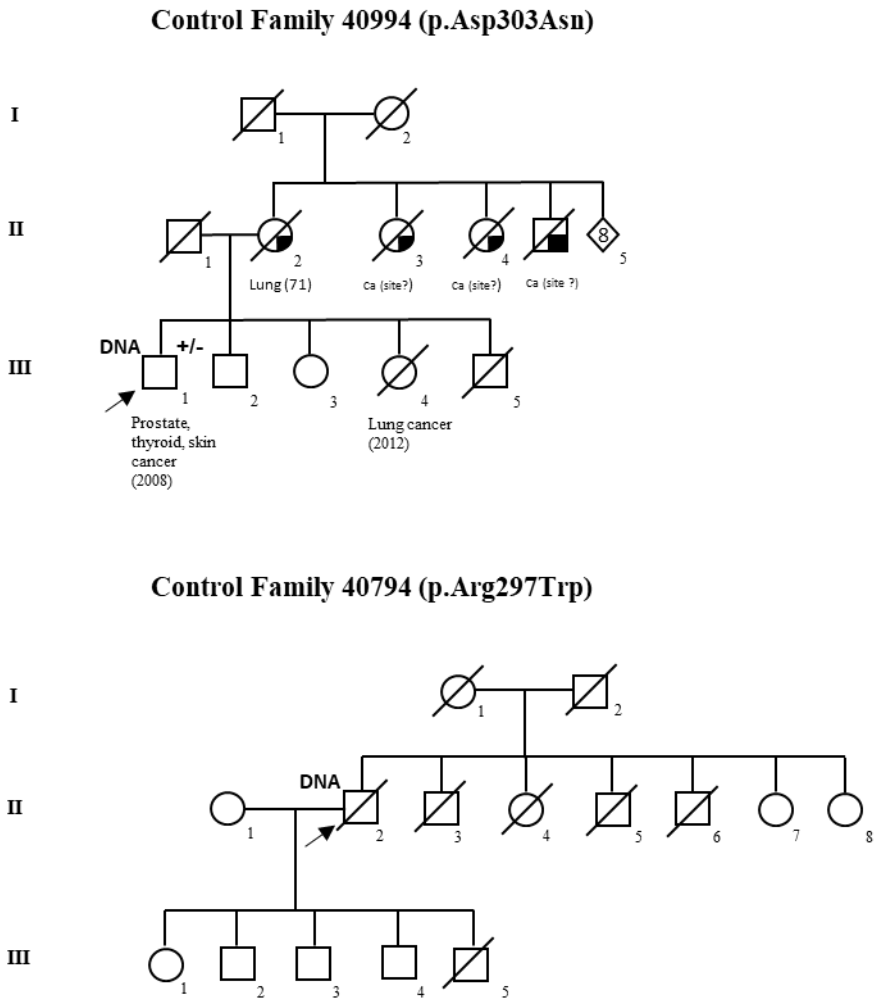
Using SIFT, a variant was considered deleterious if it scored below 0.05. For Polyphen-2, variants scoring 0.15-0.85 were considered 'possibly damaging', and variants above 0.85 were considered 'probably damaging'. CADD scores above 15 were considered deleterious. Finally, we considered a REVEL score above 0.5 indicated a deleterious variant, which corresponds to a sensitivity of 0.754 and specificity of 0.891. We obtained sensitivity and specificity data from the website provided by program authors (https://sites.google.com/site/revelgenomics/). Scores were rounded up to the nearest 0.005 to obtain these sensitivity and specificity values.

Variant Penetrance and Maximum Credible Allele Frequencies

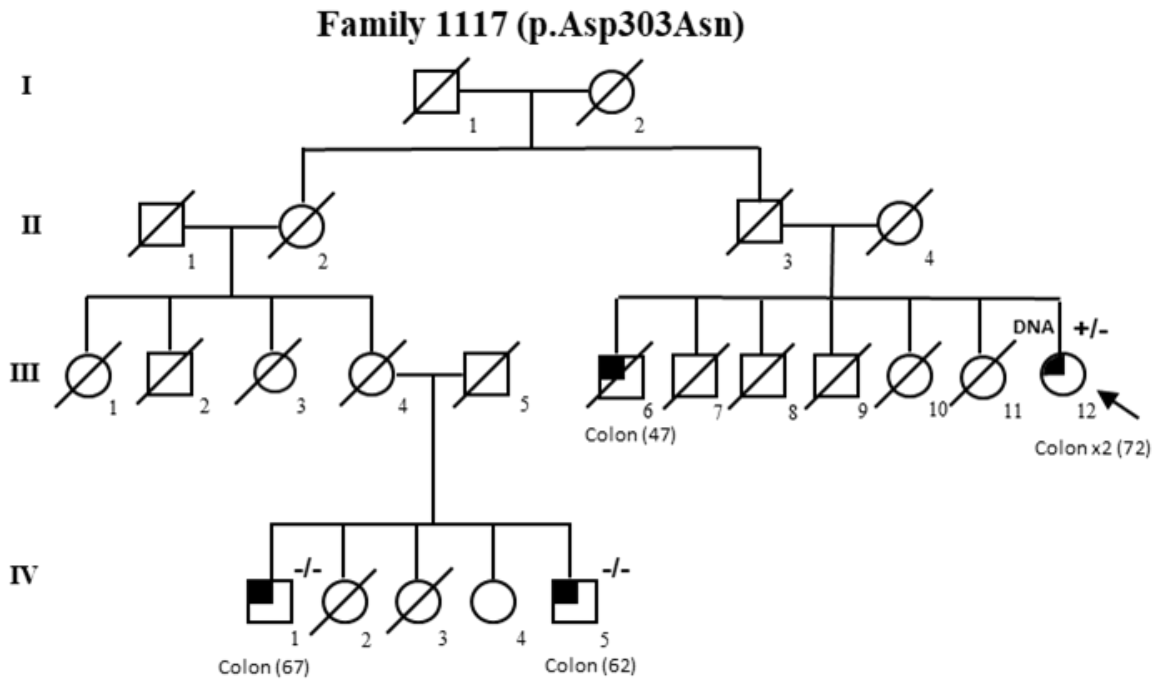
We estimated penetrance on March 22nd 2018 using the online application provided (http://cardiodb.org/allelefrequencyapp/). The incidence rate of CRC in NL is 112.2 cases (per 100,000), based on Canadian Cancer Statistics (2013). The population size of NL was 526,702 in 2013. To calculate total prevalence of CRC in NL we multiplied incidence by average CRC survival rates in NL (6 years). Dividing the total population by this value gave an estimated prevalence for CRC in NL of 1 in 148.5 people. Given that CRC is a late onset disease, we also used estimated lifetime CRC risk to estimate penetrance for comparison. Canadian lifetime risk for CRC was used as no data for NL lifetime risk was available. Under these assumptions, a total prevalence of 1 in 148.5 (estimated NL prevalance) or 1 in 13 (Canadian lifetime risk

for CRC) was the input value to estimate variant penetrance using the online application. The number of *GALNT12* case alleles (from NFCCR screening) was 1 for all variants except p.Asp303Asn (with 3 alleles detected). The total number of case alleles was 958. *GALNT12* population control allele frequencies were obtained from the gnomAD browser (gnomAD r2.0.2) (http://gnomad.broadinstitute.org/).

To calculate maximum credible allele frequency, we accessed the online application (http://cardiodb.org/allelefrequencyapp/) on March 22nd 2018. For input values, we used our estimated NL CRC prevalence (1 in 148.5) or the Canadian CRC lifetime risk (1 in 13). A previous population-based study (Woods *et al.,* 2010) identified MSH2 as the most commonly mutated gene, accounting for 2% of cases incident cases sequenced (genetic heterogeneity). The *MSH2* p.Val265_Gln314del accounted for 71% of cases with *MSH2* mutations (10 out of 14 *MSH2* mutations). Accordingly, genetic 'heterogeneity' was set to 0.2 and allelic heterogeneity was set to 0.71. We used 0.95 for confidence intervals, and a reference population size of 277264 (total possible coverage in GnomAD). We tested maximum credible allele frequencies under 0.1, 0.5 and 1.0 for prevalence (Supp Table 6).



Supp Figure 4 Probands from control families 40994 and 40794 are positive for GALNT12 p.Asp303Asn and p.Arg297Trp variants, respectively. CRC-free control families are negative for personal and family history of CRC, but can include other types. Reprinted from Evans *et al.,* (2018) with copyright permission.



Supp Figure 5 Segregation testing for the p.Asp303Asn allele in Family 1117. Reprinted from Evans *et al.,* (2018) with copyright permission. Additional DNAs were only available for family members on an extended branch of this family. The p.Asp303Asn allele was not present in extended relatives of the proband (first cousins once-removed). RefSeqs: NM_024642.4; NP_078918.3; NC_000009.12.

Supp Table 5 All *GALNT12* variants identified in a cohort of colorectal cancer patients from the Newfoundland Colorectal Cancer Registry. Reprinted from Evans *et al.,* (2018) with copyright permission.

GALNT12 variant	Variant Type	dbSNP rsID	Patients of NFCCR Cohort	dbSNP 1000 genomes (MAF/AC)	NHLBI ESP6500 (MAF/AC)	ExAc (MAF/AC/total alleles)	GnomAD (MAF/AC/total alleles)	SIFT (prediction/median information content)	Polyphen-2 HumDiv (prediction/sensitivity /specificity)	Polyphen-2 HumVar (prediction/sensitivity/ specificity)
c.136G>A (p.Gly46Arg)	SNV	rs10987768	17/479	0.0781/391	NA	0.0/1/2†	0.03608/923/25584	0.43 (T/2.59)	0.002 (B/0.99/0.30)	0.001 (B/0.99/0.09)
c.303C>G (p.His101Gln)	SNV	rs201926457	1/479	NA	0.000315/4	0.0005469/8/14628†	0.0001516/27/178098	0.01 (D/2.37)	0.995 (PRD,0.99/0.30)	0.979 (PRD/0.57/0.94)
c.356A>T (p.Glu119Val)	SNV	rs1137654	68/479	0.0286/143	0.057856/732	0.1102/3334/30262†	0.06002/12993/216482	0 (D/2.37)	0.791 (POD,0.68/0.97)	0.198 (B/0.88/0.73)
g.98823249T>G (c.372-7T>A)	intronic/splice variant	rs2295923	1/479	0.0156/78	0.000461/6	0.006583/799/121366	0.006497/1801/277198	NA	NA	NA
c.425T>C (p.Ile142Thr)	SNV	rs757214097	1/479	NA	NA	0.000008237/1/121404	NA	0 (D/2.37)	1 (PRD,0.85/0.93)	0.998 (PRD/0.18/0.98)
c.715G>C (p.Glu239Gln)	SNV	rs777144221	1/479	NA	NA	0.0000727/2/27512†	0.00008728/17/194766	0.05 (D/2.37)	1 (PRD/0/1)	0.99 (PRD/0.51/0.95)
g.98831764G>T (c.732-8 G>T)	intronic/splice variant	rs763682300	1/479	NA	NA	0.000008252/1/121182	0.000004061/1/246220	NA	NA	NA
c.781G>A (p.Asp261Asn)	SNV	rs41306504	3/479	0.0068/34	0.008073/105	0.01142/1385/121322	0.01188/3294/277212	0.18 (T/2.37)	1 (PRD/0/1)	0.968 (PRD/0.61/0.93)
c.857C>T (p.Thr286Met)	SNV	rs548915885	1/479	0.0004/2	NA	0.0000416/5/120196	0.00002844/7/246124	0.01 (D/2.37)	1 (PRD/0/1)	0.999 (PRD/0.09/0.99)
c.868G>T (p.Val290Phe)	SNV	rs371949942	1/479	NA	0.000154/2	0.0001003/12/119656	0.00009383/26/277090	0 (D/2.37)	0.991 (PRD/0/1)	0.94 (PRD/0.66/0.91)
c.889C>T (p.Arg297Trp)	SNV	rs149726976	1/479	0.0004/2	0.000308/4	0.0002725/32/117444	0.0002744/76/276924	0 (D/2.37)	0.999 (PRD/0.71/0.97)	0.908 (POD/0.69/0.9)
c.907G>A (p.Asp303Asn)	SNV	rs145236923	3/479	0.0008/4	0.000923/12	0.001233/138/111936	0.001244/344/276526	0.47 (T/2.37)	1 (PRD/0/1)	0.999 (PRD/0.09/0.99)
g.98835391G>T	intronic variant	rs377266726	1/441	NA	0.000077/1	0.00006592/8/121356	0.00007942/22/277012	NA	NA	NA
c.1131 G>A (p.Leu377=)	nSNV	rs148125332	1/455	NA	0.000077/1	0.0003047/37/121412	0.0003932/109/277232	NA	NA	NA
g.98840194G>T	intronic variant	rs3824516	34/473	0.0543/272	NA	NA	0.04641/1436/30942	NA	NA	NA
c.1392C>G (p.Pro464=)	nSNV	rs35616709	4/473	0.0018/9	0.004152/54	0.00299/363/121386	0.003071/851/277152	NA	NA	NA
g.98844267G>T	intronic variant	rs1885608	381/473	0.2244/1124	NA	NA	0.7413/22920/30918	NA	NA	NA
c.1497C>T (p.Asn499=)	nSNV	rs35632007	1/435	0.0010/5	NA	0.001822/221/121278	0.001786/495/277166	NA	NA	NA
c.1563C>T (p.Cys521=)	nSNV	rs143860974	1/435	NA	0.000077/1	0.00005766/7/121408	0.00006856/19/277118	NA	NA	NA
c.1707G>C (p.Ser569=)	nSNV	rs2273846	42/470	0.1416/709	0.083192/1082	0.08826/10713/121384	0.08736/24211/277142	NA	NA	NA
g.98849263A>G	3' UTR	rs2273847	39/470	0.1424/713	NA	NA	0.09733/3014/30968	NA	NA	NA

Legend

The RefSeq (NM_024642.4) represents *GALNT12* (ENSG00000119514, ENST00000375011.3) nucleotide coding annotation, while NP_078918.3 represents protein annotation for coding variants. Genomic coordinates (NC_000009.12) describe intronic and untranslated region (UTR) variants. Abbreviations: SNV (single nucleotide variant); nSNV (nonsynonymous single nucleotide variant); NFCCR (Newfoundland Colorectal Cancer Registry); NL (Newfoundland & Labrador); SIFT (sorting tolerant from intolerant); T (tolerated); D (damaging); B (benign); POD (possibly damaging); PRD (probably damaging); NA (not applicable); MAF (minor allele frequency); AC (allele count); dbSNP (the single nucleotide polymorphism database) (available at https://www.ncbi.nlm.nih.gov/SNP/) (Assembly GRCh38.p7); rsID (reference SNP cluster ID); NHLBI ESP (National Heart, Lung, Blood Institute Exome Sequencing Project) (available at http://evs.gs.washington.edu/EVS/); ExAC (Exome Aggregation Consortium) (available at http://exac.broadinstitute.org/) (Version 0.3.1); GnomAD (Genome Aggregation Database) (available at http://gnomad.broadinstitute.org/) (version r2.0.2). Symbols: † (potential low quality site in ExAC as coverage fewer than 80% of individuals). Note: allele counts represent total alleles detected among all sub-populations within databases.

Supp Table 6. Estimates of maximum credible allele frequencies for a putative CRC variant of variable penetrance using NL CRC prevalence or Canadian lifetime CRC risk. Reprinted from Evans *et al.,* (2018) with copyright permission.

Program Parameters	Maximum Credible Allele Frequency/Allele Count (Penetrance 0.1)	Maximum Credible Allele Frequency/Allele Count (Penetrance 0.5)	Maximum Credible Allele Frequency/Allele Count (Penetrance 1.0)
NL Prevalence (1 in 148.5)	0.000478/152	0.0000956/35	0.0000478/20
Genetic Heterogeneity (0.02)			
Allelic Heterogeneity (0.71)			
Total Possible GnomAD alleles (277264)			
Canadian Lifetime Risk (1 in 13)	0.00546/1579	0.00109/332	0.000546/172
Genetic Heterogeneity (0.02)			
Allelic Heterogeneity (0.71)			
Total Possible GnomAD alleles (277264)			

Legend.

Maximal credible allele frequency estimates as per Whiffin et al. (2017) (available online at http://cardiodb.org/allelefrequencyapp/).

Abbreviations: NL (Newfoundland & Labrador); CRC (colorectal cancer); GnomAD (Genome Aggregation Database). Further information (genetic and allelic heterogeneity estimation) available in Supp. Methods.

Supp Table 7 Estimated penetrance of *GALNT12* variants using GnomAD database. Reprinted from Evans *et al.*, (2018) with copyright permission.

GALNT12 Variant	NFCCR AC	NFCCR total alleles	Population database AC	Population database total alleles	Penetrance under NL CRC prevalence (95% CI)	Penetrance under Canadian lifetime CRC risk (95% CI)	Reduction of GALNT12 transferase activity
c.303C>G (p.His101Gln)	1	958	27	178098	0.046 (0.0056-0.38)	0.53 (0.064-1)	S
c.425T>C (p.Ile142Thr)	1	958	1	121404	0.85 (0.027-1)	1 (0.3-1)	S
c.715G>C (p.Glu239Gln)	1	958	17	194766	0.081 (0.0089-0.73)	0.92 (0.1-1)	NS
g.98831764G>T (c.732-8 G>T)	1	958	1	246220	1 (0.054-1)	1 (0.62-1)	S
c.857C>T (p.Thr286Met)	1	958	7	246124	0.25 (0.021-1)	1 (0.24-1)	S
c.868G>T (p.Val290Phe)	1	958	26	277090	0.075 (0.009-0.62)	0.86 (0.1-1)	S
c.889C>T (p.Arg297Trp)	1	958	76	276924	0.026 (0.0036-0.18)	0.29 (0.041-1)	NS
c.907G>A (p.Asp303Asn)	3	958	344	276526	0.017 (0.0052-0.055)	0.19 (0.059-0.63)	NS

Legend

Penetrance estimates as per Whiffin et al (2017) (available online at http://cardiodb.org/allelefrequencyapp/), and modelled under different the projected NL CRC prevalence (1 in 148.5) as well as the Canadian lifetime CRC risk (1 in 13). Population database allele counts and total alleles represent aggregate allele frequencies from the gnomAD browser (http://gnomad.broadinstitute.org/), with exception to allele frequencies for the p.Ile142Thr variant (not covered in GnomAD), for which ExAC (http://exac.broadinstitute.org/) was used. Abbreviations: NFCCR (Newfoundland Colorectal Cancer Registry); NL (Newfoundland & Labrador); CI (Confidence Intervals); S (significant P<0.05); N (not significant, P >0.05); AC (allele count); MAF (minor allele frequency); CRC (colorectal cancer); GnomAD (Genome Aggregation Database); ExAC (Exome Aggregation Consortium); RefSeqs: NM_024642.4; NP_078918.3; NC_000009.12. Further information is available in Supplementary Methods.