WHOLE EXOME SEQUENCING ANALYSIS OF INTRACRANIAL ANEURYSM IN MULTIPLEX FAMILIES FROM NEWFOUNDLAND AND LABRADOR

by

© Amy E. Powell

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science

Discipline of Genetics, Faculty of Medicine

Memorial University of Newfoundland

May 2016

St. John's Newfoundland and Labrador

Abstract

Intracranial aneurysm (IA) is a vascular condition characterized as a saccular dilatation of the cerebral artery wall. The purpose of this study was to identify genetic variants that cause susceptibility to IA in two multiplex families from Newfoundland and Labrador. Whole exome sequencing was completed for 12 affected individuals from families R1352 and R1256. A filtering strategy was then implemented to identify and prioritize rare variants that were shared by multiple affected family members. In family R1352, two variants were identified as top candidates: *C4orf6* c.1A>G, and *GIGYF2* c.3494A>G. Both were present in 6/7 exomes from the family, and passed all filtering steps. In family R1256, *SPDYE4* c.103C>T was identified as a variant of interest, as it segregated in 10/11 affected individuals. Though each variant exhibited incomplete segregation, all three were absent from 100 local population controls. The absence of a definitive candidate variant in the exome suggests that further study is necessary to gain better understanding of the genetic etiology of this disease.

Acknowledgements

First, I would like to thank my supervisor, Dr. Michael Woods, for his guidance, knowledge and mentorship throughout my research project. I would also like to thank my supervisory committee members, Drs. Bridget Fernandez and Sevtap Savas, for their support and assistance throughout the past two years.

Thank you to Dr. Fernandez, Barbara Noble, and the entire team involved in patient recruitment and the collection of clinical information for this project. Thank you to the study participants and their families, without whom this research would not be possible. As well, thank you to our funding agencies: the Heart and Stroke Foundation of Canada, and the Medical Research Fund from Memorial University, for providing financial support.

I would also like to thank the former and current members of the Woods lab, Robyn Byrne and Daniel Evans, for their valued input and help with learning laboratory techniques. Thank you to the students, staff and faculty of the Discipline of Genetics, especially Deborah Quinlan, for their support and advice. Finally, thank you to my parents and family, for their continued encouragement throughout my academic career.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A: Adenine

AAA: Abdominal Aortic Aneurysm

ACA: Anterior Cerebral Artery

ACOMM: Anterior Communicating Artery

AD: Autosomal Dominant

ADPKD: Autosomal Dominant Polycystic Kidney Disease

AP-1: Transcription Activator Protein 1

AR: Autosomal Recessive

BP: Base Pair

BWA: Burrows-Wheeler Aligner

C: Cytosine

Chr: Chromosome

COPD: Chronic Obstructive Pulmonary Disease

CT: Computed Tomography

CTA: Computed Tomography Angiography

DAVID: Database for Annotation, Visualization and Integrated Discovery

dbNSFP: Database of Non-Synonymous Functional Predictions

dbSNP: Single Nucleotide Polymorphism Database

DNA: Deoxyribonucleic Acid

dNTPs: Deoxynucleotide Triphosphates

ECM: Extracellular Matrix

EDN1: Endothelin 1

ExAC: Exome Aggregation Consortium

EXO: Exonuclease

FIA: Familial Intracranial Aneurysm

G: Guanine

GATK: Genome Analysis Toolkit

GERD: Gastroesophageal Reflux Disease

GERP: Genomic Evolutionary Rating Profile

GO: Gene Ontology

GWAS: Genome-Wide Association Study

IA: Intracranial Aneurysm

ICA: Internal Carotid Artery

IEL: Internal Elastic Lamina

IL1β: Interleukin 1 Beta

INDEL: Insertion/Deletion

JDP2: Jun Dimerization Protein 2

L: Left

LOD: Logarithm of Odds

MAF: Minor Allele Frequency

MCA: Middle Cerebral Artery

MCP-1: Monocyte Chemoattractant Protein 1

MMP: Matrix Metalloproteinase

MRA: Magnetic Resonance Angiography

MRI: Magnetic Resonance Imaging

mRNA: Messenger Ribonucleic Acid

MUGQIC: McGill University and Genomic Quebec Innovation Centre

NFCCR: Newfoundland Colorectal Cancer Registry

NGS: Next Generation Sequencing

NHLBI: National Heart, Lung and Blood Institute

NL: Newfoundland and Labrador

NPL: Non-Parametric Linkage

OMIM: Online Mendelian Inheritance in Man

PCA: Posterior Cerebral Artery

PCOMM: Posterior Communicating Artery

PCR: Polymerase Chain Reaction

PICA: Posterior Inferior Cerebellar Artery

PPA: Posterior Probability of Association

R: Right

RNA: Ribonucleic Acid

SAH: Subarachnoid Hemorrhage

SAP: Shrimp Alkaline Phosphatase

SIFT: Sorting Intolerant From Tolerant

SM: Smooth Muscle

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variation

T: Thymine

TAA: Thoracic Aortic Aneurysm

TAAD: Thoracic Aortic Aneurysm and Dissection

TGF-β: Transforming Growth Factor Beta

TNF-α: Tumor Necrosis Factor Alpha

UTR: Untranslated Region

VCF: Variant Call Format

VCT: Variant Comparison Tool

VSMC: Vascular Smooth Muscle Cell

WES: Whole Exome Sequencing

List of Appendices

1. Introduction

1.1 Intracranial Aneurysm

1.1.1 Aneurysm, Subarachnoid Hemorrhage, and Stroke

An intracranial aneurysm (IA) is a dilatation or balloon-like growth of the wall of a cerebral artery. These arterial lesions can vary in shape and size, and have the potential to expand over time. Saccular or berry-shaped IA is the most common form and will be discussed throughout this thesis (Tromp, Weinsheimer, Ronkainen, & Kuivaniemi, 2014). Saccular aneurysms often occur at bifurcations of arteries at the base of the brain, and as a result, IA development is concentrated in the area of the cerebral vasculature known as the Circle of Willis (Figure 1) where junctions are prevalent (Williams & Brown, 2013). IA location varies across affected individuals, though usual target sites include the internal carotid artery (ICA), middle cerebral artery (MCA), anterior and posterior communicating arteries (ACOMM and PCOMM), anterior and posterior cerebral arteries (ACA and PCA), posterior inferior cerebellar artery (PICA), ophthalmic artery, vertebral artery, and the tip of the basilar artery (Brown & Broderick, 2014).

IAs can be difficult to diagnose, as their formation can be asymptomatic. However, the weakened vessel walls of IAs have a high potential to rupture, resulting in bleeding in the subarachnoid space (Frosen et al., 2012). Uncontrolled bleeding in this space between the brain and skull can cause a type of hemorrhagic stroke called subarachnoid hemorrhage (SAH). The main symptom of SAH is a sudden severe headache that is often followed by nausea, vomiting, and neck stiffness, and can rapidly

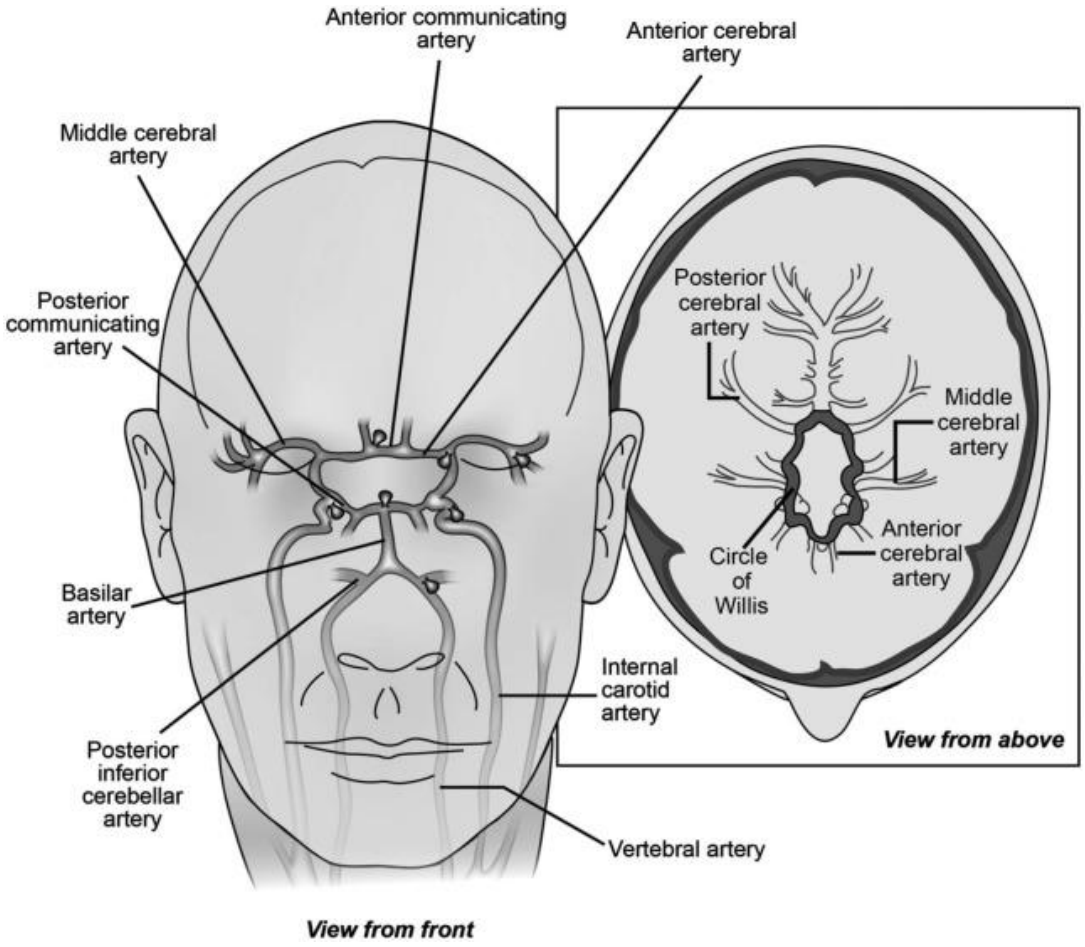


Figure 1. Common sites for saccular intracranial aneurysms. Reprinted from Williams & Brown (2013) with copyright permission.

The Circle of Willis is a known site of aneurysm development. Several major arteries interact and communicate at this juncture, to ensure proper blood supply throughout the brain.

progress to widespread neurological and physical effects (de Oliveira Manoel et al., 2015).

Hemorrhagic stroke, which encompasses both intracerebral hemorrhage and SAH, contributes significantly more to global stroke burden than ischemic strokes caused by obstructed cerebral vessels (Krishnamurthi et al., 2014). Of the 5.3 million cases of hemorrhagic stroke that occurred worldwide in 2010, 3.0 million resulted in death. The majority of stroke cases occur in low to middle income countries, which may be attributed to differences in diagnostic capabilities and access to treatment (Krishnamurthi et al., 2014). However, stroke is still the third leading cause of death in North America, and is responsible for 14,000 deaths in Canada every year (Mukherjee & Patil, 2011; Statistics Canada, 2011). IA rupture is responsible for about 85% of SAH incidences, and SAH itself has a high mortality rate of approximately 50% (Starke, Chalouhi, Ding, & Hasan, 2015). Due to the devastating nature of IA and the impact of stroke on the global population, the study of this condition is of great importance. The identification of risk factors is a step toward reducing the burden of IA and its complications.

1.1.2 Risk Factors for IA Development and Rupture

Several factors play a role in the formation of IA, and consequently, the severity and outcome after diagnosis. Unruptured IA has a prevalence of 3.2% in the general population, as estimated by a meta-analysis of 68 studies in 83 study populations, with an equal proportion of men and women, a mean age of 50, and an absence of comorbidity (Vlak, Algra, Brandenburg, & Rinkel, 2011). However, individuals between the ages of 40 and 60 have the highest risk for IA development (International Study of Unruptured

Intracranial Aneurysms Investigators, 1998). IA is very rare in children, as these lesions are typically acquired later in life (Frosen, 2014). Females are also affected more often than males. Vlak et al. (2011) determined that the prevalence of unruptured IA in women was 6%. In their analysis, the female/male prevalence ratio increased from 1.61 to 2.2 when they considered only individuals over the age of 50. As a result, some studies have suggested that there could be a physiological determinant that is influencing this sex difference, which increases with age. It has been hypothesized that a decrease in estrogen levels and cerebrovascular estrogen receptor density in post-menopausal women might result in increased IA risk, due to the role of estrogen in inflammatory-related processes (Harrod, Batjer, & Bendok, 2006).

Though some factors are beyond our control, many modifiable risk factors also exist. Smoking, hypertension, and excessive alcohol use are all known to increase the risk of IA development as well as progression to rupture (Hussain, Duffis, Gandhi, & Prestigiacomo, 2013). Therefore, lifestyle changes are an integral part of aneurysm prevention and management. Finally, genetic predisposition has been identified as a risk factor for both IA development and rupture. This is suggested based on the fact that, while IA can occur sporadically, it can also be concentrated in families. Research has shown that the first-degree relatives of IA patients are at a 4 time greater risk of developing the disease (Hussain et al., 2013). Familial aneurysms have an earlier age of onset, on average, and pose a greater rupture risk (Broderick et al., 2009). Over the past 20 years, the genetic contribution to both sporadic and familial IA has been explored using various research techniques. Despite these efforts, the genetic etiology of this disease is not well understood, and further research is necessary to explain IA at the

molecular level. Evidence regarding the genetic etiology of IA will be explored in detail in section 1.2.

1.1.3 Screening and Treatment for IA

If an individual is at high risk for IA, or has a family history of the disease, a physician can suggest screening by magnetic resonance angiography (MRA) or computed tomography angiography (CTA) to establish a diagnosis. Both methods have a high degree of accuracy for aneurysms larger than 3 mm (Williams & Brown, 2013). Unfortunately, approximately 50% of IA cases are diagnosed only after SAH has occurred (Brown & Broderick, 2014). More than 30 years ago, Fox (1982) demonstrated the impact of screening when family history is known, in a short case report. He described a family with multiple cases of IA, with two siblings who would not agree to undergo elective screening for undiagnosed IA. A couple of years after the original study was published, one of these siblings suffered a severe SAH, and died shortly thereafter from complications.

Following confirmation of unruptured IA, or an SAH episode, one of several treatment options may be considered. Surgical clipping and endovascular coiling are the two most common interventions, though each carries significant risk (Figure 2). One of these strategies could be employed to prevent rupture, or to reduce the chance of re-bleeding and the accompanying neurological complications of SAH (de Oliveira Manoel et al., 2015).

Surgical clipping is an invasive procedure that involves a craniotomy, followed by the placement of a metal clip around the aneurysmal neck. This action separates the IA

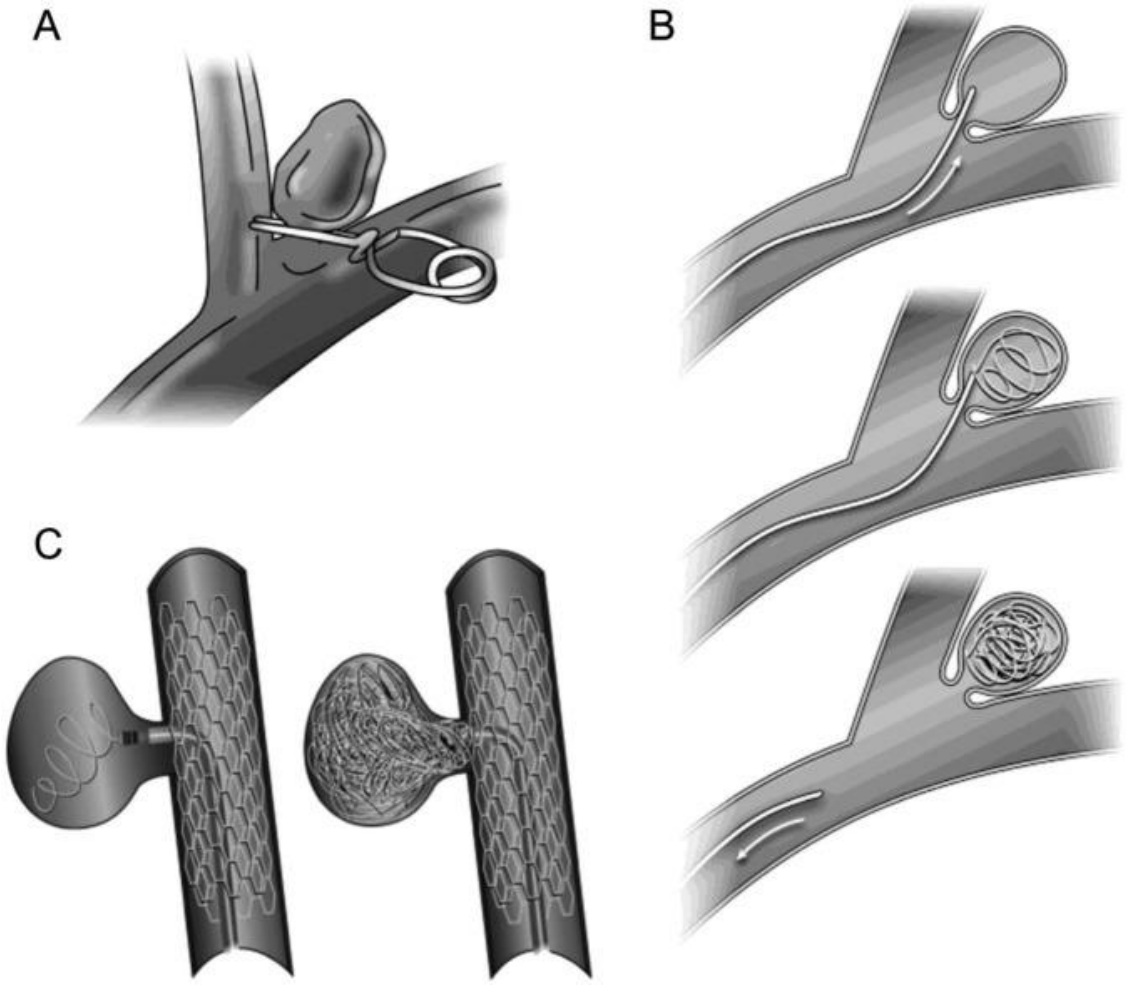


Figure 2. Types of unruptured intracranial aneurysm interventions: (A) aneurysm clipping, (B) endovascular coiling, (C) endovascular coiling with stent assistance. Reprinted from Williams & Brown (2013) with copyright permission.

from the rest of the main artery, to prevent the entry of blood (Williams & Brown, 2013). In contrast, endovascular coiling is a common and relatively new treatment, which is less invasive than clipping. In this method, platinum coils are delivered into the aneurysm to isolate it from the affected artery (Williams & Brown, 2013). A stent may be used in conjunction with this procedure in some difficult cases, such as when the aneurysmal neck is wide. A neurosurgeon must weigh the risks of IA complication with the risks of treatment, to decide the best course of action (Brown & Broderick, 2014). In addition to aneurysm size and location, the age of the patient, family history, and medical history must be considered. For example, an older patient with a small IA and other serious health issues may not be an ideal candidate for intervention. Alternatively, patients may be recommended for "conservative management", where they are screened periodically to track the progression of the IA (Bederson et al., 2000).

1.1.4 Pathophysiology

Several hypotheses have been presented regarding the pathophysiology of IA, but many questions still remain. The exact order of steps in the development and rupture of an IA is much debated, though most research points to a shared set of events. A study by Chalouhi et al. (2012) stated that the common pathway for intracranial aneurysm development consists of endothelial injury, inflammatory response, vascular smooth muscle cell modulation, extracellular matrix remodeling, and apoptosis, leading to the degradation of the artery wall.

Normally, cerebral arteries consist of three layers (Figure 3): the outer adventitia, the media, and the inner intima (Y. Wang et al., 2014). The hemodynamic stress of

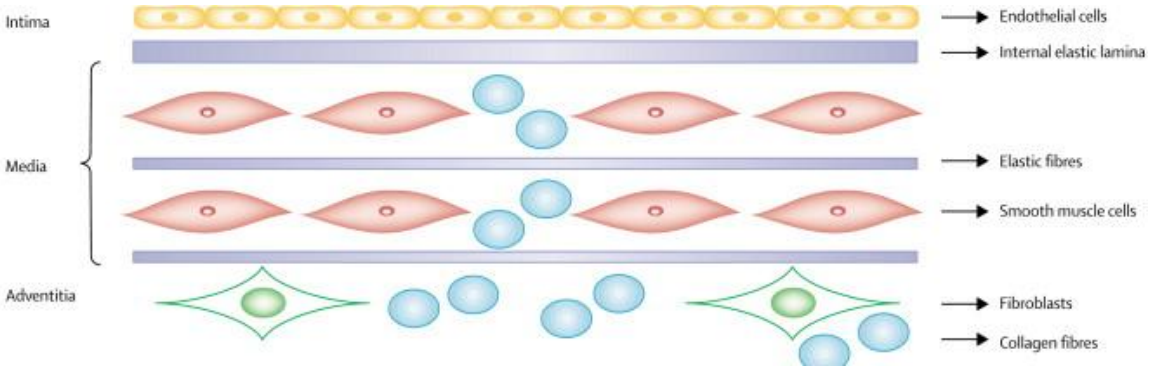


Figure 3. Intracranial arterial wall. Reprinted from Ruigrok, Rinkel & Wijmenga (2005) with copyright permission.

turbulent blood flow at arterial junctures, in conjunction with modifiable and non-modifiable IA risk factors, initiates a cascade of events that alters the structure of these layers. Experimental models of IA have shown that this stress causes the dysfunction of endothelial cells, which become elongated and experience a change in density. In addition to physical changes, signaling in the endothelium is altered (Chalouhi, Hoh, & Hasan, 2013). For example, Aoki et al. (2011) demonstrated that shear stress causes the activation of the prostaglandin E_2/prostaglandin E receptor 2 (PGE_2/EP_2) pathway in endothelial cells, which is involved in the promotion of inflammation. Modified endothelial cells also express monocyte chemoattractant protein-1 (MCP-1), which causes the recruitment of macrophages and additional leukocytes into the cerebral artery (Chalouhi et al., 2013).

Another known feature of IA pathology is the phenotypic modulation of vascular smooth muscle cells (VSMCs). VSMCs are located in the media layer of the artery, and are normally involved in contraction and dilation (Chalouhi et al., 2013). During aneurysm development, VSMCs migrate to the intima, as a response to endothelial injury. These VSMCs respond by synthesizing collagen and thickening the wall, which is referred to as intimal hyperplasia and is a known response to vessel damage and wounds (Frosen, 2014). VSMCs in the intima become less compacted and weaker than their media counterparts, thus undergoing several functional changes to enter a pro-inflammatory differentiated state. Intimal VSMCs stimulate the production of matrix metalloproteinases (MMPs), which are enzymes that degrade collagenous fibers and result in the loss of the extracellular matrix (Chalouhi et al., 2012). Ali et al. (2013) used western blot analysis to demonstrate that MMP-3 and MMP-9 are expressed in rat

cerebral VSMCs that have been treated with tumour necrosis factor alpha (TNF-α), a pro-inflammatory cytokine. TNF-α is part of the immune system's response to stress, and has increased expression in the wall of IAs. As well, Ali et al. (2013) showed that the smooth muscle cell proteins SM-MHC (smooth muscle myosin heavy chain), SM-α-actin, and SM-22-α had reduced expression in rat VSMCs following TNF-α treatment. A decrease in the expression of these contractile proteins would provide evidence for the role of TNF-α in the modulation of VSMCs from a muscle contraction role to a pro-inflammatory role in the body.

As demonstrated above, the role of inflammation is predicted to be integral to IA pathogenesis. Both VSMC and endothelium modification is involved in the production of an inflammatory response and the loss of essential layers of the arterial wall. Like intimal VSMCs, macrophages that have been recruited to the site of vascular injury have been shown to produce MMPs, cytokines and other inflammatory agents (Figure 4). For example, the cytokine interleukin-1 beta (IL1β) has been detected following induced IA formation in mice (Moriwaki et al., 2006). Specifically, IL1β activity was involved in the promotion of VSMC apoptosis in this model. In addition to the death of VSMCs and consequently the degradation of the ECM, inflammatory mediators also disrupt the internal elastic lamina. The internal elastic lamina (IEL) is located between the intima and media layers, and helps to maintain elasticity in the artery and inhibit over-expansion of the wall (Frosen, 2014). However, IAs have a characteristic lack of IEL, which increases the workload of collagen fibers. Strain on the remaining collagen fibers in conjunction with VSMC apoptosis results in a loss of collagen and overall thinning of the

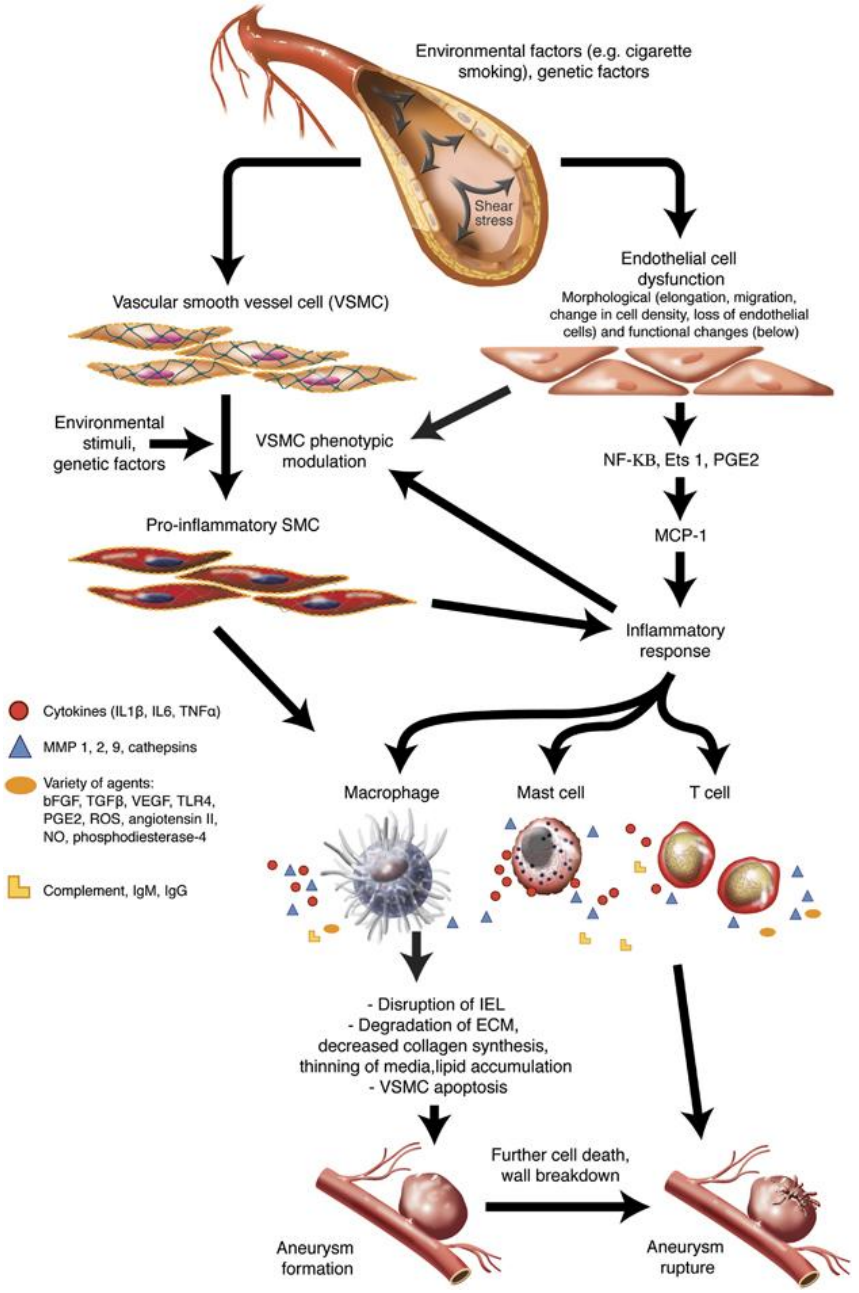


Figure 4. Inflammatory reaction associated with intracranial aneurysm formation and rupture. Reprinted from Chalouhi et al. (2012) with copyright permission.

This hypothesized pathway shows that the action of various inflammatory agents triggers the progression from weakened artery wall to aneurysm formation. The highlighted processes are influenced by environmental and genetic risk factors, demonstrating the multifactorial nature of IA.

media layer (Chalouhi et al., 2012). Overall, it is clear that a stimulated loss of structural capacity of the artery wall is key to the development of IA.

Following initial IA formation, these lesions often continue to grow or expand with the stress of blood flow and distension of the artery wall (Frosen, 2014). Frosen et al. (2012) succinctly stated that aneurysm rupture occurs "when blood pressure-induced tension of the wall exceeds it strength". An ongoing active inflammatory response may also contribute to the eventual collapse of an IA (Chalouhi et al., 2013). Starke et al.(2015) recently suggested that aspirin and other anti-inflammatory medications should be explored as treatments to prevent aneurysm rupture. Better understanding of the exact biological processes that lead to IA formation, expansion and progression to rupture could lead to improved treatment, management, and even prevention. This understanding may come from the identification of genes, and consequently proteins, that are underlying the pathogenesis of this disease.

1.2 The Genetic Contribution to IA

1.2.1 Preliminary Evidence

It is widely acknowledged that there is a genetic component to the development of IA. One significant piece of evidence is the association of IA development with heritable connective tissue disorders that have a vascular component, including Marfan syndrome, type IV Ehlers-Danlos syndrome, and autosomal dominant polycystic kidney disease (ADPKD) (Schievink, Michels, & Piepgras, 1994). The genes involved in the inheritance of these disorders are well described (Table 1). Most notably, ADPKD is caused by mutations in *PKD1* and *PKD2*, which encode the extracellular matrix proteins

Table 1. Heritable connective tissue disorders that have been associated with IA, and related genetic risk factors.

Syndrome	Causative Gene(s)	Corresponding Proteins
Autosomal dominant polycystic kidney disease	*PKD1*, *PKD2*	Polycystin-1, Polycystin-2
Marfan syndrome	*FBN1*	Fibrillin-1
Neurofibromatosis type I	*NF1*	Neurofibromin
Ehlers-Danlos syndrome type IV	*COL3A1*	Collagen alpha 1 (III) chain
Pseudoxanthoma elasticum	*ABCC6*	Multidrug resistance-associated protein 6
Supravalvular aortic stenosis	*ELN*	Elastin
Multiple endocrine neoplasia type I	*MEN1*	Menin
Hereditary hemorrhagic telangiectasia	*SMAD4*, *ENG*, *ACVRL1*	Mothers against decapentaplegic homolog 4, Endoglin, Serine/threonine-protein kinase receptor R3

polycystin-1 and polycystin-2 respectively, and approximately 10% of ADPKD patients will develop IA (Chapman et al., 1992; Zhang & Claterbuck, 2008). IA is known as the symptom of ADPKD that contributes the most to disease-related mortality (Rossetti & Harris, 2013). Other genetic syndromes have also been associated with an elevated risk of cerebral aneurysm, including Moyamoya disease, sickle-cell disease, and fibromuscular dysplasia (Brown & Broderick, 2014).

Additionally, IA can occur in patients afflicted by other types of aneurysms. Aneurysm growth is not limited to the cerebral vasculature, and lesions on the thoracic and abdominal regions of the aorta can occur as isolated or familial incidences. Regalado et al. (2011) described multiple families with autosomal dominant inheritance of thoracic aortic aneurysms and dissections (TAAD), where some affected relatives also had IA and/or abdominal aortic aneurysms (AAA). Through the study of five of these families, four rare mutations in the *SMAD3* gene were identified. They were able to conclude that *SMAD3* mutations can explain 2% of familial TAAD cases, and more significantly, that there may be a common genetic mechanism that connects all arterial aneurysms (Regalado et al., 2011).

However, to date, no causative genes for non-syndromic IA have been defined. Review of the published literature has revealed an abundance of research aiming to uncover the gene or genes involved in IA development and progression. Functional studies have yet to show a connection between identified genetic variants and the manifestation of IA. Evidence including the co-occurrence of IA with other genetic syndromes and aneurysms, the observation of IA clustering in families, and the increased

risk associated with a family history of the disease have all catalyzed further investigation.

1.2.2 Family-Based Studies

Twin studies are a well-known and highly useful approach for exploring the influence of genetics on disease risk. Leung et al. (2011) completed a review of the literature, and found that only 15 twin pairs with IA had been described before 2011. Seven of these pairs had IAs in the same location in the vasculature. On average, the time between diagnosis of the first twin, and development of IA in the second twin was 2.4 years. Leung et al. (2011) also reported on a newly identified monozygotic twin pair with IA, where the first twin was diagnosed with an aneurysm of the left ICA at the age of 57. The other twin underwent screening, which resulted in the diagnosis of a right ICA aneurysm. Interestingly, these twins demonstrated IA at "mirror sites", which is another feature seen in 3/15 of the pairs previously described. In a recent article by Mackey et al. (2015), additional twin-pairs were identified from the Familial Intracranial Aneurysm Study cohort. These participants are part of a large-scale international study of familial IA (Broderick et al., 2005). Of the 11 monozygotic twin pairs enrolled in their study, 8 demonstrated IA location concordance. The same phenomenon was only observed in 1 of 5 dizygotic twin pairs, providing evidence for a genetic contribution to IA location and incidence (Mackey et al., 2015).

Along with twin-pairs affected by IA, many multiplex families have been described. These studies have provided integral insight into the mode of inheritance of familial IA. In particular, a study conducted by Wills et al. (2003) describes 346 Finnish

families and their predicted inheritance patterns. Each family had at least 2 affected members, while 14 of these 346 families had particularly strong family history, with 6 or more affected members. Wills et al. (2003) reported that 57.2% of the families had a pattern of inheritance that was characteristic of an autosomal recessive mode of inheritance, and 36.4% were consistent with an autosomal dominant mode. They indicated that 5.5% of the pedigrees were consistent with an autosomal dominant incompletely penetrant mode, and the remaining 0.9% of families did not fit any Mendelian pattern of inheritance. It was emphasized that across their cohort there was father-to-son, father-to-daughter, mother-to-son, and mother-to-daughter transmission of IA, providing evidence for autosomal inheritance in this disease (Wills et al., 2003). In addition to showing the degree to which IA aggregates in families, this large-scale study showed that familial IA might follow multiple inheritance patterns.

In order to further explore the genetic contribution to IA in families, many researchers have turned to DNA linkage analysis. In a recent review article, Tromp et al. (2014) highlighted six chromosomal loci that were identified in at least two independent linkage studies of IA, and had significant logarithm of odds (LOD) scores above 2.0. These regions are 1p34-36, 4q32, 7q11, 19q13, and Xp22. A detailed description of the respective publications and study designs is shown in Table 2.

Ruigrok et al. (2008) were able to replicate two of these previously reported loci, 1p36.11-p36.13 and Xp22, in a single consanguineous family from the Netherlands. First, they used a linkage panel of 5,861 single nucleotide polymorphisms (SNPs) to genotype 17 family members. Twenty-four additional microsatellite markers were then genotyped corresponding to regions that were of particular interest. Using the GENEHUNTER

Table 2. Genomic regions described in two or more previous linkage analyses of familial IA.

Locus	Publications	Cohort	Study Design
1p34-36	Nahed et al.(2005)	1 North American family (AD)	Parametric
	Ruigrok et al. (2008)	1 Dutch family	Non-parametric
4q32	Foroud et al. (2008)	192 International families	Non-parametric
	Foroud et al. (2009)	333 International families	Parametric
7q11	Onda et al. (2001)	85 Japanese families	Sibling pairs
	Farnham et al. (2004)	13 North American families (AR)	Parametric
19q13	Olson et al. (2002)	85 Finnish families	Sibling pairs
	van der Voet et al. (2004)	333 Finnish families	Sibling pairs
	Yamada et al. (2004)	29 Japanese families	Non-parametric
	Mineharu et al. (2007)	9 Japanese families (AD)	Parametric
Xp22	Olson et al. (2002)	85 Finnish families	Sibling pairs
	Yamada et al. (2004)	29 Japanese families	Non-parametric
	Ruigrok et al. (2008)	1 Dutch family	Non-parametric

Abbreviations: AD = autosomal dominant, AR = autosomal recessive

software, they selected a nonparametric or model-free approach, as the inheritance pattern for this family was unclear. Only known affected family members were included in this analysis. This software produces statistical estimates in the form of nonparametric linkage (NPL) scores, with scores above 3.18 being suggestive of linkage, and scores above 4.08 indicating significant linkage. For a series of markers at the 1p36 locus, there was a maximum NPL score of 3.18. For Xp22, there was a maximum NPL score of 4.54. It is possible that one or both of these linked regions could be of importance in familial IA. This study is of interest, as these results could also indicate that IA is digenic in this kindred.

Similarly, Verlaan et al. (2006) used linkage analysis to identify a susceptibility locus for IA in a large French-Canadian family. The French-Canadian population from Quebec is known to exhibit founder effects, making haplotype analysis a useful method for investigating IA in this family. A total of 531 microsatellite markers were genotyped in nine affected and three unaffected family members. Based on the disease segregation in this pedigree, a parametric, affecteds-only approach was taken by the research team, assuming an autosomal dominant mode of inheritance. A locus on chromosome 5p had the strongest evidence for linkage, and fine mapping revealed a common disease haplotype between 5p15.2-14.3 that segregated with affected family members. The authors also mention that there are two known individuals in the family who are non-penetrant for IA, and have the disease haplotype. Most of the affected family members are known smokers, whereas these two non-penetrant individuals are non-smokers. Therefore, it is possible that the inheritance of a disease locus alone may not be sufficient to result in IA manifestation in this family. However, as IA is known to be a late-onset

disease, the non-penetrant individuals could still develop IA later in life. Follow-up diagnostic testing of these individuals over time could provide answers regarding the relevance of this shared haplotype.

One of the main challenges following linkage analysis is the identification of a specific gene or variant of interest, given a linked region. This issue of fine-mapping is exasperated by the fact that familial IA may have a high degree of genetic heterogeneity, and genomic regions identified through linkage analysis may be specific to certain families. Verlaan et al. (2006) identified two possible candidate genes in the 5p15.2-14.3 region, *CTNND2* and *TRIO*, both of which are involved in cell modeling. These are only two of 25 known genes in this region, and further in-depth study of these genes would be necessary to connect them to IA susceptibility.

1.2.3 Candidate Gene Studies

Linkage analyses have led to the identification of several candidate genes for IA within linked genomic regions. These genes are generally selected for their functional relevance. For example, the elastin (*ELN*) gene has been identified as a candidate gene, for its proximity to the 7q11 linked region and the contribution of its protein product to the structure and elasticity of vessel walls (Tromp et al., 2014). Candidate gene study designs involve the selection of common polymorphisms found in or near a gene of functional interest. These polymorphisms are typically genotyped in sporadic IA cases to determine statistical association with disease occurrence, compared to genotyped controls (Tromp et al., 2014).

Krischek et al. (2010) completed one such study, by focusing their attention on the *JDP2* gene and its potential role in IA. They had previously identified the 14q22 locus as a target for further exploration, through linkage analysis of affected sibling pairs from Japan (Onda et al., 2001). Initial genotyping of 100 SNPs located up and downstream from 14q22 was completed for 148 Japanese sporadic IA cases and 190 controls, and statistical association was detected near the *JDP2* gene (Krischek et al., 2010). After genotyping more individuals (403 cases and 412 controls), a single SNP in the intronic region of *JDP2* was associated with IA, with an odds ratio of 1.44. Ten SNPs in the *JDP2* gene were then genotyped in Japanese, Korean, and Dutch case-control cohorts. A total of three SNPs were associated with IA in the Japanese cohort, with P-values reaching significance ($P < 0.05$). One of these intronic SNPs was also associated with IA in the Korean patient cohort (Krischek et al., 2010). They explained the relevance of this candidate by describing the jun dimerization 2 (JDP2) protein as a repressor of transcription activator protein 1 (AP-1). The AP-1 protein initiates apoptosis and, in the absence of a repressor, could cause vascular remodeling via cell death (Krischek et al., 2010).

In order to identify strongly associated SNPs, Alg et al. (2013) completed a meta-analysis of IA candidate gene and genome-wide (discussed in section 1.2.4) association studies, enabling them to evaluate the accuracy of previously reported findings. If a SNP was associated with IA in more than study, a fixed-effect statistical model was used to pool odds ratios. They determined that eight SNPs, each found in at least two candidate gene studies, were significantly associated with IA with either a dominant, recessive or additive disease model. These SNPs were found in the genes *SERPINA3* (*rs4934*),

COL1A2 (*rs42524*), *COL3A1* (*rs1800255*), *HSPG1* (*rs3767137*), *CSPG2* (*rs251124* and *rs173686*), *ACE (rs4646994*) and *IL6* (*rs1800796*), all of which have integral functions in vascular biology and inflammation (Alg et al., 2013). Two variants in the *CSPG2,* or versican, gene were associated with IA in diverse cohorts, including Dutch and Japanese individuals with sporadic IA. The versican protein is a component of the extracellular matrix, and previous studies of AAA tissue have revealed decreased mRNA expression of versican, making it a gene of significant interest (Handley, Samiric, & Ilic, 2006; Alg et al., 2013) *COL1A2* and *COL3A1* are also clearly strong candidates, as the collagen proteins are a known structural component of the vascular adventitia (Ruigrok et al., 2005). Unfortunately, to date, the relationship between IA candidate genes and their biological role in disease development has not been established. To uncover more loci that may be associated with IA, many researchers have turned to broader genome-wide investigations.

1.2.4 Genome-Wide Association Studies

Genome-wide association studies (GWAS) have been instrumental to the evolving understanding of IA genetics, by providing a more comprehensive assessment of genomic variation. Several genomic regions have been associated with IA in more than one independent GWAS (Table 3), suggesting their potential roles in IA development, while others have not been replicated. The chromosomal region that has garnered the most attention for its association with IA is the 9p21 locus. Helgadottir et al. (2007) first reported an association between a common SNP at this locus, *rs10757278*, and myocardial infarction in a large sampling of the Icelandic population. This polymorphism

Table 3. Loci associated with IA in two or more GWAS.

Locus	Publications	Study Population	Odds Ratio (OR)	Candidate Genes in Region
2q33.1	Bilguvar et al. (2008)	Japan & Europe	OR=1.24	*BOLL*, *PLCL1*, *ANKRD44*
	Kurki et al. (2014)	Finland, Netherlands	OR=1.27	
4q31.23	Yasuno et al. (2011)	Japan	OR=1.22	*EDNRA*
	Low et al. (2012)	Japan	OR=1.25	
8q11.23-12.1	Bilguvar et al. (2008)	Japan & Europe	OR=1.36	*SOX17*
	Yasuno et al. (2010)	Japan & Europe	OR=1.28	
	Foroud et al. (2012)	Europe & USA	OR=1.25	
9p21.3	Bilguvar et al. (2008)	Japan & Europe	OR=1.29	*CDKN2A*, *CDKN2B*
	Yasuno et al. (2010)	Japan & Europe	OR=1.32	
	Low et al. (2012)	Japan	OR=1.21	
	Foroud et al. (2012)	Europe & USA	OR=1.35	

is adjacent to the *CDKN2A* and *CDKN2B* tumor suppressor genes, which have been classified for their role in apoptosis, proliferation, and other cellular functions (Magrane & Consortium, 2011). To investigate any further link between 9p21 and arterial disease, Helgadottir et al. (2008) genotyped *rs10757278* in individuals affected by AAA, IA, or large artery atherosclerotic/cardiogenic stroke, and corresponding controls of European descent. They determined that *rs10757278-G* is a risk allele for both AAA and IA, in multiple populations. This association has been replicated in the four largest GWAS of IA to date (Bilguvar et al., 2008; Yasuno et al., 2010; Low et al., 2012; Foroud et al., 2012). Each of these groups utilized large sporadic case-control cohorts in both the discovery and replication phases of their research, and were able to achieve results reaching genome-wide significance (Hussain et al., 2013).

Bilguvar et al. (2008) performed the first large GWAS of IA utilizing both a discovery and replication cohort, which included more than 2,100 IA cases and 8,000 control samples. During the discovery phase genotyping of Finnish and Dutch individuals, they found that SNPs at chromosomal regions 1q, 2q, 8q and 9p were associated with IA. Following replication in a Japanese cohort, eight SNPs were significantly associated with IA, at 2q33.1, 8q11.23 and 9p21.3, with P-values below $1x10^{-8}$, the determined cut-off for genome-wide significance (Yasuno et al., 2010). To increase the power of this initial study, Yasuno et al. (2010) used the same Illumina platform to genotype 832,000 SNPs in two additional case cohorts and five control cohorts from Europe. A second Japanese replication cohort was also added. They used Bayesian statistics to calculate the posterior probability of association (PPA) for each SNP. They were able to confirm the associations at 8q11.23-q12.1 and 9p21.3, and

identify SNPs at three new loci that were associated with IA risk with PPA values greater than 0.5 (50%). These loci include 10q24.32, 13q13.1, and 18q11.2 (Figure 5). Given that these loci account for only a small percentage of overall IA risk, Yasuno et al. (2011) predicted that SNPs with a PPA value between 0.1-0.5 in the discovery cohort might also be significant following replication in the Japanese cohort. They genotyped 25 SNPs with a PPA in the 0.1-0.5 range in two Japanese case-control groups. One SNP in particular, *rs6841581*, was significantly associated with IA, with a final PPA of 0.986 following replication. This SNP is located at the 4q31.23 locus, which is close to the *EDNRA* gene. *EDNRA* encodes a G-protein coupled receptor for endothelins, including EDN1. EDN1, or endothelin-1, is involved in vasoconstriction and dilation of blood vessels, including most notably, cerebral arteries (Yasuno et al., 2011). The 4q31.23 locus grew in interest when an association was confirmed by Low et al. (2012) in their GWAS of the Japanese population.

Low et al. (2012) performed a large-scale GWAS using a cohort of cases who had experienced SAH due to aneurysmal rupture. They successfully genotyped 565,149 SNPs in their discovery cohort of 1,383 SAH patients and 5,484 controls, and determined that several SNPs showed suggestive association with SAH. They genotyped 36 of these loci, and 7 that were previously associated with IA in other GWAS, in their replication cohort of 1,048 IA patients and 7,212 controls. One of these SNPs, *rs6842241*, was significantly associated with IA, with a P-value of 9.58×10^{-9}. This SNP is located at 4q31.22, and also falls within the boundaries of the *EDNRA* gene. As stated previously, they also confirmed that *rs10757278-G* is a risk allele for IA development in the Japanese population (Low et al., 2012).

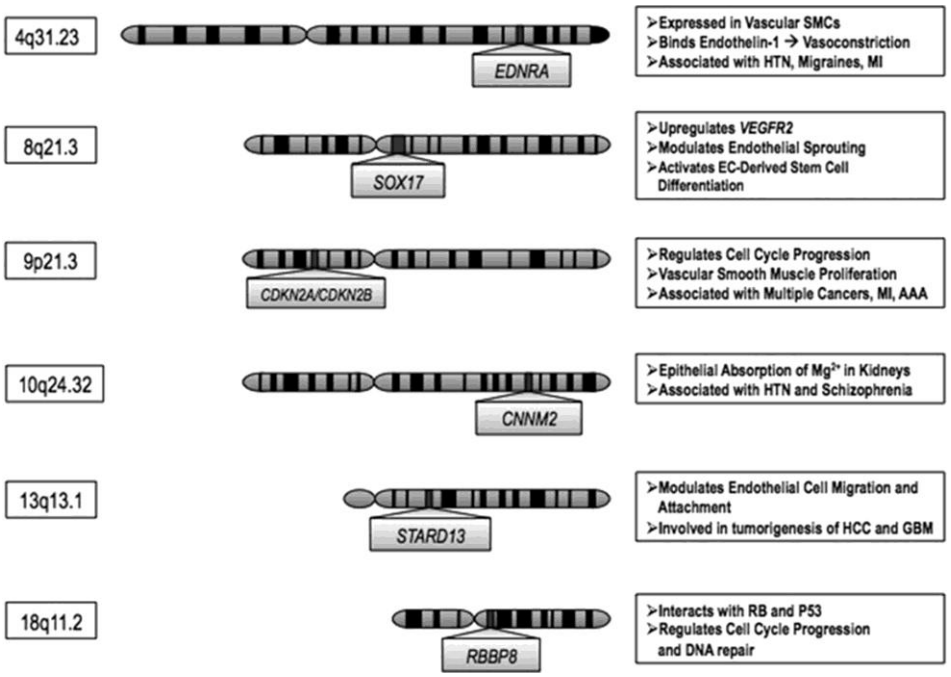


Figure 5. Functional candidate genes located at or near loci that are significantly associated with IA (P-value <1x10^{-8}). Reprinted from Hussain et al. (2013) with copyright permission.

In their comprehensive study of the genetic etiology of IA, the Familial Intracranial Aneurysm (FIA) Study Investigators also completed a large-scale GWAS (Foroud et al., 2012). The GWAS described above involved the recruitment of patients with sporadic IA or SAH, without consideration of familial disease occurrence. One of the limitations of such a study design is that risk factors for sporadic IA may be separate from the factors contributing to familial IA development and rupture. To ameliorate this design, Foroud et al. (2012) created a study with two new discovery cohorts, with participants of European descent. Discovery Sample 1 included 388 cases from the FIA study, all with a family history of the disease. This sample also contained 397 control individuals. Discovery Sample 2 contained a mixture of 1,095 familial and sporadic IA cases, with 27% reporting a family history. This sample also included 1,286 controls. A SNP array was used for genotyping, and imputation was used to produce a common set of SNPs across both samples. Use of a logistic regression model showed that there were no risk variants unique to familial IA, as no associations reached genome-wide significance. This finding was interesting, as it suggests that the distinction between sporadic and familial IA cases might not influence experimental findings in GWAS. Multiple additional studies with sporadic and familial cohorts would be necessary to determine if family history has any impact on which loci are associated with IA.

Meta-analysis of the results was performed across the two discovery samples, which led to the successful identification of a SNP reaching genome-wide significance. Foroud et al. (2012) confirmed a significant association between the SNP *rs6475606* and IA in Caucasian individuals. This SNP is found in the *CDKN2BAS* gene region of 9p21. They also attempted to replicate the association between 8q11.23 and IA, which did not

achieve genome-wide significance. The 8q chromosome has been mentioned in multiple small and large-scale GWAS, and contains the candidate gene *SOX17*. Under-expression of SOX17 protein in a mouse model resulted in mutants with defective endothelial cell sprouting, which is a component of angiogenesis (Matsui et al., 2006). Many of the loci replicated throughout large-scale GWAS contain genes that are connected to vascular biology (Figure 5).

In summary, linkage analyses and association studies have provided some potential functional candidates and common SNPs that are statistically associated with IA risk. A chromosomal map of these IA loci is provided in Figure 6. However, the translation of this knowledge has not yet been successful. No definitive genetic variant has been identified that causes IA. Therefore, the inclusion of new approaches that focus on rare and deleterious genetic variation in the genome should be explored. New innovative approaches to human genetics research could help determine what genetic factors are involved in the etiology of this disease.

1.2.5 Recent Strategies for IA Research

In recent years, several new technologies have arisen for the study of genetic diseases. Foroud & FIA Study Investigators (2013) guided a new approach to IA research by using next generation sequencing to study the Familial Intracranial Aneurysm Study cohort. Specifically, they pursued whole exome sequencing (WES): a method used to capture and sequence the known protein-coding portion of the genome, which contains approximately 85% of the mutations described in Mendelian disorders (Ng et al., 2009; Z. Wang, Liu, Yang, & Gelernter, 2013). Exons and exon/intron boundaries only

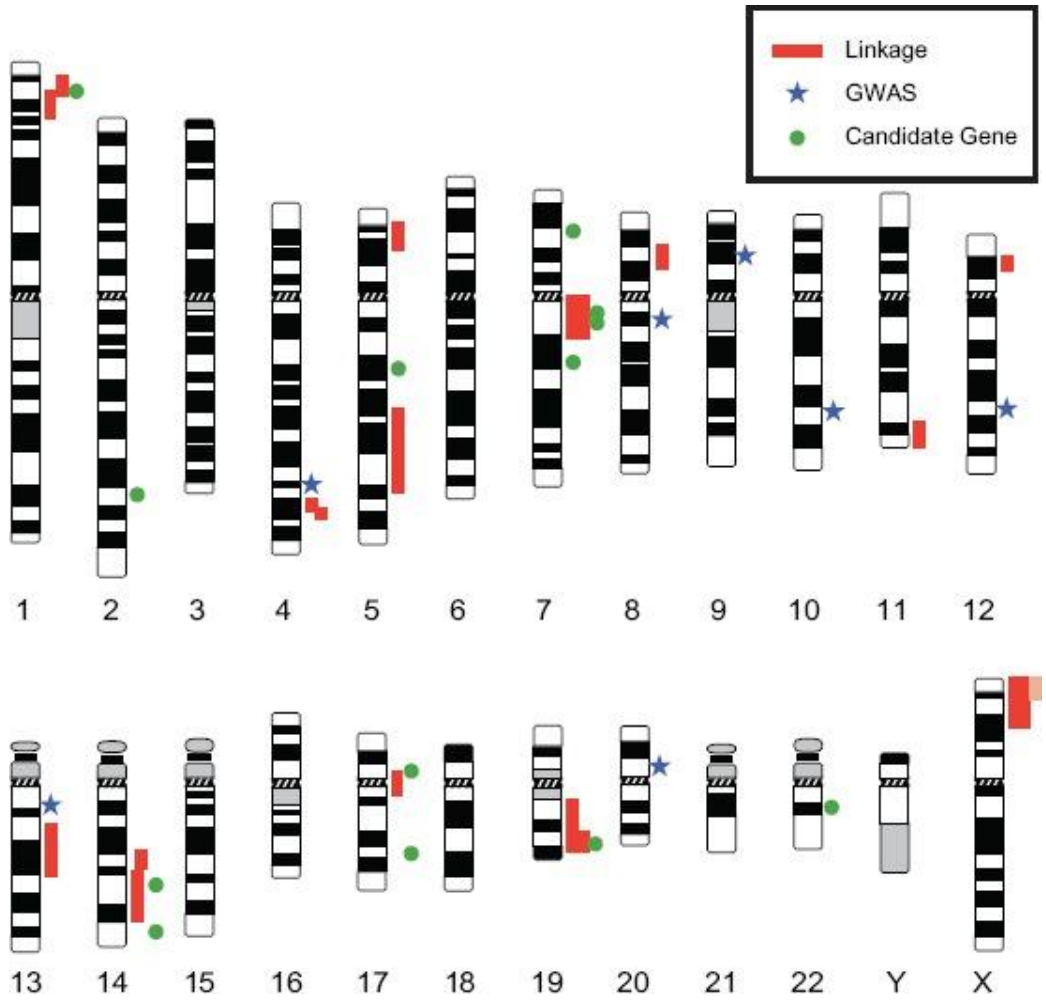


Figure 6. Chromosomal map of IA loci identified through linkage, candidate gene and GWAS studies, and prioritized by Tromp et al. (2014). Used with permission under the Creative Commons license CC-BY-NC-ND 3.0 for the open access article by Tromp et al. (2014).

comprise about 1% of the human genome, but contain a vast amount of genetic variation with potential functional roles. A single exome of European-American descent contains an average of 20,000 coding variants, the majority of which are non-novel (Bamshad et al., 2011). Despite the abundance of known, common polymorphisms, an increasing large number of research groups have been able to use WES to describe rare, pathogenic variants in heritable diseases.

Although it is difficult to identify causative variants in late-onset heritable disorders, pathogenic variants have been successfully elucidated through WES. Zimprich et al. (2011) employed WES to identify the causative variant for late-onset Parkinson's disease in a large Austrian family, with autosomal dominant inheritance. This neurological disease affected sixteen members of the family, and seven participated in this study. The mean age of onset for the cohort was 53, and two affected second cousins were chosen for WES. Following filtering for rare heterozygous variants, a variant in the *VPS35* gene, c.1858G>A, was prioritized. It was present in all seven affected individuals, and was later shown to segregate in two additional Austrian families with Parkinson's disease (Zimprich et al., 2011). This example demonstrates that WES-based studies have great potential where candidate gene sequencing has yielded negative results. WES is now widely used as a method to solve these cases, and explore diseases with genetic heterogeneity.

Boileau et al. (2012) used WES to uncover the cause of familial TAAD in two unrelated families exhibiting autosomal dominant inheritance. Some family members without TAAD had a vascular phenotype in the form of IA and SAH. Mutations in the *SMAD3* gene have been previously identified in individuals with similar phenotypes of

TAAD in five other families (Regalado et al., 2011). Additionally, *TGFBR2* and *FBN1* are also known genes for thoracic aortic aneurysm (TAA) susceptibility (Faivre et al., 2009; Loeys et al., 2006). However, in the Boileau et al. (2012) study, candidate gene sequencing did not yield any causative variants. Genome-wide linkage analysis revealed a significant locus at 1q41 in both families, which was then followed by WES to further explore missing heritability. Two affected individuals from an American family, TAA288, were exome sequenced, along with three affected and one unaffected individual from a French family, MS239. Following filtering of the exome data, both families had rare variants in the *TGFB2* gene, which is located in the linked 1q41 region. The 5-bp deletion (TAA288) and nonsense variant, p.C229* (MS239), were absent from control exomes, and were predicted to cause TAAD by haploinsufficiency of *TGFB2*.

Given the success of WES in exploring adult-onset and vascular phenotypes, Foroud and colleagues used this technology to study familial IA (Foroud & FIA Study Investigators, 2013; Farlow et al., 2015). They sequenced the exomes of 45 individuals from seven families. The multiplex families published in this manuscript had a pedigree appearance of either autosomal dominant or recessive inheritance of IA. Through the use of the Genome Analysis Toolkit (GATK) Unified Genotyper software, a list of genetic variants was detected for each exome. A series of biological filters were then used to prioritize these variants, by factors such as minor allele frequency and gene function. The research group published a list of 68 rare variants in 68 genes that segregated with IA in at least one family (Farlow et al., 2015). Based on RNA expression data for 51 of these 68 candidate genes, only the *TMEM132B* gene showed differential expression in IA tissue as compared to controls. In future analyses of the data, the authors plan to prioritize

the *TMEM132B* c.3050C>T (p.S1017L) variant, to determine its possible contribution to disease.

A second WES-based study of IA was performed by Yan et al. (2015), and involved 42 individuals from 12 multiplex Japanese families. Following a systematic filtering approach, 78 single nucleotide variants were prioritized across this cohort. Only two of these variants were found in affected individuals from more than one family: c.578T>A (p.Y193F) in *GPR63* and c.425C>T (p.R142H) in *C10orf122*. In addition, they completed an association study of 10 out of 78 variants, which were found in genes with gene ontology (GO) terms relating to angiogenesis. A variant in *ADAMTS15*, c.397G>C (p.E133Q), had the only significant association with IA following Bonferroni correction, with a P-value of 0.00013. Silencing of this gene in human umbilical venous endothelial cells caused increased endothelial cell migration, though further functional analysis at the cellular level is necessary (Yan et al., 2015). Both of these WES studies of IA have provided a starting point for delving into further exome-based research, though functional data will be required to determine if any of these variants are actually pathogenic and involved in IA development. These studies emphasize the use of advanced techniques in next generation sequencing to complement widely used genetics methodologies. Similarly, my research involves the application of WES to familial IA in Newfoundland and Labrador, where certain genetic diseases have been deeply phenotyped.

1.3 Studying IA in Newfoundland and Labrador

1.3.1 Heritable Disease Research in NL

The province of Newfoundland and Labrador has a unique population and history. The current population has largely grown from a group of initial immigrants, mainly of Irish and English descent, dating back to the mid-1700s (Rahman et al., 2003). The settlement of small outport communities around the province, combined with minimal immigration over time, has led to both geographic and genetic isolation (Rahman et al., 2003). As a result, several rare monogenic diseases have an increased prevalence in this province, including Bardet-Biedl syndrome (Young et al., 1999) and arrhythmogenic right-ventricular cardiomyopathy (Merner et al., 2008). In addition, several complex diseases have a high prevalence in the NL population, including familial colorectal cancer (Green et al., 2007). Given the successful identification of mutations implicated in genetic disorders, the population of Newfoundland and Labrador has become attractive for genetics research (Rahman et al., 2003).

The recruitment process for IA families has been a long-term effort in this province. In our collaboration between the Discipline of Genetics at Memorial University and Eastern Health, 53 Newfoundland families with a strong family history of IA have been identified and enrolled in our study. Three families from Newfoundland and Labrador affected by familial IA were first described in a case study published by Maroun et al. (1986). The authors state that there is an increased prevalence of familial central nervous system disorders in the province. Six of seven SAH patients from the three families were diagnosed with ruptured IA, and the seventh individual was suspected

of having an IA. The average age of onset for these patients was relatively low, at 41.6 years of age. A summary of the IA study cohort will be described in the Materials and Methods section of this thesis. Due to the previous success of heritable disease research in NL, and the genetic isolation of the population, this province is an ideal place to investigate the genetic etiology of IA.

1.3.2 Previous Work Completed

Prior to the current study, several experiments were conducted in the Woods Laboratory using the IA cohort from Newfoundland and Labrador. In 2009, a study was published that explored the possibility of common genetic mechanisms between aortic aneurysms and IA (Santiago-Sim et al., 2009). Mutations in transforming growth factor beta (TGF-β) receptor genes are known to play a role in some familial cases of thoracic aortic aneurysm and dissection. To investigate the role of disrupted TGF-β signaling in IA risk, several TGF-β receptor genes, and genes related to this pathway, were sequenced in 44 familial IA patients, including *TGFBR1*, *TGFBR2*, *TGFBR3*, *TGFB1*, *ACVR1*, and *ENG* (Santiago-Sim et al., 2009). Novel variants in *ENG* (endoglin) and *TGFBR3* (transforming growth factor, beta receptor III) were identified in a subset of the cohort, but the pathogenicity of these variants could not be determined without further functional analysis (Santiago-Sim et al., 2009). Following the publication of this article, a research assistant in our laboratory completed Sanger sequencing of these two candidate genes in all probands from our IA families (53 in total) and in our cohort of sporadic individuals (N=33). However, no rare variants were identified in either of these genes. It was decided that sequencing of candidate genes might not be the best approach for our study of IA. As

we continue to investigate this disease, it is clear that the presence of genetic heterogeneity, and a general lack of knowledge surrounding the genetics of IA, may hinder the effectiveness of this method in the Newfoundland population.

In addition to candidate gene sequencing, SNP genotyping was outsourced for seven of our IA families using the Illumina Human610-Quad chip, followed by genome-wide linkage analyses. The selected families (R1256, R1352, R1357, R1276, R1277, R1400 and R1888) have a particularly strong incidence of IA, and are highlighted in section 2.2. In total, 64 individuals were genotyped, for 620,901 SNPs. The final dataset contained 59 samples as five samples had a low call rate or were too distantly related for linkage considerations. The final number of SNPs was 574,441, as SNPs with a call rate below 99% were removed. Two-point and multipoint linkage analyses were performed, and both dominant and recessive modes of inheritance were considered. For the dominant inheritance model, family R1256 had several markers on chromosomes 6, 10, and 14 that had a two-point LOD score greater than 2.0. For this report, LOD scores greater than or equal to 2.0 were considered to be suggestive of linkage, with theta values equal to 0. For the multi-point analysis on this same family, only three markers in a 22 cM region of chromosome 14 had a LOD $\geq$ 2.0. For the other families, neither had LOD scores $\geq$ 2.0, using the dominant model. With a recessive model, no LOD scores $\geq$ 2.0 were identified in either of the families. Finally, a multi-point analysis using all seven families revealed a LOD score of 1.9 on chromosome 14.

Interestingly, two previous IA studies have identified loci that overlap with this region on chromosome 14 (near rs17105585, rs11158743 and rs1956534). Ozturk et al. (2006) found significant linkage to chromosome region 14q23-31 in a Japanese family,

while Mineharu et al. (2008) found a SNP (rs767603) at Chr 14q23 that was significantly associated with IA in a larger cohort from Japan. However, no candidate genes were identified in either of these studies. This previous work has provided our laboratory with some preliminary data and the insight to employ a new approach for the continuation of this project. Any candidate variants on chromosome 14 will be of higher priority if they are detected through whole exome sequencing in our current study.

1.3.3 Utilizing Whole Exome Sequencing

The introduction of next generation sequencing has provided a new way to explore disease-gene relationships, and has broadened the scope and success of genetic research. As mentioned previously, whole exome sequencing is a next generation technology that targets the protein-coding portion of the genome, of which we have the greatest understanding. Any variants that cause IA predisposition will likely impact the structure and proper functioning of specific proteins. As no strongly penetrant genes have been connected to the non-syndromic occurrence of IA, WES theoretically provides a way to visualize all the pathogenic exonic variants present in an individual affected by this disease.

As with all sequencing technologies, several points need to be considered in a WES-based experiment. Different commercial platforms are currently available for massively parallel sequencing of the exome, including workflows by Illumina and Agilent. The steps of the Illumina HiSeq 2000 workflow will be discussed in section 2.3. Each platform is similar, but may yield varying coverage of the exome. Bamshad et al. (2011) reported that as much as 5-10% of the exome may have insufficient coverage or

may be skipped completely in a WES experiment, depending on the chosen commercial platform. The data analysis pipeline that is used following WES also greatly influences the identification of causal variants. Accuracy in alignment, quality control and variant calling is crucial, especially given the large amount of data produced from a single exome. At this point in time, only the detection of SNPs and some small insertion/deletions (INDELs) is possible using WES. Therefore, our focus lies within the current capabilities of this technology.

Ultimately, WES may allow us to develop a better picture of the genetic make-up of familial IA patients than would be possible by traditional sequencing methods. By analyzing the exomes of multiple affected individuals within a family, I will be able to draw connections between shared genetic variants. The commercialization and increased use of WES in research has led to an abundance of user-friendly bioinformatics tools and variant databases that are now publically accessible. These tools will assist in evaluating variant pathogenicity and relevance to IA.

1.4 Hypothesis

Based on current knowledge of this disease, it is predicted that the multiplex families in this study have a genetic component contributing to the incidence of IA. I hypothesize that the families in this study will have one or more strongly penetrant variants that cause IA. Furthermore, I hypothesize that these genetic variants will be rare or novel, and located in the exome or intron/exon boundaries of the genome.

1.5 Objectives and Relevance of Research

The purpose of my study is to identify genetic variants that cause IA susceptibility, through the study of families from Newfoundland and Labrador that have a strong predisposition. My study is designed around the following objectives:

1. To design and implement a strategy to filter the genetic variants identified in the exomes of 12 IA patients from 2 multiplex families.
2. To analyze filtered low, moderate and high impact variant lists separately, based on categorization by McGill University and Genome Quebec Innovation Centre. Also, to compare these variant lists to the results produced by NextGene®, and determine the potential utility of this desktop software in our laboratory.
3. To use Sanger sequencing to validate prioritized genetic variants, and test for segregation in the families.
4. To identify several candidate variants, discuss their potential relevance to the IA phenotype, and determine their prevalence in a NL population control cohort.

Through the completion of this study, I hope to contribute to our growing understanding of the genetic etiology of IA. Due to genetic heterogeneity, there is likely to be more than one mode of inheritance for familial IA. The knowledge obtained from this research could lead to the identification of individuals with unruptured IAs, which may then be treated. Successful treatment can reduce the risk of hemorrhagic stroke and

its complications in IA patients. As well, the identification of genetic risk factors could assist in genetic counseling of patients and their families. It is also possible that genetic variants identified through this study will have importance in other populations globally. Furthermore, my research may lead to further genetic studies involving the cohort of IA patients from Newfoundland and Labrador.

2. Materials and Methods

2.1 Study Design, Patient Recruitment and Clinical Information

To test my hypothesis, a research plan with corresponding rationale was designed (Figure 7). Step one of the methods, participant recruitment, was completed prior to the commencement of this project. The recruitment of IA patients for this study was a collaboration between Dr. Falah Maroun and neurosurgeons from Eastern Health, and Dr. Bridget Fernandez, chair of the Discipline of Genetics. Neurosurgeons in the Division of Surgery referred patients with both ruptured and unruptured IA. Diagnosis occurred through the use of computed tomography scan (CT) or magnetic resonance imaging (MRI) of the Circle of Willis. Eastern Health has the only unit of neurosurgery for Newfoundland and Labrador, thus providing us with a patient cohort from across the province.

Each affected individual was placed into one of three categories: familial, sporadic, or equivocal. Familial cases are patients who have at least one first or second-degree relative with IA, and sporadic cases are patients with no known family history of the disease. Finally, patients are classified as equivocal if there is not yet enough evidence to determine whether their IA is sporadic or familial in nature. A three-generation pedigree was taken for each participant. The active participation of IA patients was instrumental to the recruitment of additional affected and unaffected relatives to this study. The medical histories of any consenting relatives were reviewed, and they were offered screening of the Circle of Willis through computed tomography angiography, to identify any unknown cases of IA. A blood sample was collected from all affected and

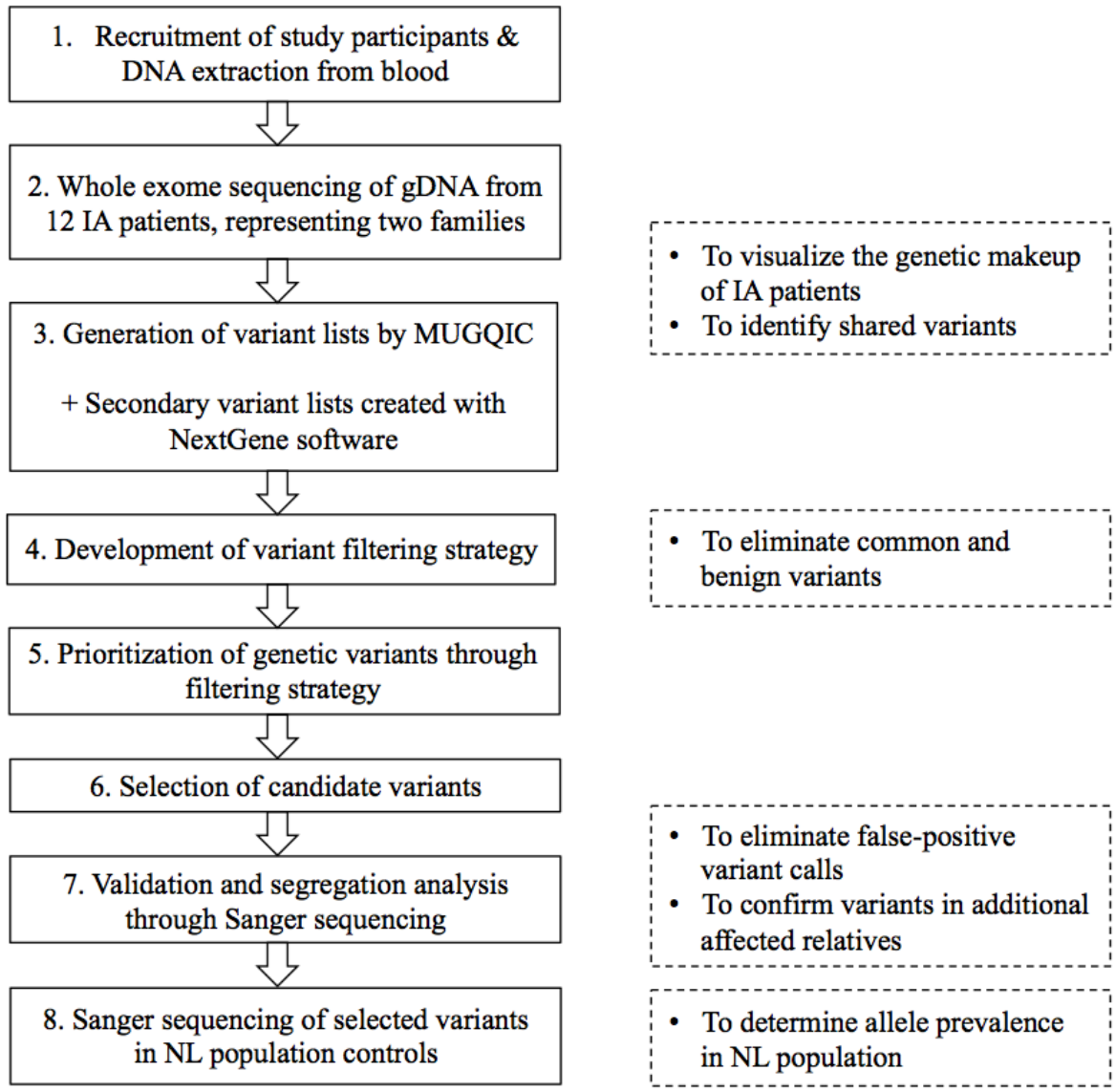


Figure 7. Overall study design, with corresponding rationale.

Abbreviations: gDNA = genomic DNA, NL = Newfoundland and Labrador, MUGQIC = McGill University and Genome Quebec Innovation Centre

unaffected study participants, for the extraction of DNA. Blood samples were sent to the Discipline of Genetics, where DNA extraction from whole blood was performed using the Promega Wizard® Genomic DNA Purification Kit, which is described in detail in Appendix A. Genomic DNA was catalogued in the Woods Lab DNA Bank, and samples were stored at 4°C.

In addition, a thorough collection of phenotypic information was recorded for all affected participants (familial, sporadic and equivocal). This included the patient's sex, age at diagnosis, whether diagnosis was made through clinical screening or following rupture, number of IAs, location of IA in vasculature, size of IA, age at rupture, and any treatments administered. The detailed phenotypes were provided to our laboratory, and could be highly valuable at the later stages of this study, for example for investigating the connection between genetic etiologies and the varying severity and presentation of the disease. For all study participants, any additional medical history was also documented. Several risk factors for IA were also recorded, including hypertension, smoking status and smoking history.

As of April 2014, 137 affected individuals had been recruited to the study, including 99 familial IA patients and 396 of their unaffected relatives representing 53 families. DNA samples are available for 92 of these familial cases. The remaining affected individuals were categorized as 33 sporadic IA patients and five equivocal cases. These individuals also consented to have a blood sample drawn for the purpose of DNA extraction. Eight of the participating families are of particular interest, as they each have three or more affected family members for whom DNA samples are available (Table 4).

Table 4. Families from NL cohort with more than three affected individuals.

Family ID	# of Reported Affected Family Members	# Affected, with DNA available
R1256	12	11
R1276	4	4
R1277	8	6
R1352	11	9
R1357	4	4
R1381	5	3
R1400	6	3
R1888	5	4

The number of reported affected family members is based on provided pedigrees and clinical data as of 2014. DNA was not available for several affected relatives who did not provide consent to participate in the research study. As well, several individuals were deceased prior to the commencement of patient recruitment.

2.2 Study Families and Mode of Inheritance

For my thesis, I focused on two of the recruited families that have a particularly strong family history of the disease: R1256 and R1352. In this way, my thesis research has represented a pilot project of WES with our IA cohort. Family R1256 is a large kindred with origins in Happy Valley-Goose Bay (Figure 8, 9). In contrast, family R1352 has origins in the St. Mary's Bay region of the island (Figure 10). A detailed clinical summary of the affected participants is provided in Tables 5 and 6. In addition to family history, several of these individuals have relevant modifiable risk factors, including cigarette smoking, hypertension, and alcohol use, which have been noted for their potential influence in IA development. All affected individuals were diagnosed between the ages of 29 and 79, with a mean age of initial diagnosis of 50.2 for R1256, and 54.9 for R1352. In both families, the number of females with IA was greater than the number of males with the disease. In family R1352, eight females and three males were affected, and in family R1256, nine females and three males were affected.

Given that clinical information was available for each family over several generations, predictions can be made regarding the mode of inheritance. The pedigree appearance for family R1256 is consistent with an autosomal recessive mode of inheritance, as most of the affected individuals (11/12) are in a single generation (Figure 9). If there is incomplete penetrance, as well as unknown aneurysm diagnoses in deceased family members, the mode could also be autosomal dominant with variable penetrance. In family R1352, the presence of consanguinity affects interpretation of the pedigree. Two affected individuals in this pedigree (Z1496 and Z1495) are the parents of

eight children, and six of the children have the disease. Individual Z1495 has another son by a different partner, Z1507, who also has the disease phenotype. Given the consanguinity between Z1496 and Z1495, the mode of inheritance in this family could be either autosomal dominant or recessive. If the mode is autosomal dominant, all affected individuals could be either heterozygous or homozygous for the causative variant. For the purposes of our study, both modes of inheritance will be considered in the interpretation of genetic variants in these families. The complex nature of IA including penetrance and the influence of modifiable risk factors will be explored further in section 4.

Figure 8. Geographic representation of families R1256 and R1352 in province of Newfoundland and Labrador. Adapted from NordNordWest (2009) with copyright permission, under Creative Commons license CC BY-SA 3.0.

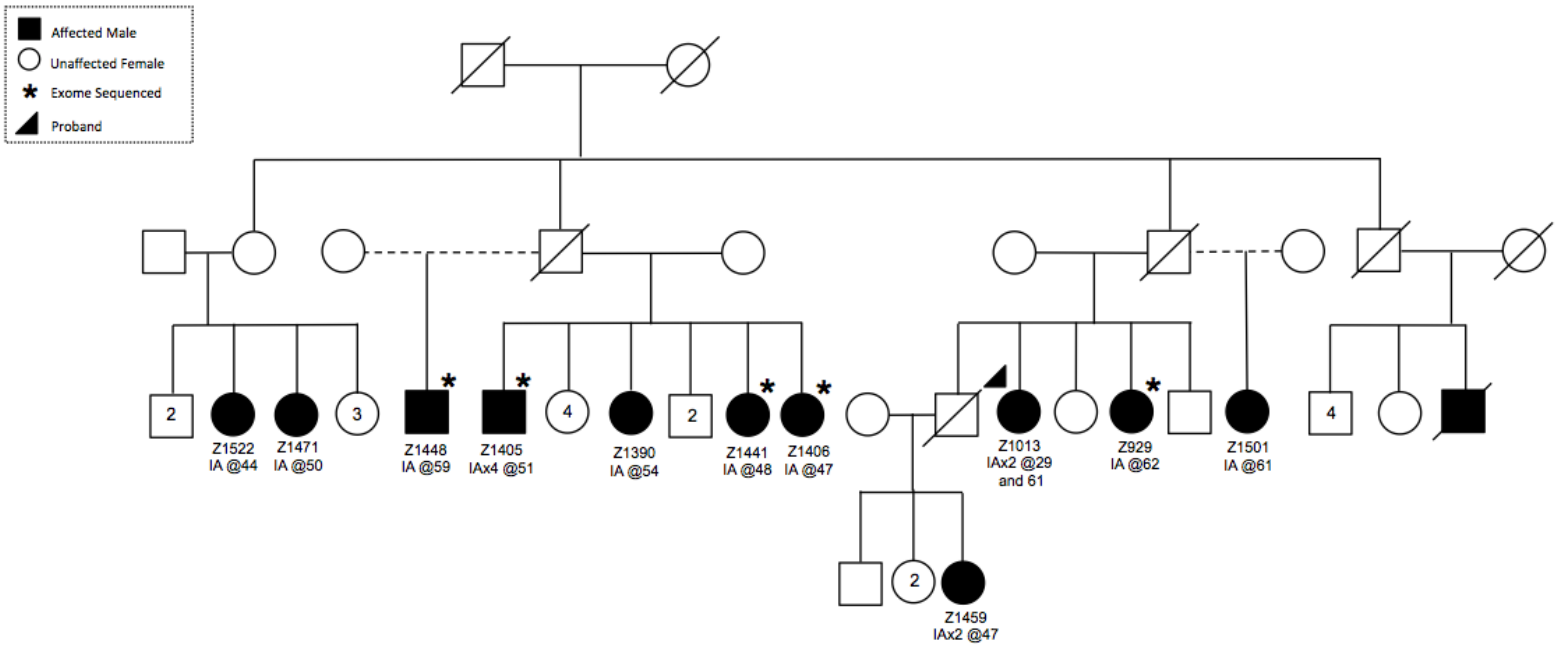


Figure 9. Condensed pedigree for family R1256.

DNA extraction was completed for individuals with a Z identification number. IA affected status including the patient age at diagnosis and number of aneurysms is indicated. An asterisk is used to indicate individuals selected for whole exome sequencing.

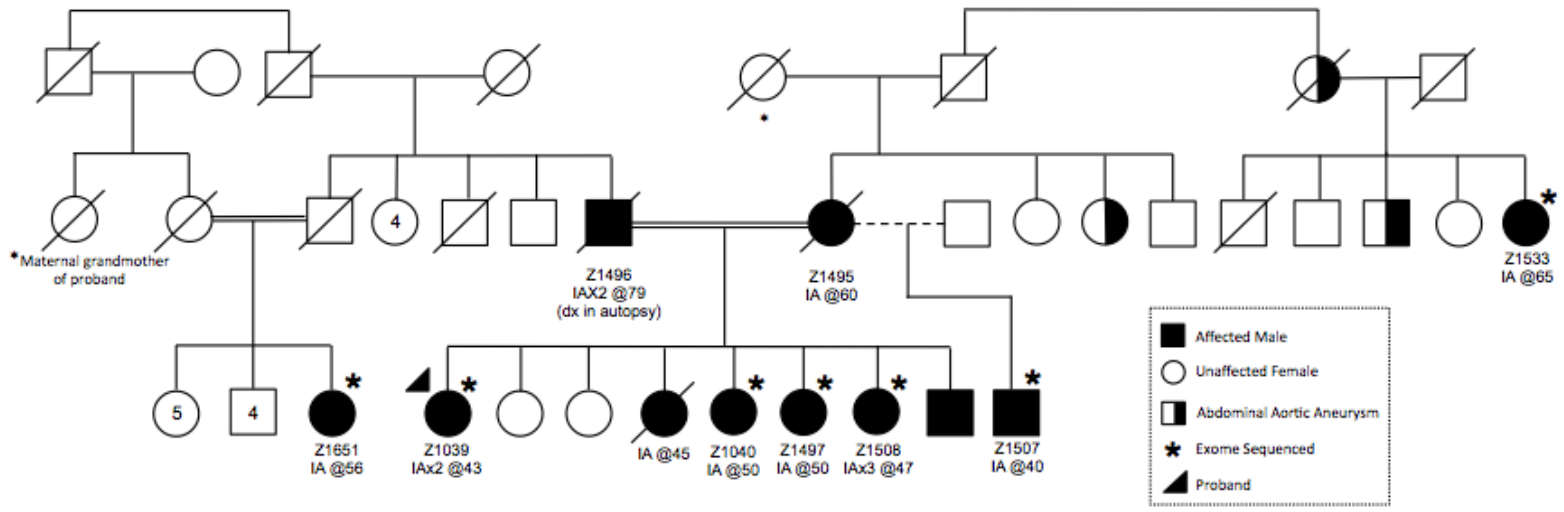


Figure 10. Condensed pedigree for family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affected status including the patient age at diagnosis and number of aneurysms is indicated. An asterisk is used to indicate individuals selected for whole exome sequencing.

Table 5. Phenotypic summary of affected family members from family R1256.

Patient ID	# of IAs	Size of IA(s)	Site of IA(s)	Type of Treatment Given	IA Rupture (Yes/No)	Risk Factors & Known Medical History
Z929	1	4.5 mm	R-MCA	No surgery (observation)	No	Hypertension, coronary artery bypass surgery
Z1013	2	Unknown; 4.5x5.5 mm	R-MCA x2	Clipping	Yes	Not available
Z1390	1	1.5x2.5 mm	A1-A2 Junction (L)	No surgery (observation)	No	Hypertension, smoker, heart murmur
Z1405	4	13-14 mm; 4-5 mm; 2 mm; 2 mm	Paraophthalmic; L-paraclinoid; R-ICA x2	Coiling	No	Ex-smoker, hypercholesterolemia, allergies, asthma, GERD, fibromyalgia
Z1406	1	5x10 mm	L-ACOMM	Clipping	Yes	Hypertension, hyperlipidemia
Z1441	1	3.6x3.2 mm	L-ICA	No surgery (observation)	No	Hypertension, smoker
Z1448	1	1.5 mm	R-MCA	No surgery (observation)	No	Hypertension, ex-smoker, hypercholesterolemia
Z1459	2	7x8 mm; 9 mm	L-ACOMM; R-PICA	Coiling; Clipping	No	Hypertension, ex-smoker, diabetes mellitus, obesity, asthma
Z1471	3	1.9x1.3 mm; 2.4x1.7 mm; 1.3x0.8 mm	PCOMM; Bilateral Distal ICA; Unknown	No surgery (observation)	No	Smoker
Z1501	1	3x4 mm	ACOMM	No surgery (observation)	No	Hypertension, ex-smoker, hypercholesterolemia, diabetes mellitus, osteoarthritis, Burkitt's lymphoma, diverticulosis
Z1522	1	10 mm	R-ICA	Coiling	No	Smoker, hypothyroidism, migraines, reflux

Abbreviations: ACOMM = anterior communicating artery, GERD = gastroesophageal reflux disease, ICA = internal carotid artery, L = left, MCA = middle cerebral artery, PCOMM = posterior communicating artery PICA = posterior inferior cerebellar artery, R = right

Table 6. Phenotypic summary of affected family members from family R1352.

Patient ID	# of IAs	Size of IA(s)	Site of IA(s)	Type of Treatment Given	IA Rupture (Yes/No)	Risk Factors & Known Medical History
Z1039	2	10x7 mm; 3 mm	L-ICA; R-Distal ICA	Clipping	No	Hypertension
Z1040	2	Unknown	L-ICA x2	No surgery (observation)	No	Reflux, depression
Z1495	1	Moderate	R-MCA	Clipping	Yes	AAA, smoker, COPD
Z1496	2	Small; Small	L-MCA; R-MCA	No surgery (observation)	No	Smoker, heavy alcohol intake, Steele-Richardson Olszewski syndrome, COPD, peptic ulcer disease
Z1497	1	Small	Paraclinoid	No surgery (observation)	No	Hypertension, hypercholesterolemia, back pain
Z1507	1	2.8x4.8 mm	ACOMM	No surgery (observation)	No	Smoker, myocardial infarction, atrial fibrillation, splenic artery pseudoaneurysm, alcoholic pancreatitis, back pain
Z1508	2	Small; 4 mm	L-MCA; R-ICA	Coiling	No	Smoker, GERD
Z1533	1	9.4x5.6 mm	R-MCA	Clipping	No	Smoker
Z1651	1	1.5-2 mm	L-Paraophthalmic	No surgery (observation)	No	Not available

Abbreviations: AAA = abdominal aortic aneurysm, ACOMM = anterior communicating artery, COPD = chronic obstructive pulmonary disease, GERD = gastroesophageal reflux disease, ICA = internal carotid artery, L = left, MCA = middle cerebral artery, R = right

Several family members were subsequently selected for whole exome sequencing. For a single complete WES run, six samples can be accommodated. For this pilot project, 12 affected individuals were chosen between both families for a total of two runs: five from R1256 (Figure 9), and seven from R1352 (Figure 10). Where possible, a diverse array of relatives was selected for sequencing. Affected family members who are more distantly related, such as cousins, can be highly informative as they share fewer variants than closely related individuals, such as siblings (Bamshad et al., 2011). Often, WES experiments involve the inclusion of at least one unaffected relative as a control sample to help elucidate pathogenic variants. Several features of IA as a disease prevented the use of this methodology. IA is a late-onset condition and within each of these two families the penetrance is unknown. It is possible that an included "unaffected" relative could develop IA later in life, or may carry an IA risk allele but not display disease characteristics due to decreased penetrance. However, many WES studies do not have access to multiplex families of this size. The comparison of five or more affected exomes within a single family is expected to assist in the identification of shared rare variants, and should help overcome the lack of an unaffected control exome. It is expected that the affected family members will have a shared genetic risk factor. Our laboratory has also performed exome sequencing for other genetic diseases including colorectal cancer and idiopathic pulmonary fibrosis; therefore we have access to data from unrelated NL samples that can be used as controls in this study if necessary.

2.3 Whole Exome Sequencing and Bioinformatics Analyses

Whole exome sequencing and preliminary bioinformatics analyses outlined in this section were completed at the McGill University and Genome Québec Innovation Centre (MUGQIC). The following descriptions are adapted from instructional videos and materials provided by MUGQIC (MUGQIC Bioinformatics, 2014; Schwartzentruber, 2012), technical documents from Illumina®, and a review by Bamshad et al. (2011).

2.3.1 Library Preparation and Sequencing

Initially, it is necessary to amplify and isolate the exons and intron/exon boundaries of the target DNA samples. MUGQIC employs the "Exome Seq – Sure Select Agilent" protocol for shotgun sequencing library preparation. DNA is required to meet specific quality standards, including sample concentration and volume. For this project, concentrations were adjusted to 100 ng/µl, and 70 µl of each sample was provided. Each DNA sample was sheared into fragments of approximately 200 base pairs (bp) in length, through ultrasonication. Fragments were then ligated with 5' and 3' adapters to repair the sequence ends, and amplified through PCR. Next, a hybridization step was used to capture the exome; ligated fragments were hybridized to biotinylated RNA baits, which are designed to target the exon sequences. Baited fragments were "pulled down" with streptavidin covered magnetic beads, which bind to biotin. This step separated out any uncaptured non-coding DNA.

For our study, massively parallel sequencing was completed using an Illumina® HiSeq 2000 sequencer. Ligated fragments were loaded into a flow cell, where they were bound to complementary oligos. Bridge amplification was used to create sample clusters,

each of which produced a paired-end read. The Illumina® system utilizes sequencing by synthesis (SBS) technology to generate reads, which are lengths of A, T, C and G bases. Base calls were made using the Illumina® CASAVA pipeline.

2.3.2 Alignment and Quality Control Methods

The next steps in the WES process concern data analysis, which is integral to the accurate annotation and interpretation of genetic variants. Raw reads were trimmed and adapters were removed, so that filtered reads were at least 50 base pairs in length. Sequencing reads were then aligned to the human reference genome version GRCh37 (hg19), using Burrows Wheeler Alignment (BWA) software version 0.6.2 (H. Li & Durbin, 2009). Several quality control methods were then employed to ensure high sequence quality. This included the realignment of reads in areas with multiple base mismatches, which was done with Genome Analysis Toolkit (GATK) software version 2.7.2 (McKenna et al., 2010). During the sequencing process, PCR duplicates can occur when some sheared fragments become localized to multiple beads of the flow cell. This results in over-amplification of some regions of the exome. Duplicate reads were marked and excluded using Picard software version 1.108. Overall base quality was then reassessed, in preparation for variant calling and annotation.

2.3.3 Variant Calling and Annotation

Any discrepancies between the exome sequence and reference genome sequence were classified as genetic variants. Both SNPs and INDELs were detected using the programs samtools and bcftools. Samtools (version 0.1.19) (H. Li et al., 2009) collects

summary information, and bcftools completes the actual variant calling. Through the use of Bayesian statistics, information such as the number of reads and sequence quality was considered, to predict whether a mismatch was a true variant or a false-positive call. Bcftools was then used to transform this information into the variant call format (VCF) (H. Li, 2011). The ANNOVAR program (K. Wang, Li, & Hakonarson, 2010) uses these VCF files to annotate the variant type - for example: non-synonymous SNPs, nonsense, frameshift, splicing, etc. This program includes SnpSift software (Cingolani, Patel et al., 2012) to provide annotations from dbSNP (Smigielski, Sirotkin, Ward, & Sherry, 2000). As well, ANNOVAR includes annotations from the Database for Non-Synonymous SNPs Functional Predictions (dbNSFP version 2.0). This database integrates the scores from several functional predictive algorithms, for all possible non-synonymous SNPs in the human genome (Liu, Jian, & Boerwinkle, 2013). Our results included scores compiled from SIFT (Kumar, Henikoff, & Ng, 2009), Polyphen2 (Adzhubei et al., 2010), LRT (Chun & Fay, 2009) and MutationTaster (Schwarz, Rodelsperger, Schuelke, & Seelow, 2010). These four algorithms each provide a predictive score that can be used to assess the likelihood that a particular amino acid change is pathogenic. As well, PhyloP (Siepel & Haussler, 2004), GERP (Cooper et al., 2005) and SiPhy (Garber et al., 2009) scores were included, which can be used to predict the evolutionary conservation of an amino acid change. Finally, the SnpEff software (version 3.3) was used to predict the effect or impact of a variant on an overall gene (Cingolani, Platts et al., 2012). Variants were placed in one of three categories: high, moderate or low impact (Table 7).

Bioinformaticians at MUGQIC grouped the VCF files and annotations for each of our exomes by family, and also by impact. Thus, we were provided with separate high,

moderate and low impact lists for R1256 and R1352. As an initial filtering step, the bioinformatics team at MUGQIC recommended the removal of variants with a read depth lower than 10 in all exomes from a single family. This step is a quality control measure used in next generation analysis to remove false-positive calls and variants with low coverage. The VCF files were subsequently transferred to Microsoft Excel spreadsheets for ease of data retrieval and filtering. These lists included any genetic variants that were detected in one or more exomes from a single family.

For the purposes of my study, I will be filtering all three variant impact categories, but considering only *variants of high and moderate effect* for validation and further analysis at this time (Table 7). It is expected that a variant contributing to IA development and rupture risk would have a significant impact on both the gene and protein level. To deal with this large amount of data, the next step in my methodology involved the development of a variant filtering strategy.

Table 7. Variant impact categories.

	Variant Types Included
High Impact	Splice site acceptor, splice site donor, start-lost, exon deleted, frame-shift, stop-gained, stop-lost, rare amino acid substitution
Moderate Impact	Non-synonymous coding (missense), codon change, codon insertion, codon deletion, codon change plus codon insertion, codon change plus codon deletion, 5'UTR deleted, 3'UTR deleted
Low Impact	Synonymous start, non-synonymous start, start-gained, synonymous coding, synonymous stop

Criteria determined by McGill University and Genome Quebec Innovation Centre, through the use of the SNPEff software (Cingolani et al., 2012).

Abbreviations: UTR = untranslated region

2.4 Filtering of Low, Moderate, and High Impact Variants

2.4.1 Previously Identified Candidate Genes for IA

As a first step, I searched the moderate and high impact lists for variants in genes that have been significantly associated with IA in previous studies. I selected a recent review article by Tromp et al. (2014) that lists the statistically significant associations with IA that have been identified in published GWAS or candidate gene association studies. Specifically, the 20 genes found in or near these chromosomal loci are highlighted in Table 8. I used the list of genes featured in this article to query the moderate and high impact lists for each family. Any variants identified in IA candidate genes would be highlighted and further analyzed for their relevance to this study.

2.4.2 Variant Filtering Strategy

In order to identify genetic variants that could play a role in familial IA, I developed a strategy to filter out variants that were likely extraneous in this study. This strategy was designed based on a review of WES publications, and considerations regarding the features of IA. First, I began by using the annotations provided by MUGQIC to filter the collection of low, moderate, and high impact variants from all 12 exomes. As only affected individuals were included in this analysis, the assumption was made that any variants of interest would be shared by all exomes from a single family. To account for the possibility of phenocopies and sequencing coverage error, I chose to also keep variants that were shared by 6/7 members of R1352, or 4/5 members of R1256. This step is not a common practice in family-based WES studies, but I believed it would be suitable considering the role of environmental factors in IA development.

Table 8. Summary of genes found in or near variants that have been significantly associated with IA.

Gene	Chr Region	Associated SNPs	Variant Context (Proximity to Gene)	Initial Reference for Association
ACE	17q23.3	*rs4646994*	intronic	Chen et al. (2013)
CDKN2B-AS1	9p21	**rs1333040*	intronic	Bilguvar et al. (2008)
		**rs10757278*	intergenic	Helgadottir et al. (2008)
		**rs6475606*	intronic	Foroud et al. (2012)
COL1A2	7q21	**rs42524*	intragenic (p.P549A)	Yoneyama et al. (2004)
COL3A1	2q32	**rs1800255*	intragenic (p.A698T)	Hua et al. (2008)
CNNM2	10q24.3	*rs12413409*	intronic	Yasuno et al. (2010)
EDNRA	4q31.23	**rs6841581*	intergenic	Yasuno et al. (2011)
ELN	7q11.2	*rs8326*	3' UTR	Akagawa et al. (2006)
FGD6	12q22	**rs6538595*	intronic	Yasuno et al. (2011)
HSPG2	1p36.1	**rs3767137*	intronic	Ruigrok et al. (2006)
IL6	7p15	**rs1800796*	intergenic	Sun et al. (2008)
JDP2	14q24	*rs175646*	intronic	Krischek et al. (2010)
KLK8	19q13.3	*rs1722561*	intronic	Weinsheimer et al. (2007)
LIMK1	7q11.2	*rs6460071*	intergenic	Low et al. (2011)
RRBP1	20p12.1	**rs1132274*	intragenic (p.R891L)	Yasuno et al. (2011)
SERPINA3	14q32	**rs4934*	intragenic (p.A9T)	Slowik et al. (2005)
SOX17	8q11.23	**rs10958409*	intergenic (5'end)	Bilguvar et al. (2008)
		**rs9298506*	intergenic (3')	Bilguvar et al. (2008)
STARD13	13q13	*rs9315204*	intronic	Yasuno et al. (2010)
TCN2	22q12	*rs1801198*	intragenic (p.R259P)	Semmler et al. (2008)
TNFRSF13B	17p11.2	*rs4985754*	promoter	Inoue et al. (2006)
		rs2274892	intronic	Inoue et al. (2006)
		rs34562254	intragenic (p.P251L)	Inoue et al. (2006)
		rs11078355	intragenic, synon.	Inoue et al. (2006)
VCAN (CSPG2)	5q14	**rs173686*	intronic	Ruigrok et al. (2006)
		**rs251124*	intronic	Ruigrok et al. (2006)

Abbreviations: Chr = Chromosome, SNP = single nucleotide polymorphism, synon = synonymous

*These variants were also reported in a meta-analysis by Alg et al. (2013) as being significantly associated with IA. They used random-effects and sensitivity analyses to determine the robustness of previously-reported associations from candidate gene association studies and GWAS for IA.

Since IA is not a highly common disorder in the general population, I chose to focus on variants that were not previously reported, or had a low minor allele frequency (MAF) in dbSNP (Smigielski et al., 2000). The exclusion of variants with a MAF greater than or equal to 1% is commonly used in the study of recessive disorders, while a cutoff lower than 1% may be applied in autosomal dominant cases (Bamshad et al., 2011). To avoid an overly stringent interpretation, the 1% cutoff was maintained in this study. Bamshad et al. (2011) referred to this step as "discrete filtering", to provide a manageable list of prioritized rare variants that can be further assessed for functional relevance. Variants with a MAF ≥1% in a control set of 848 exomes from Genome Quebec were also excluded.

While this first phase of filtering eliminated a large number of variants, further reduction of the data was necessary. In recent years, the accessibility and decreased cost of WES has led to its frequent use in gene discovery. As a result, a number of exome variant databases have emerged, most notably, the NHLBI Exome Variant Server and the ExAC Browser. The most recent National Heart Lung and Blood Institute (NHLBI) dataset includes 6,503 exomes from American projects targeting heart, lung and blood disorders (Exome Variant Server, 2015). The ExAC Browser from the Broad Institute currently contains exome data from 60,706 individuals, who were sequenced in various large-scale genetics projects internationally (Exome Aggregation Consortium, 2015). This browser, in particular, provides an invaluable resource of control exomes with variant frequencies. Similarly, Ensembl is a continually updated genome browser that unites frequency data from dbSNP, 1000 Genomes (1000 Genomes Project Consortium et al., 2012) and NHLBI, as well as individual laboratories (Cunningham et al., 2015).

After the first phase of filtering, I manually searched for the presence of remaining variants in these three browsers, beginning with ExAC. Any variants with a reported MAF greater than or equal to 1% were eliminated. A flow chart clearly depicting these two phases of filtering methodology is provided in Figure 11.

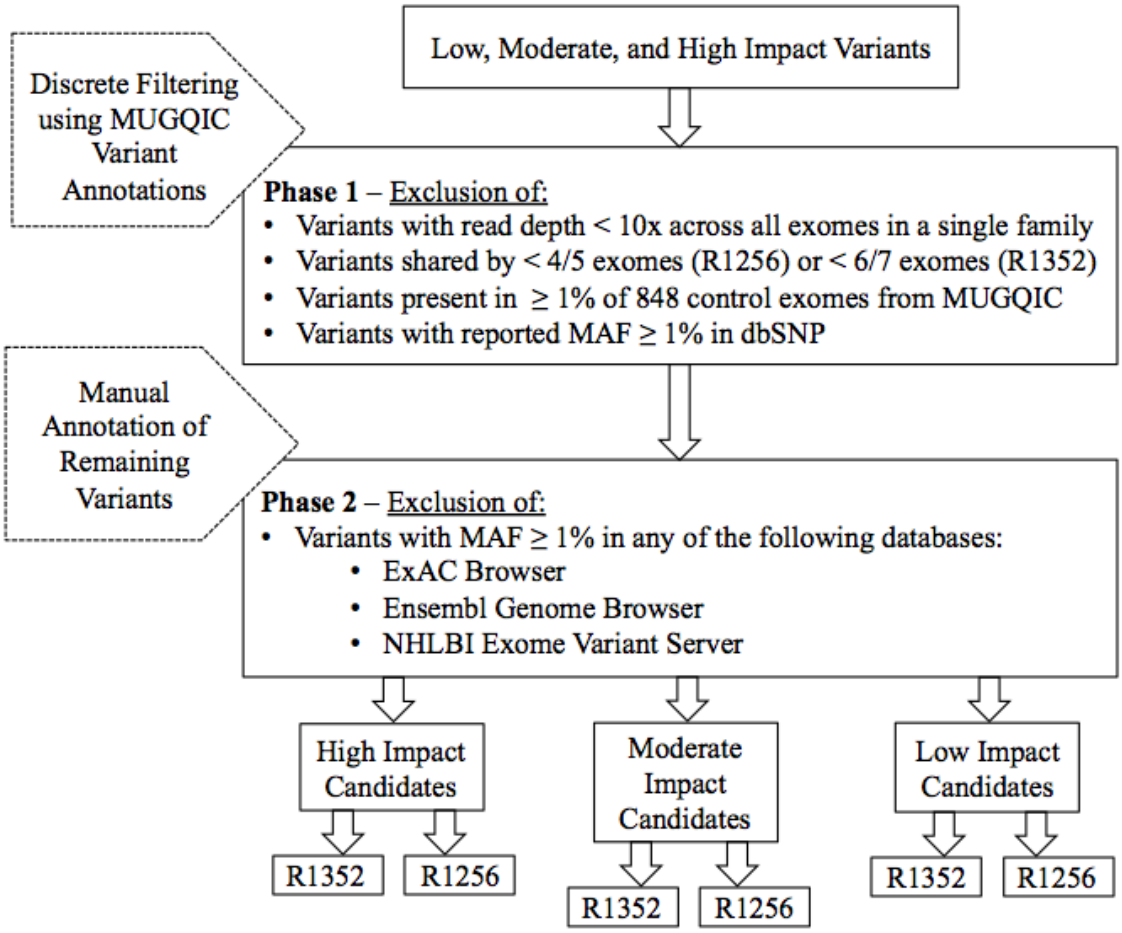


Figure 11. Preliminary variant filtering strategy for 12 IA exomes.

2.5 NextGene® Software

In addition to the data analysis pipeline used by MUGQIC, I performed some general bioinformatics analysis of our raw exome data using the NextGene® desktop software produced by SoftGenetics®. NextGene® was designed to enable researchers to quickly assess genetic variants from NGS data, without additional bioinformatics support. This stage of my project was completed prior to receiving variant calls from MUGQIC, mainly to evaluate the ease of use and applicability of this platform for use in exome analysis in our laboratory.

In spring 2014, MUGQIC uploaded the raw exome data for this project to the Nanuq external server. I was able to access and download these .fastq files, for the forward and reverse direction of each exome, to a desktop computer in the Woods laboratory. Files in the .fastq format can be directly uploaded to NextGene®, where they are subsequently converted to .fasta files. The forward and reverse files for each exome were then aligned to the human reference genome version GRCh37 (hg19). Following alignment, the software offers several options for quality control, such as selection of an optimal sequencing coverage cut-off. The default parameters were chosen for my analysis, to assess the basic functioning of the software. The coverage option was set to >10x for each read, and reads were required to be at least 50 base pairs in length. The software recognizes discrepancies between the exome and reference genome, which are called as variants.

With regards to variant annotation, NextGene® has a Track Manager tool, which enables users to import information from different variant databases such as dbSNP, and

1000 Genomes. This information includes the rs ID number for previously reported variants, the overall minor allele frequency, and frequencies specific to different populations. Annotations from dbNSFP 2.0 (Liu et al., 2013) can also be imported, including scores retrieved from SIFT (Kumar et al., 2009), Polyphen2 (Adzhubei et al., 2010), LRT (Chun & Fay, 2009), MutationTaster (Schwarz et al., 2010), MutationAssessor (Reva, Antipin, & Sander, 2011) and FATHMM (Shihab et al., 2013). Values for PhyloP (Siepel & Haussler, 2004), GERP (Cooper et al., 2005) and SiPhy (Garber et al., 2009) are also accessible by NextGene®. The variant list for a single exome may be queried through the use of these various tracks. For example, variants with Polyphen2 scores in the "probably damaging" or "possibly damaging" categories can be selected, which effectively eliminates scores that are predicted to be benign. Each predictive algorithm category can be filtered in this manner, to tailor a variant list to reflect a specific study design.

The most relevant application provided by NextGene® for family-based studies is the variant comparison tool (VCT). This tool allows a user to compare variant calls and annotations from multiple exomes. The VCT also provides a filter to view heterozygous or homozygous variants exclusively, as well as the option of specifying family relationships and affection status. For my project, I used this tool to compare the five exomes from R1256 with two unrelated control exomes selected from other projects in the Woods Laboratory. One exome was from a familial colorectal cancer patient (CRC:RD218), and the other was from an idiopathic pulmonary fibrosis patient (IPF:Z37). I was able to cross-reference variants that were found in all exomes from family R1256, but not in these controls. Both heterozygous and homozygous variants

were queried, to reflect the unknown inheritance pattern of IA in this family. All variants with an unreported MAF or value less than 1% were kept in the analysis. I also used the VCT to compare all seven exomes from R1352 with these two control exomes. As a second option, I compared the exomes of the four siblings (Z1039, Z1040, Z1497, Z1508) from R1352 with the selected controls. These closely related individuals were selected due to potential genetic heterogeneity within the extended family. As well, this option allowed me to visualize variants shared by four, five or six relatives, which provided a greater dataset for this large family. The resultant variant lists were reviewed to locate any genes previously described in IA association studies. As well, these lists were compared to the group of candidates that passed my initial variant filtering strategy. Any overlap would provide a level of increased confidence in the variant calls made by MUGQIC.

2.6 Additional Variant Prioritization

Following the filtering strategy outlined in section 2.4, a limited number of candidate variants remained for each family. From this point forward, I only considered moderate and high impact variants for further analysis. During the filtering stage of my project, the article "Lessons learned from whole exome sequencing in multiplex families affected by a complex genetic disorder, intracranial aneurysm" by Farlow et al. (2015) was released. This study summarized preliminary data from a whole exome sequencing project by the Familial Intracranial Aneurysm Study investigators. As mentioned previously, they used WES to identify a list of 68 variants of interest across seven families, which are potentially involved in IA risk. Due to the timely release of this

article and relevance to my research, I decided to compare their list of 68 variants to my lists of filtered moderate and high impact variants.

Lastly, the number of moderate impact variants for R1256 was still high after filtering, so further analysis was necessary to achieve a manageable number for the scope of my thesis project. Moderate impact variants that were found in 5/5 exomes, and were unreported in dbSNP and ExAC, were selected for further consideration in this study. Additional variants from this family may be validated as part of future research.

2.7 Validation and Segregation Analyses

In order to verify that the moderate and high impact candidate variants identified by WES were not false-positives, Sanger sequencing was used for the validation stage of this project. Additional affected individuals from R1256 were included as well (Table 9). This step allowed me to assess whether a variant was present in all or most relatives with IA. Any variants that did not appear in the majority of affected individuals were discarded from further analysis.

After a number of top candidates were identified, a selection of unaffected relatives were also sequenced, to provide a better picture of the possible mode of inheritance and penetrance. To further assess top candidates, Polyphen2, SIFT and GERP scores were also evaluated as a prediction for pathogenicity, as they were readily available for most non-synonymous variants. Polyphen2 is an algorithm that is used to predict the impact of an amino acid substitution on the overall protein, and its function and structure (Adzhubei et al., 2010). Scores can range from 0-1, or "benign" to "probably damaging". Intermediate scores on this scale may be classified as "possibly

Table 9. Affected individuals who were Sanger sequenced in each family.

R1352	R1256
Z1039	Z929
Z1040	Z1013[+]
Z1497	Z1390[+]
Z1507	Z1405
Z1508	Z1406
Z1533	Z1441
Z1651	Z1448
Z1496[+]	Z1459[+]
Z1495[+]	Z1471[+]
	Z1501[+]
	Z1522[+]

[+]These individuals were not used in WES analysis, but were included in the validation stage of the project, when prioritized variants were sequenced. DNA for Z1496 and Z1495, in family R1352, did not produce successful Sanger sequencing results for the majority of tested variants, due to low quality DNA extraction.

damaging". Sorting Intolerant From Tolerant (SIFT) provides a similar prediction, as scores from 0-0.05 are considered to be predictive of "damaging" or deleterious substitutions. Scores above 0.05 to 1 are classified as "tolerated" or benign (Kumar et al., 2009). Genome Evolutionary Rate Profiling (GERP) scores can range from -12.3 to 6.17, with scores greater than 3.00 being indicative of evolutionary conservation of an amino acid, as a measure of functional constraint (Cooper et al., 2005).

2.8 Sanger Sequencing Protocol

Multiple steps are involved in the execution of a successful Sanger sequencing experiment. The following standard protocols are used by the Woods Laboratory, and were consistently followed throughout this project.

2.8.1 Polymerase Chain Reaction

Prior to PCR, highly concentrated genomic DNA samples were diluted to aliquots of approximately 50 ng/µl. Each individual reaction contained a mixture of 1.5 µl of 10x PCR reaction buffer, 0.375 µl of deoxyribose nucleotide triphosphates (dNTPs) at 100ng/µl, 0.5 µl of forward primer at 10 µM, 0.5 µl of reverse primer at 10 µM, 0.75 µl of MgCl at 50 Mm, 0.15 µl of Platinum Taq Polymerase, 10.225 µl of distilled H_2O, and 1 µl of genomic DNA. A negative control, which excludes genomic DNA, was included in each PCR procedure. Master PCR mixes were prepared in 0.5 or 1.5 ml microcentrifuge tubes, and individual reactions were pipetted into 96-well PCR plates. Plates were covered with plastic PCR caps, vortexed, and centrifuged for 20 seconds. Plates were then placed in either an Eppendorf Mastercycler® or Biometra® T1 thermocycler, which

were programmed to a specific protocol. Prior to amplification of patient DNA, protocols were optimized for each primer set, using control DNA. Protocol design was based on the melting temperature of primers, and experimentation with different known procedures. Primer sequences and corresponding thermocycler protocols are described in Appendix B.

Following amplification, PCR reactions were verified using agarose gel electrophoresis. Agarose gels were prepared using a mixture of 50 ml of 1x TAE buffer and 1 g of Ultrapure agarose in an Erlenmeyer flask. This mixture was then heated in a microwave for 75 seconds. Subsequently, 3.75 µl of SYBR® Safe DNA Gel Stain was added to the mixture to allow the visualization of the gel under UV light. The gel mixture was poured into a mold, and combs were placed to create wells for PCR products. To prepare for electrophoresis, 3.5 µl of 5x loading dye was mixed with 3 µl of each PCR product, and these samples were then pipetted into individual wells. As a reference point, the first well of each row was filled with 3.5 µl of loading dye and 1 µl of 100 bp DNA ladder. The agarose gel was then placed in a gel chamber and covered with 1x TAE buffer. A power source was connected to the chamber and each gel was run at 120 V for 25-30 minutes. Gels were viewed under UV light using an AlphaImager EP light cabinet, to determine if DNA was successfully amplified, and confirm that reactions were not contaminated.

2.8.2 ExoSAP

The next step in preparation for Sanger sequencing involves the removal of excess dNTPs, using a mixture of exonuclease and shrimp alkaline phosphatase (ExoSAP). Each

reaction consisted of 0.5 µl of exonuclease, 0.5 µl of shrimp alkaline phosphatase, 7.5 µl of distilled H_2O, and 8 µl of PCR product. The master mix of ExoSAP was prepared in a 1.5 ml microcentrifuge tube, over ice. Individual reactions were pipetted into standard PCR plates, and capped. Plates were mixed with a vortex, and centrifuged for 20 seconds. Plates were then placed on a thermocycler, which was programmed to the "ExoSAP" protocol, and is described in detail in Appendix C.

2.8.3 ABI Cycle Sequencing

Following PCR product clean-up with ExoSAP, ABI cycle sequencing reactions were prepared. Each reaction contained 0.5 µl of ABI cycle sequencing mix, 2 µl of 5x sequencing buffer, 0.67 µl of 10 µM primer, and 15.83 µl of distilled H_2O. Two reaction mixtures were created for each PCR product, one using a forward primer, and one using a reverse primer. Each 19 µl mixture, and 4 µl of PCR product, was added to a well of a 96-well sequencing plate. Plates were capped, mixed with a vortex and centrifuged for 30 seconds at 300 rpm. Plates were then placed in a thermocycler, and programmed to the "ABIseq" protocol, as described in Appendix C.

The next step was to complete ethanol precipitation of the cycle sequencing reactions. After removing the plates from the thermocycler, 65 µl of 95% EtOH and 5 µl of 0.125 mM EDTA was pipetted into each well. Plates were re-capped, mixed quickly with a vortex, and centrifuged for another 30 seconds at 300 rpm. Plates were placed in the refrigerator at 4°C for 30 minutes, or alternatively, covered at room temperature, overnight. Next, plates were centrifuged for 30 minutes at 3000 g. Caps were removed from the plates, which were then inverted to remove the ethanol. Plates were inverted and

placed on paper towel in the centrifuge, where they were spun at 300 rpm for 30 seconds to remove any excess ethanol. A multichannel pipette was then used to add 70% EtOH to each well. Plates were capped and centrifuged for 15 minutes at 3000 g. Once again, plates were inverted to remove excess ethanol, and blotted on paper towels. The plates were then allowed to dry for 30 minutes in a drawer at room temperature. Finally, 10 µl of deionized formamide (HiDi) was pipetted into each well. A clean septa was placed onto each plate, which was then vortexed and centrifuged briefly before being placed on a thermocycler set to the "Denature" program (Appendix C).

Finally, plates were loaded onto an ABI 3130 Cycle Sequencer. Each sequencing run, which covers two columns of a 96-well plate, is approximately one hour in length. Following sequencing, result files were uploaded to the Sequencing Analysis desktop software. Sequencing Analysis files were then opened using the Sequencher 5.0 program, which allows visualization of chromatograms and corresponding base calls.

2.9 Population Control Testing

Following Sanger sequencing of prioritized variants, select top candidate variants were Sanger sequenced in a control cohort of 100 individuals representing the Newfoundland and Labrador population. Any variants with a high frequency in NL population controls are unlikely to be causative in this study, as some polymorphisms with a low MAF globally may have an increased prevalence localized to this province. Only 100 controls were used, as we were only interested in determining whether selected candidate variants are common, as opposed to estimating population frequency. The control DNA samples are part of the Newfoundland Colorectal Cancer Registry

(NFCCR) project. They represent residents from throughout the province who had no medical history of cancer, and were recruited through random-digit dialing (P. Wang et al., 2009).

3. Results

3.1 Variant Calls by MUGQIC

3.1.1 Data Summary Following Variant Calling

Following WES and initial bioinformatics analysis, MUGQIC provided a total of six variant lists for the IA cohort, as well as some general statistics regarding sequencing coverage and run metrics for the exomes. For the 12 IA exomes, 40-63 million reads were generated per library. On average, the ratio of surviving reads following trimming to the number of raw reads from the sequencer was 91.85% for the 12 exomes. The mean coverage of the whole genome was 2.8% (total number of aligned reads/size of the genome). Approximately 97% of the bases in the capture region of the exome had a read depth of at least 10x.

As an initial requested filtering step, MUGQIC generated lists that contained only variants with a read depth above 10, across all exomes within a family. The low impact category contained the most variants, with 10,899 low impact variants detected in family R1256, and 9,324 in R1352. These values include variants found in any member of a family. As expected, the lists of moderate and high impact variants were extensive; 8,401 moderate and 339 high impact variants were called in the R1352 exomes. Similarly, 9,814 moderate and 410 high impact variants were called in family R1256. Throughout my results, the archived genome browser GRCh37.p13 from Ensembl was used to fill in any incomplete variant annotations from the MUGQIC variant lists. This measure was taken to avoid error, as the GRCh37 version of the human reference genome was used for read alignment.

3.2 Initial Variant Filtering

3.2.1 Candidate Genes from Previous Studies

Initially, I searched both the moderate and high impact lists from each family for the presence of genes found in or near regions that had been significantly associated with IA, as reviewed by Tromp et al. (2014). A total of 20 genes were identified from known candidate gene and genome-wide association studies of IA, and were summarized in Table 8. My results were interpreted with the knowledge that these studies are reporting statistical association, and not causation between certain SNPs and the IA phenotype. Associated SNPs may be in linkage disequilibrium with a SNP in another gene, or may possibly impact a molecular pathway involved in IA development.

In both families (R1256 and R1352) none of the 20 candidate genes were detected in the high impact variant lists. However, in the moderate impact lists, there were several variants identified within these genes. Notably, 10 missense variants in IA-associated genes were shared between the two families, and had been previously reported in dbSNP (Table 10). Some of these variants were shared by multiple exomes within a single family, whereas others were found in only two or three affected individuals, and are not likely implicated in the IA phenotype. All 10 variants were assessed in the Broad Institute's ExAC Browser for minor allele frequency. The lowest MAF for one of these entries was 20.5%, for the *RRBP1* variant *rs1132274*. Therefore, all of these variants are very common in the general population and were assumed to be non-pathogenic.

Additionally, there were several variants that were detected in only one of the study families. In R1256, 4 missense variants in the *HSPG2* gene were called. One of

these variants, *rs2291827*, had a MAF value above 17% in the ExAC Browser. The other three variants included: *rs139500146* (MAF=1.07%), *rs114851469* (MAF=1.10%) and *rs116788687* (MAF=1.76%), all of which had frequencies below 2%. These variants were only detected in 1 or 2 exomes out of 5, and were unlikely to be causative of IA. A missense variant in *TNFRSF13B* (*rs34562254*) was identified in 1 exome from R1256, and had a MAF of 14.2%. Finally, all 5 exomes were heterozygous for *rs3625* in the *JDP2* gene, which had a MAF value of 53% in the ExAC Browser.

In R1352, 7 missense variants and 1 inframe insertion were detected, which were not shared by the R1256 exomes. Two different missense variants in *HSPG2* were identified: *rs17459097* (MAF=4.2%) and *rs146309392* (MAF=0.07948%). Despite its low population frequency, *rs146309392* was not pursued as a candidate variant as it was only present in 1/7 exomes. Two common variants in *TCN2*, 1 in *FGD6*, and 1 in *VCAN* were also detected, but were also only present in a single family member. Three family members also shared a heterozygous inframe insertion in *SOX17*, which has not been previously reported in dbSNP.

Finally, 4/7 exomes were heterozygous for *rs72553883* in the *TNFRSF13B* locus, which has an allele frequency of 0.536%, and was found in 648/120,888 alleles from the ExAC Browser cohort. This variant could be explored further in future studies, though it does not segregate completely in this family. Functionally, *TNFRSF13B* has a potential role in immunity (Tromp et al., 2014). Four SNPs near or within this gene were significantly associated with IA in a Chr 17-specific exploration of IA cases and controls by Inoue et al. (2006). None of the variants in IA-associated genes were categorized as rare and shared by most exomes within a family, and thus were not pursued as part of this

Table 10. Variants in IA-associated genes detected in families R1256 and R1352.

Gene	Variant Details (*rs#* from dbSNP)	Family (# Exomes with Variant)	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in ExAC Browser (alleles)
COL1A2	c.1645C>G, p.P549A (*rs42524*)	R1352 (7/7) R1256 (5/5)	R1352: 1 / 6 / 0 R1256: 4 / 1 / 0	76.9% (93340/121334)
COL3A1	c.2092G>A, p.A698T (*rs1800255*)	R1352 (4/7) R1256 (2/5)	R1352: 2 / 2 / 3 R1256: 2 / 0 / 3	32% (22129/69064)
ELN	c.1222G>A, p.G408S (*rs2071307*)	R1352 (5/7) R1256 (2/5)	R1352: 5 / 0 / 2 R1256: 2 / 0 / 3	32.6% (39600/121386)
RRBP1	c.2672G>T, p.R891L (*rs1132274*)	R1352 (4/7) R1256 (4/5)	R1352: 4 / 0 / 3 R1256: 4 / 0 / 1	20.5% (23922/116400)
	c.1829T>A, p.L610H (*rs6034867*)	R1352 (7/7) R1256 (5/5)	R1352: 0 / 7 / 0 R1256: 0 / 5 / 0	*99.9% (86841/86848)
STARD13	c.45C>G, p.N15K (*rs876133*)	R1352 (3/7) R1256 (1/5)	R1352: 2 / 1 / 4 R1256: 1 / 0 / 4	44.1% (5427/12302)
VCAN	c.5477G>A, p.R1826H (*rs188703*)	R1352 (6/7) R1256 (5/5)	R1352: 5 / 1 / 1 R1256: 3 / 2 / 0	38.1% (45930/120534)
	c.1045A>G, p.K349E (*rs61749613*)	R1352 (1/7)	1 / 0 / 6	3.2% (3771/119722)
TCN2	c.67A>G, p.I23V (*rs9606756*)	R1352 (1/7)	0 / 1 / 6	11.7% (14139/121344)
	c.962C>T, p.S321F (*rs9621049*)	R1352 (1/7)	1 / 0 / 6	11.2% (13582/121384)
TNFRSF13B	c.542C>A, p.A181E (*rs72553883*)	R1352 (4/7)	4 / 0 / 3	0.54% (648/120888)
	c.752C>T, p.P251L (*rs34562254*)	R1256 (1/7)	1 / 0 / 4	14.2% (16736/118038)
FGD6	c.2255G>T, p.R752L (*rs117209224*)	R1352 (1/7)	1 / 0 / 6	0.80% (966/121370)
SOX17	c.68_69insGCACCA p.Q1325_H1326insQH (*Unreported*)	R1352 (3/7)	3 / 0 / 4	N/A

HSPG2	c.10918G>A, p.V3640I (*rs17459097*)	R1352 (2/7)	2 / 0 / 5	4.2% (5134/120852)
	c.7235G>A, p.S2412N (*rs146309392*)	R1352 (1/7)	1 / 0 / 6	0.08% (77/96878)
	c.9790A>G, p.I3264V (*rs139500146*)	R1256 (1/5)	1 / 0 / 4	1.08% (1297/120388)
	c.9766C>T, p.H3256Y (*rs2291827*)	R1256 (2/5)	2 / 0 / 3	17.3% (20740/120026)
	c.8929C>T, p.R2977W (*rs114851469*)	R1256 (2/5)	2 / 0 / 3	1.10% (1068/97498)
HSPG2, continued	c.6114C>G, p.I2038M (*rs116788687*)	R1256 (1/5)	1 / 0 / 4	1.76% (2061/117092)
	c.4508C>T, p.A1503V (*rs897471*)	R1352 (7/7) R1256 (5/5)	R1352: 0 / 7 / 0 R1256: 0 / 5 / 0	74.6% (68707/92062)
	c.2294A>C, p.N765S (*rs989994*)	R1352 (7/7) R1256 (5/5)	R1352: 0 / 7 / 0 R1256: 0 / 5 / 0	*98.11% (119030/121326)
	c.1912A>G, p.M638V (*rs1874792*)	R1352 (7/7) R1256 (5/5)	R1352: 0 / 7 / 0 R1256: 0 / 5 / 0	*99.3% (96055/96720)
JDP2	c.37A>G, p.T13A (*rs3625*)	R1256 (5/5)	5 / 0 / 0	53.0% (60318/113860)

In the fourth column, the values in brackets indicate how many affected individuals have each variant (out of a possible five exomes in R1256, and seven exomes in R1352).

*For *rs6034867*, *rs989994*, and *rs1874792*, the reference assembly contains the very rare minor allele at these loci, which was then labeled as the reference allele. As a result, the MAF values at these sites are extremely high in the general population.

thesis. As more association studies are published, the list of IA candidate genes will be updated.

3.2.2 Data Summary Following Filtering of Whole Exome Variants

After searching for variants in genes previously connected to IA, the direction of my project was steered toward novel genetic factors for IA predisposition within the whole exome. The step-wise implementation of the filtering strategy outlined in Figure 10 significantly reduced the number of variants for each family, providing a more manageable dataset. The number of variants remaining after each step is summarized in Table 11.

Table 11. Number of variants remaining following filtering of variant lists.

	R1352: Low	R1352: Moderate	R1352: High	R1256: Low	R1256: Moderate	R1256: High
All variants with read depth >10x	9,324	8,401	339	10,899	9,814	410
Variants shared by ≥ 6/7 family members (R1352) or ≥ 4/5 (R1256)	3501	2,988	202	5219	4,551	211
Variants found in <1% of 848 MUGQIC control exomes	8	15	2	56	106	3
Variants with MAF unreported or <1% in dbSNP	8	15	2	52	92	2
Variants with MAF unreported or <1% in additional databases (ExAC Browser; Ensembl Genome Browser; NHLBI Exome Variant Server)	5	5	1	41	66	2

3.2.3 High Impact Variants: Family R1352

As expected, few high impact variants passed the selected filters used in my study. Variants categorized as "high impact" have a greater predicted effect on DNA sequence and overall protein assembly. For family R1352, a single variant remained that was present in 6/7 exomes at a read depth above 10, was detected in fewer than 1% of internal exome controls from MUGQIC, and had a MAF less than 1% in public variant databases. This variant in the *C4orf6* gene results in the loss of the start codon, and has been previously reported in dbSNP under the identification number *rs144117694* (Table 12).

3.2.4 Moderate Impact Variants: Family R1352

Five moderate impact variants remained in the dataset following filtering in R1352. Four of these variants were classified as missense, and the fifth was determined to be an inframe insertion (Table 13). Previously reported variants in *ATP1A4*, c.1798C>T, and *GIGYF2*, c.3494A>G, were present in 6/7 exomes. Additionally, the *RP1L1* variant c.202C>T was detected in 6 exomes, and appears to be novel in this family. The *MUC16* variant c.40588G>A and the *HSPBP1* insertion c.78_79insGGCGGCGGA were called in all 7 exomes, and were not previously reported in variant databases. All 5 of these variants required further consideration through Sanger sequencing validation.

Table 12. High impact variant in family R1352 that passed filtering steps.

Gene	Chr. Position	Status in dbSNP (*rs#)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
C4orf6	4p16.2	Reported (*rs144117694*)	c.1A>G; p.M1V Start-lost	6 / 0 / 1	0% (0/848)	0% (0/21384)

Table 13. Moderate impact variants in family R1352 that passed filtering steps.

Gene	Chr. Position	Status in dbSNP (*rs#)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
ATP1A4	1q23.2	Reported (*rs142338502*)	c.1798C>T; p.P600S Missense	5 / 1 / 1	0.59% (5/848)	0.2002% (246/122948)
GIGYF2	2q37.1	Reported (*rs72554081*)	c.3494A>G; p.H1165R Missense	6 / 0 / 1	0.59% (5/848)	0.1586% (195/122946)
RP1L1	8p23.1	Unreported	c.202C>T; p.L68F Missense	6 / 0 / 1	0% (0/848)	N/A
MUC16	19p13.2	Unreported	c.40588G>A; p.G13530S Missense	7 / 0 / 0	0% (0/848)	N/A
HSPBP1	19q13.42	Unreported	c.78_79insGGCGGCGGA; p.G25_G26insAAD Inframe Insertion	7 / 0 / 0	0.12% (1/848)	N/A

Column 5 of each table lists the number of exomes from family R1352 that were heterozygous (Het) for each given variant, followed by how many were homozygous (Hom) for the minor allele, and the number that were homozygous for the reference allele. For the MUGQIC MAF values, percentages are given based on 848 control exomes. For ExAC Browser MAF values, the fractions in brackets indicate the number of minor alleles in the dataset over the number of available sequenced alleles for this locus.

3.2.5 Low Impact Variants: Family R1352

In family R1352, 5 synonymous variants remained after filtering (Table 14). Three of these synonymous changes are in the same gene – *MUC16*. This gene was previously mentioned, as a missense variant in *MUC16*, c.40588G>A, also passed filtering. One of the synonymous *MUC16* variants was previously reported in dbSNP, but was detected at an extremely low frequency in the ExAC Browser (0.000828%). The other two variants were found in all 7 exomes, and are apparently novel. Additionally, synonymous changes in *COL6A3*, c.702C>T, and *APBB2*, c.231G>A passed the filtering strategy. Both variants were called as heterozygous in 6/7 exomes, and were detected at low minor allele frequencies in exome control populations.

Table 14. Low impact variants that passed filtering strategy in family R1352.

Gene	Chr. Position	Status in dbSNP (*rs #)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC controls (/848)	MAF in ExAC Browser (/alleles)
COL6A3	2q37	*rs142876356*	c.702C>T p.L234	6 / 0 / 1	0.12% (1/848)	0.00505% (6/118902)
APBB2	4p13	*rs116718314*	c.231G>A p.A77	6 / 0 / 1	0.24% (2/848)	0.1441% (174/120764)
MUC16	19p13.2	*rs111468097*	c.41742C>A p.P13914	6 / 0 / 1	0% (0/848)	0.000828% (1/120766)
		Unreported	c.40587T>C p.S13529	7 / 0 / 0	0% (0/848)	N/A
		Unreported	c.37254G>A p.R12418	7 / 0 / 0	0% (0/848)	N/A

Abbreviations: Het = heterozygous, Hom = homozygous

3.2.6 High Impact Variants: Family R1256

Two high impact variants passed the implemented filtering strategy in R1256 (Table 15). A previously reported splice-site donor in the *OCIAD1* gene was detected, as well as an unreported single-base deletion in *CCDC3*, leading to a frameshift. Both variants were detected in 4/5 exomes from the family and are heterozygous.

3.2.7 Moderate Impact Variants: Family R1256

Following the application of my initial filtering criteria, 66 moderate impact variants still remained for family R1256. To reduce this list of candidates to a more manageable number, I uploaded the list of 66 variants to the online Database for Annotation, Visualization and Integrated Discovery (DAVID) program, to determine whether any of these genes had functional relevance to IA pathogenesis (Huang da, Sherman, & Lempicki, 2009a; Huang da, Sherman, & Lempicki, 2009b). DAVID provides gene ontology and pathway keywords to annotate gene lists. Results of this endeavor were inconclusive, as no gene was a clear functional candidate for IA pathophysiology. Next, I decided to focus on variants that were found in all 5 exomes, and had not been previously reported with an identification number in dbSNP. This elimination step involves the assumption that there are no phenocopies among these 5 sequenced individuals, and that any variants of interest are novel. This reduced the list of candidate variants to 15 (Table 16). In this list of 15, the *MUC16* and *HSPBP1* variants previously reported in family R1352 were present.

Table 15. High impact variants in family R1256 that passed filtering steps.

Gene	Chr. Position	Status in dbSNP (*rs#)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
OCIAD1	4p11	Reported (*rs144048911*)	c.-6+1G>A Splice Site Donor	4 / 0 / 1	0% (0/848)	0.02851% (5/17538)
CCDC3	10p13	Unreported	c.425delA; p.Tyr267ThrfsTer21 Frameshift	4 / 0 / 1	0.35% (3/848)	0.2677% (290/108350)

Abbreviations: Het = heterozygous, Hom = homozygous

Table 16. Moderate impact variants in family R1256 that passed filtering criteria, were detected in 5/5 exomes, and were unreported in dbSNP.

Gene	Chr. Position	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
POU3F1	1p34.1	c.82_84delGCG; p.A28del Inframe Deletion	5 / 0 / 0	0.94% (8/848)	N/A
MUC4	3q29	c.7165G>A; p.D2389N Missense	5 / 0 / 0	0% (0/848)	N/A
MUC4	3q29	c.7166A>G; p.D2389G Missense	5 / 0 / 0	0% (0/848)	N/A
MUC4	3q29	c.7168G>C; p.A2390P Missense	5 / 0 / 0	0% (0/848)	N/A
MUC4	3q29	c.2989G>A; p.D997N Missense	5 / 0 / 0	0.12% (1/848)	N/A
DSPP	4q22.1	c.2053_2054insGTAGC AGTGACAGCAGCA; p.N685_S686insSSSDSS; Inframe Insertion	3 / 2 / 0	0% (0/848)	N/A
MAML3	4q31.1	c.1468_1470delCAG; p.Q490del; Inframe Deletion	0 / 5 / 0	0.83% (7/848)	N/A
DOPEY1	6q14.1	c.6902C>T; p.A2301V Missense	5 / 0 / 0	0% (0/848)	N/A
CCDC3	10p13	c.217C>G; p.L73V Missense	5 / 0 / 0	0.35% (3/848)	N/A
KNDC1	10q26.3	c.2686G>T; p.A896S Missense	5 / 0 / 0	0% (0/848)	N/A
MTG1	10q26.3	c.611C>T; p.P204L Missense	5 / 0 / 0	0% (0/848)	N/A
SPDYE4	17p13.1	c.103C>T; p.P35S Missense	5 / 0 / 0	0.35% (3/848)	N/A
ZPBP2	17q12	c.622A>T; p.T208S	5 / 0 / 0	0.12% (1/848)	N/A

		Missense			
MUC16	19p13.2	c.40588G>A; p.G13530S Missense	5 / 0 / 0	0% (0/848)	N/A
HSPBP1	19q13.42	c.78_79insGGCGGCGG; p.G25_G26insAAD Inframe Insertion	5 / 0 / 0	0.12% (1/848)	N/A

Abbreviations: Het = heterozygous, Hom = homozygous

Since *MUC16* c.40588G>A and *HSPBP1* c.78_79insGGCGGCGGA were present in all 5 exomes from family R1256 and all 7 exomes from R1352, these variants were of great interest. Both variants are also localized to Chr 19. In particular, the shared inframe insertion in *HSPBP1* is notable, as 3 amino acids are inserted into the genomic sequence in affected individuals.

The additional 51 moderate impact variants for family R1256 are described in detail in Appendix D. Thirteen of these variants were detected in 5/5 exomes and were previously reported in dbSNP, 27 were detected in 4/5 exomes and were previously reported in dbSNP, and 11 were detected in 4/5 exomes and were unreported in dbSNP. Extensive prioritization and validation of top candidates from this list of 51 variants will be pursued at a later date.

Within my variant lists, there was a high degree of disagreement between the predictive scores produced by in silico algorithms such as Polyphen2 and SIFT. This prevented me from further eliminating variants based on these scores alone. Functional predictive scores were noted for use in prioritizing variants later in the study, and for use as evidence to support candidate genes and their pathogenicity. Due to the manageable number of variants that passed my filtering strategy, I was able to avoid further filtering my data based on predicted functional relevance at this juncture. Therefore, all variants described in Tables 12, 13, 15 and 16 were Sanger sequenced to determine if they were true variants or false-positives.

3.2.8 Low Impact Variants: Family R1256

My filtering strategy was also applied to the low impact variant list from family R1256, and 41 synonymous variants successfully passed filtering. The 11 variants that were detected in all 5 exomes from R1256 are highlighted in Table 17. The 30 additional synonymous variants present in 4/5 exomes are described in Appendix E. The 3 *MUC16* variants from R1352 were also found in family R1256, along with 2 other unreported variants in this gene, c.39099C>T and c.390996C>A. This gene would not be prioritized in any hypothetical further analyses, due to the highly polymorphic nature of this locus. An unreported change in *HLA-DRB1* would also be removed from further study, as it was called in an alternate haplotype of this Chr 6 region, and this gene is known to be a highly polymorphic "super SNP" gene in the general population (Ju et al., 2010). "Super SNP" genes are enriched with non-synonymous variation, and may not be relevant to IA pathogenesis. Ju et al. (2010) state that most of their classified "super SNP" genes are involved in sensory and immune function, such as olfactory receptor and *HLA* genes.

Table 17. Low impact variants that passed filtering strategy, and were detected in all 5 exomes from family R1256.

Gene	Chr. Position	Status in dbSNP (*rs #)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC controls (/848)	MAF in ExAC Browser (/alleles)
TNRC18	7p22.1	Unreported	c.3807C>T p.P1269	5 / 0 / 0	0% (0/848)	*0.008% (1/12503)
OTOGL	12q21.31	*rs145834039*	c.3486C>T p.G1162	3 / 2 / 0	0.35% (3/848)	0.2827% (341/120626)
PCDH9	13q21.32	*rs146618643*	c.60C>T p.S20	4 / 1 / 0	0.35% (3/848)	0.1431% (170/118834)
FAM83G	17p11.2	Unreported	c.2208A>G p.P736	5 / 0 / 0	0% (0/848)	N/A
GPR179	17q12	*rs202228440*	c.5868C>T p.S1956	5 / 0 / 0	0.35% (3/848)	0.07619% (92/120746)
MUC16	19p13.2	*rs111468097*	c.41742C>A p.P13914	5 / 0 / 0	0% (0/848)	0.000828% (1/120766)
		Unreported	c.40587T>C p.S13529	5 / 0 / 0	0% (0/848)	N/A
		Unreported	c.39099C>T p.S13033	5 / 0 / 0	0% (0/848)	N/A
		Unreported	c.39096C>A p.G13032	5 / 0 / 0	0% (0/848)	N/A
		Unreported	c.37254G>A p.R12418	5 / 0 / 0	0% (0/848)	N/A
HLA-DRB1	[+]6_ssto_hap7	Unreported	p.R99	0 / 5 / 0	0% (0/848)	N/A

Abbreviations: Het = heterozygous, Hom = homozygous

*MAF was not reported in ExAC Browser, but was present in NHLBI Exome Variant Server.
[+]6_ssto_hap7 is an alternate haplotype for chromosome 6.

3.3 NextGene® Results

3.3.1 Family R1352 Variants

Through the use of the Variant Comparison Tool (VCT), it was determined that 7 homozygous missense and 51 heterozygous missense variants were shared by the 7 exomes from family R1352 and were absent from control exomes IPF:Z37 and CRC:RD218. No high impact variants were detected with this tool, and all 7 homozygous missense variants had dbSNP MAF values above 15%. There was no overlap between the list of heterozygous missense variants with my filtered MUGQIC data. Therefore, the *MUC16* p.G13530S and *HSPBP1* p.G25_G26insAAD variants were not detected through NextGene®. Functional annotation with GeneCards and DAVID did not yield any promising hits for IA-related keywords in this list of 51 variants. A list of keywords was generated containing words and phrases that appear frequently in descriptions of the IA phenotype and pathophysiology in published articles (Appendix F). To expand the scope of the data output and possibly find novel or rare variants not detected by the MUGQIC pipeline, I then used the VCT to compare the exomes of siblings Z1039, Z1040, Z1497 and Z1508 to the 2 controls. Though the additional relatives share a phenotype (Z1507, Z1533, Z1651), they are more distantly related and could be phenocopies in this family. This was a generous filtering method, as variants shared by 4 or more exomes out of 7 are included in these results. The objective was to identify variants in functionally relevant genes that were omitted following stringent filtering.

3.3.2 Family R1352 Variants Shared by 4 Siblings

A total of 50 homozygous missense and 618 heterozygous missense variants were shared by Z1039, Z1040, Z1497 and Z1508, but were not found in the control exomes. By using the dbSNP MAF values provided by NextGene®, these values were reduced to 2 homozygous and 141 heterozygous missense variations with an unreported MAF or value less than 1%. This significant drop shows the importance of considering allele frequency when searching for disease-related variants. It is still highly likely that some of these 143 variants are common, as many of the corresponding dbSNP MAF values were unreported, and would have to be imputed manually.

Within this list of 143 variants, the 3 remaining missense calls from Table 13 were detected: *ATP1A4* c.1798C>T, *GIGYF2* c.3494A>G, and *RP1L1* c.202C>T. Each of these variants was called in 6/7 exomes by NextGene®. *ATP1A4* c.1798C>T and *GIGYF2* c.3494A>G were absent from individual Z1651, and *RP1L1* c.202C>T was absent from Z1533, which matched the MUGQIC results. The successful replication of these calls provides increased confidence and evidence to support the bioinformatics pipeline used by MUGQIC. It is interesting to note that several variants called by NextGene® were not actually discovered through MUGQIC variant calling. Discrepancy across different analysis pipelines is a known issue in NGS research, which can be partly ameliorated by using multiple pipelines to analyze raw exome reads.

Finally, the list of 143 variants was uploaded to DAVID to search for any genes with possible IA-related functionality. A rare variant in the *COL6A3* gene, c.5610C>A, was highlighted, which has been previously reported in dbSNP (*rs113153193*). This

variant has a MAF of 0.128% in the ExAC Browser, as it has been detected in 154/120,552 alleles. The COL6A3, collagen type 6 alpha 4, protein was noted in DAVID for its role in collagen biosynthesis and cell adhesion. This variant was found in 6/7 exomes by NextGene®, but was not called by MUGQIC. A second gene of interest was recognized through DAVID, the adenylate cyclase 1 (brain) gene, *ADCY1*. The *ADCY1* variant c.1837G>A was only called in the 4 siblings (4/7 exomes), and has been previously reported in dbSNP as *rs752410249*. In the ExAC Browser, this variant has a low MAF of 0.0033%, and has only been detected in 4/121412 alleles. Endothelin-1/EDNRA signaling was identified as a pathway related to *ADCY1* function in GeneCards and DAVID. Endothelin-1 is known to be involved in the dilation and constriction of blood vessels, and *EDNRA* has been previously identified as an IA candidate gene (Yasuno et al., 2011). *ADCY1* may also be involved in brain development and regulatory processes in the central nervous system. This *ADCY1* variant, c.1837G>A, was also not present in the MUGQIC variant lists. Validation of these variants will be pursued in the future.

3.3.3 Family R1256 Variants

In family R1256, 18 homozygous missense and 223 heterozygous missense variants were shared by the 5 exomes, and were not present in either control exome. The majority of the detected, shared variants have high MAF values reported in dbSNP, and are not of interest in this project. Two heterozygous nonsense variants were also shared by the 5 exomes and were absent in controls, though both were reported at high frequencies in dbSNP. Unfortunately, with this version of NextGene®, I was unable to

detect other high impact variant types including splice site and frameshift variants with accuracy. Therefore, the *OCIAD1* and *CCDC3* high impact variants detected by MUGQIC could not be replicated in this dataset. After reviewing the list of 241 missense variants, it was evident that there was a significant degree of overlap with my filtered variants that were called and annotated by MUGQIC. From the list of 15 variants in Table 16, seven calls were replicated with NextGene®. These variants included *MUC4* c.2989G>A, *DOPEY1* c.6902C>T, *CCDC3* c.217C>G, *KNDC1*, c.2686G>T, *MTG1* c.611C>T, *SPDYE4* c.103C>T, and *ZPBP2* c.622A>T. It is possible that some of the remaining 8 variants were also called by NextGene®, but were excluded by the use of 2 control exomes. Given that scenario, the NextGene® software is effectively helping to eliminate variants that are generally common or are prevalent in the NL population.

Across both families, there was a high degree of confirmation from NextGene® to support my filtered variant lists. Therefore, it can be concluded that NextGene® could be a helpful secondary analysis tool to provide additional evidence of variant discovery.

3.4 Homozygous Variation in Siblings from Family R1352

As stated previously, the majority of variants that passed the filtering strategy were heterozygous in affected participants. However, there were several incidences where the classification of homozygosity and heterozygosity was split for a given variant. For example, R1256 family members Z1405, Z1441, and Z1448 were homozygous for a c.3704C>T missense variant in *MYO18A*, while Z929 and Z1406 were heterozygous. However, there were many homozygous variants present in each family that were not shared by all or most members. In family R1352, there was a large subset of variants that

were shared by the main group of siblings (Z1039, Z1040, Z1497, and Z1508). For many of these variants, there was a split between exomes that were homozygous and heterozygous for the minor allele. Given the strong evidence for familial predisposition to IA in R1352, and the lack of a strong candidate from my filtering methodology, it was determined that investigation into homozygous variation in these four siblings would be worthwhile.

Table 18 summarizes 14 variants that were shared by at least the 4 siblings (Z1039, Z1040, Z1497 and Z1508), were called as homozygous in at least 1 exome, and have a MAF value less than 5% in internal exome controls and the ExAC Browser. The 5% cutoff was selected to see what effect increasing the MAF limit would have on the generated results. Almost all of the 14 variants had a MAF above 1%, which would have eliminated them from my previous filtering strategy. *ATP1A4* c.1798C>T was previously analyzed, as it passed the original filtering strategy for this project. *ZFPM1* c.1334_1339delCTCCGG also passed initial filtering, but was not prioritized. At the onset of this project, the *ZFPM1* variant was mis-classified by dbSNP and Ensembl as a common variant. The deletion was classified as a common variant, while the reference allele was classified as rare in public datasets. After discovering this discrepancy, this variant was highlighted as a site of interest, as it was shared by all exomes in both families R1256 and R1352. Sanger sequencing was completed to determine if this variant was a true positive.

Table 18. Variants shared by exomes Z1039, Z1040, Z1497 and Z1508 that are homozygous in at least one exome, and have a MAF less than 5%.

Gene	Chr. Position	Status in dbSNP (*rs #)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC controls (/848)	MAF in ExAC Browser (/alleles)
**ATP1A4*	1q23.2	*rs142338502*	c.1798C>T, p.P600S	5 (Z1508,1497,1039,1533,1040) / 1 (Z1507) / 1 (Z1651)	0.59% (5/848)	0.2002% (246/122948)
PTPRC	1q31.3	*rs41269905*	c.367G>C, p.D123H	4 (Z1533, 1651, 1508, 1039) / 1 (1507) / 2 (Z1040, Z1497)	2.48% (21/848)	1.39% (1684/121386)
		rs2230606	c.403A>C, p.I135L	4 (Z1533, 1651, 1508, 1039) / 1 (1507) / 2 (Z1040, 1497)	2.83% (24/848)	1.59% (1929/121330)
LY75	2q24.2	*rs78446341*	c.3740C>T, p.P1247L	4 (Z1497, 1507, 1533, 1039) / 2 (Z1508, 1040) / 1 (Z1651)	3.77% (32/848)	1.99% (2417/121340)
STX2	12q24.33	*rs137928907*	c.94T>G, p.F32V	3 (Z1507, 1039, 1497) / 2 (Z1508, 1040) / 2 (Z1533, 1651)	4.83% (41/848)	1.45% (1759/121032)
SEZ6L2	16p11.2	*rs113753753*	c.1210G>A, p.D404N	4 (Z1533, 1039, 1497, 1508) / 1 (Z1507) / 2 (Z1040, 1651)	3.66% (31/848)	1.69% (2001/118422)
**ZFPM1*	16q24.2	*rs149145771*	c.1334_1339 delCTCTGG p.L446_A447del Inframe Deletion	0 / 7 / 0	0.35% (3/848)	0% (0/118)
MUC16	19p13.2	*rs183524392*	c.9359C>A, p.T3120N	5 (Z1651, 1508, 1497, 1039, 1533) / 1 (Z1040) / 1 (Z1507)	1.77% (15/848)	0.91% (1098/120754)

Abbreviations: Het = heterozygous, Hom = homozygous

*These variants also passed the original filtering strategy from Figure 11.

3.5 Validation and Segregation Analysis

Following validation by Sanger sequencing, it was discovered that there were no shared high or moderate impact variants between the 2 families in our study. Sanger sequencing of *MUC16* c.40588G>A and *HSPBP1* c.78_79insGGCGGCGGA revealed that these genetic variants were not actually present in any of the 7 family members from R1352, nor the 5 affected members from R1256. The *MUC16* and *HSPBP1* variants were classified as false-positive calls, which is not uncommon in exome sequencing studies. This result was also foreshadowed by their absence from the NextGene variant reports.

As well, the 2 families did not share any prioritized variants in the same gene. This is consistent with genetic heterogeneity in familial IA. It is possible that these 2 families have completely different loci that are contributing to IA risk. The following sub-sections provide a summary of the Sanger sequencing results for exome-sequenced individuals and for any additional affected relatives.

3.5.1 High Impact Variants: Family R1352

Sanger sequencing confirmed the presence of the *C4orf6* variant in 6 of 7 exomes (Table 19). I was able to sequence this region in an additional affected family member, Z1496, who is the father of several study participants (Z1039, Z1040, Z1497, Z1508). The *C4orf6* variant was not detected in this individual, though sequencing chromatograms for this DNA sample were distorted and difficult to interpret with certainty. The DNA extraction process did not yield high quality product for study participant Z1496 and his spouse, Z1495. Sequencing of these samples produces poor

sequence quality and often, failed PCR. Therefore, Z1496 was not used in the segregation analysis of moderate impact variants from this family.

At the onset of this study, *C4orf6*, also known as Chromosome 4 open reading frame 6, was reported as encoding an uncharacterized protein product. More recently, *C4orf6* has been categorized as a long intergenic non-protein coding RNA gene in GeneCards. The lack of functional information associated with this gene makes it difficult to confidently create a connection to the pathophysiology of IA. In order to get a better idea of the potential involvement of c.1A>G in IA development, DNA from a selection of unaffected relatives was Sanger sequenced in R1352 (Figure 12). This test group consisted of 14 unaffected individuals. These relatives were considered to be informative and relevant, due to their presence in the 2 generations of the pedigree that contain all known affected individuals. Clinical information and DNA are available for family members in a younger generation, which is not shown in the condensed pedigree for R1352. Individuals in this generation are currently outside of the normal age range for IA development, as IA is a late-onset disorder. The current age range of these individuals is 14-36, with a mean age of 26.9 in the year 2015. It is quite possible that some of these younger relatives will develop IA in their lifetime, and variant analysis at this point would not contribute greatly toward our understanding of IA mode of inheritance and etiology.

It was determined that 5/14 unaffected relatives were carriers of this variant. Seven asymptomatic relatives did not have the variant, as they were homozygous for the reference allele at this locus. Individuals Z1537 and Z1509, who have been diagnosed with abdominal aortic aneurysm, also did not have the *C4orf6* variant. The use of in silico

predictive tools for this site provided conflicting results. This variant had a reported GERP score of 0.225, which predicts that the methionine at this site is not highly conserved. *C4orf6* c.1A>G was predicted to be benign by Polyphen2, but was predicted to be deleterious by SIFT, with a score of 0.

Though *C4orf6* c.1A>G exhibits incomplete segregation in this family, it remains the top candidate due to its classification as *high impact*. Therefore, this variant was chosen to be Sanger sequenced in a cohort of population controls from Newfoundland and Labrador. Sanger sequencing of the *C4orf6* c.1A>G variant was completed in 100 randomly selected control samples from the Newfoundland and Labrador Colorectal Cancer Registry (NFCCR) cohort. Analysis of the DNA sequence chromatograms showed that this variant was not present in any member of the control set (Figure 13). Therefore, it is assumed that the *C4orf6* c.1A>G variant is not common in the NL population.

3.5.2 Moderate Impact Variants: Family R1352

Validation by Sanger sequencing was successful for the *ATP1A4, GIGYF2* and *RP1L1* missense variants (Table 20). It was confirmed that each of these variants was present in 6/7 exomes from family R1352. In order to assess the predicted pathogenicity of these three variants, the scores from GERP, Polyphen2 and SIFT were recorded and evaluated. *ATP1A4* c.1798C>T had a significant GERP score of 4.19, and was predicted to be damaging by both Polyphen2 and SIFT. *GIGYF2* c.3494A>G had a highly conserved GERP score of 5.43 out of a possible 6. It was predicted to be damaging by Polyphen2, but benign by SIFT, with a score of 0.19 out of 1. *RP1L1* c.202C>T yielded

similar results, with a conserved GERP score of 4.19. This variant was also predicted as damaging by both Polyphen2 and SIFT. To assess their relevance to IA, a PubMed search was conducted for all three genes. Other databases including UniProt, OMIM and GeneCards were also explored to determine gene function.

ATP1A4 encodes the sodium/potassium-transporting ATPase subunit alpha-4 protein, which is involved in the creation of an electrochemical gradient across the plasma membrane. This protein is also known to have a role in sperm motility. *RP1L1*, or retinitis pigmentosa 1-like 1, has a role to play in photoreceptor cell differentiation in the human eye. This gene has been previously associated with occult macular dystrophy in Japanese families (Akahori et al., 2010). Finally, *GIGYF2* encodes the GRB10-interacting GYF protein 2. This protein has been categorized by gene ontology terms including "negative regulation of translation" and "post-embryonic development", and has fairly ubiquitous expression throughout the body. Heterozygous variants in the *GIGYF2* gene have been associated with autosomal dominant Parkinson disease type 11, which is a neuromuscular disorder (Lautier et al., 2008). This gene is also believed to play a role in tyrosine kinase receptor signaling, which is essential to the regulation of cellular processes in the body.

In order to distinguish between these 3 variants, a number of unaffected family members were Sanger sequenced. Neither variant was absent from all unaffected relatives, though incomplete penetrance remains a distinct possibility. *ATP1A4* c.1798C>T was present in 4/11 unaffected relatives, including one family member with AAA (Figure 14). However, the other family member with AAA did not have the variant. *GIGYF2* c.3494A>G was present in 5/14 unaffected relatives, including 1 individual with

AAA (Figure 15). Finally, *RP1L1* c.202C>T was present in 4/13 unaffected relatives, but was absent from both family members with AAA (Figure 16). Therefore, the results of unaffected Sanger sequencing did not provide a clear visualization of what moderate impact variant(s) should be prioritized. Functionally, *GIGYF2* appeared to have the most potential out of these three genes, and was sequenced in 100 population controls from the NFCCR cohort. This variant was absent from all 100 controls, which suggests that this substitution is not common in the NL population (Figure 17).

Table 19. Sanger sequencing results for filtered high impact variants from family R1352.

Gene	Variant Details	# of Sequenced Affected Individuals with Variant[Δ]	Affected Individuals with Variant	Affected Individuals without Variant
C4orf6	c.1A>G; p.M1V Start-lost	6/8	Z1039, Z1040, Z1497, Z1507, Z1508, Z1651	Z1533, Z1496

Table 20. Sanger sequencing results for filtered moderate impact variants from family R1352.

Gene	Variant Details	# of Sequenced Affected Individuals with Variant[Δ]	Affected Individuals with Variant	Affected Individuals without Variant
ATP1A4	c.1798C>T; p.P600S Missense	6/7	Z1039, Z1040, Z1497, Z1507 (homozygous), Z1508, Z1533	Z1651
GIGYF2	c.3494A>G; p.H1165R Missense	6/7	Z1039, Z1040, Z1497, Z1507, Z1508, Z1533	Z1651
RP1L1	c.202C>T; p.L68F Missense	6/7	Z1039, Z1040, Z1497, Z1507, Z1508, Z1651	Z1533

[Δ] WES results were consistent with results obtained by Sanger sequencing.

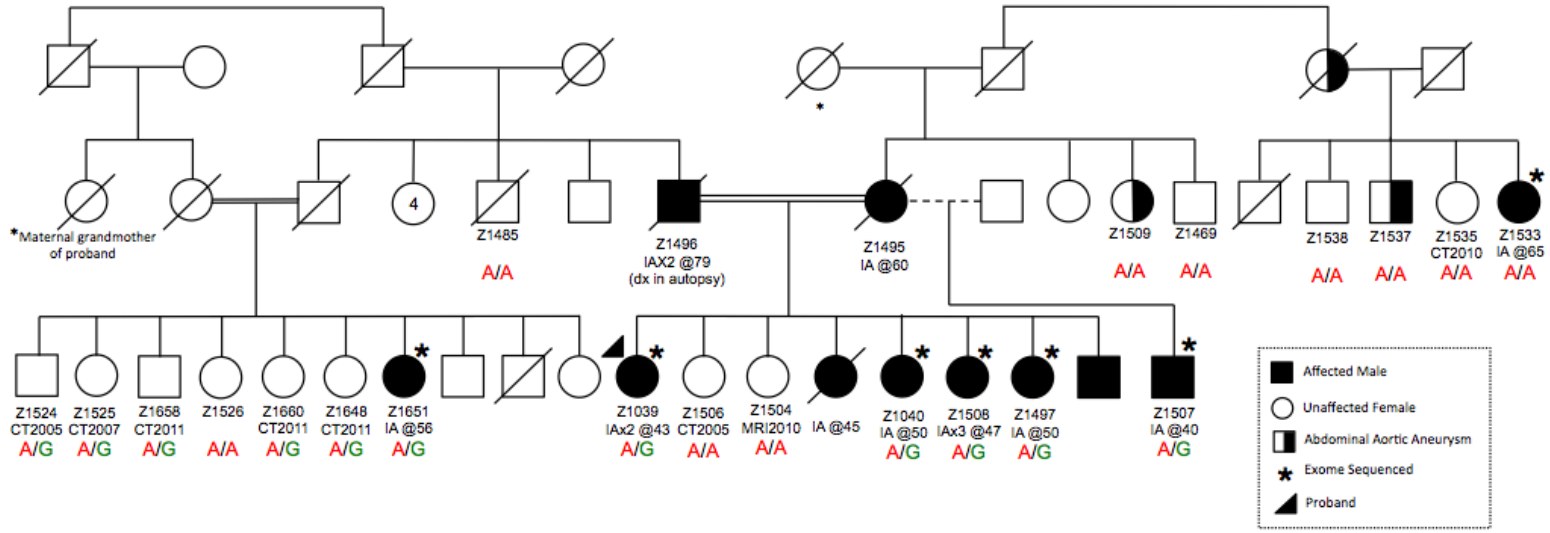
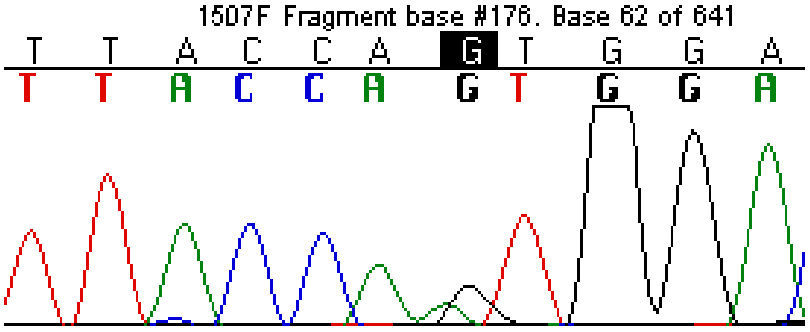


Figure 12. Segregation of *C4orf6*, c.1A>G in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. For unaffected individuals, the date of their last known negative CT scan is indicated where available.

(A)



(B)

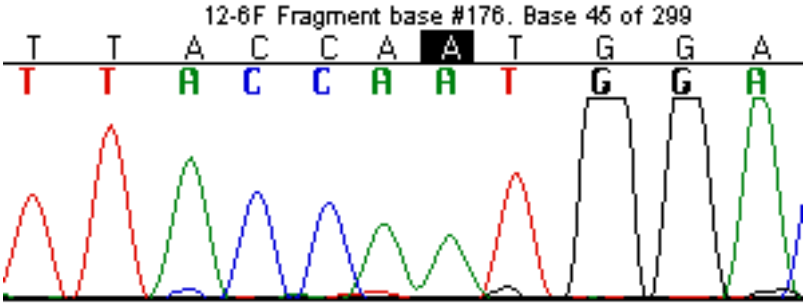


Figure 13. Chromatogram of Sanger sequencing results for *C4orf6* c.1A>G.

The top image (A) shows confirmation of the heterozygous start-lost variant in an affected member of R1352. The bottom image (B) shows results for a control individual who is homozygous for the reference allele.

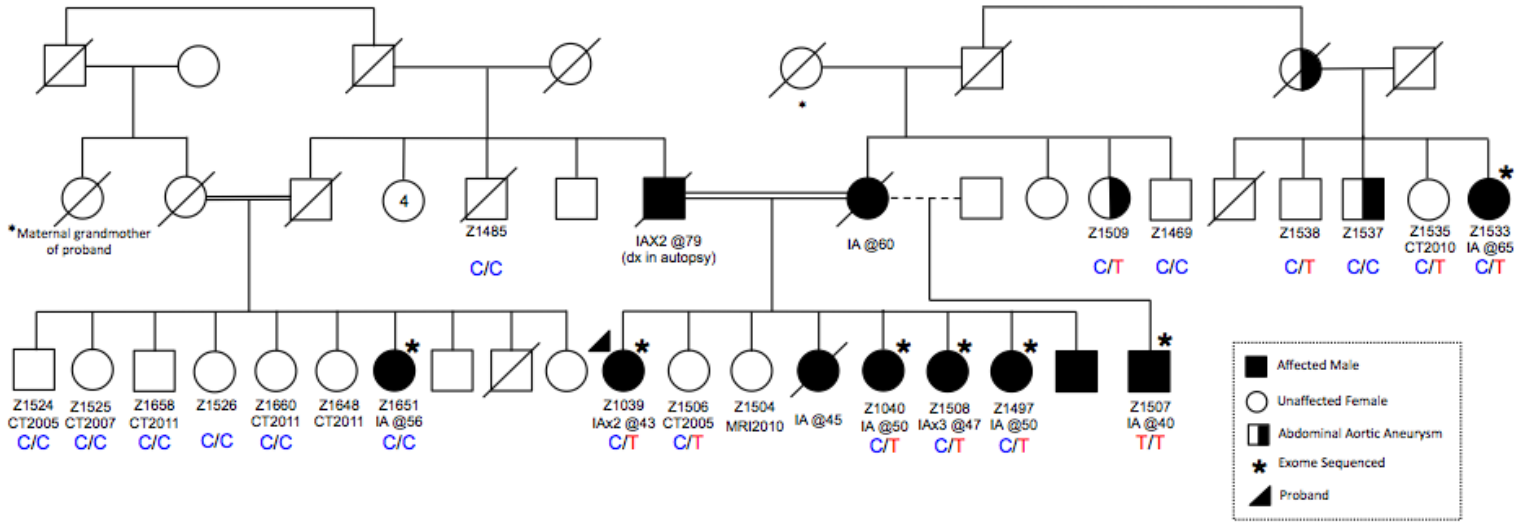


Figure 14. Segregation of *ATP1A4*, c.1798C>T in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. For unaffected individuals, the date of their last known negative CT scan is indicated where available.

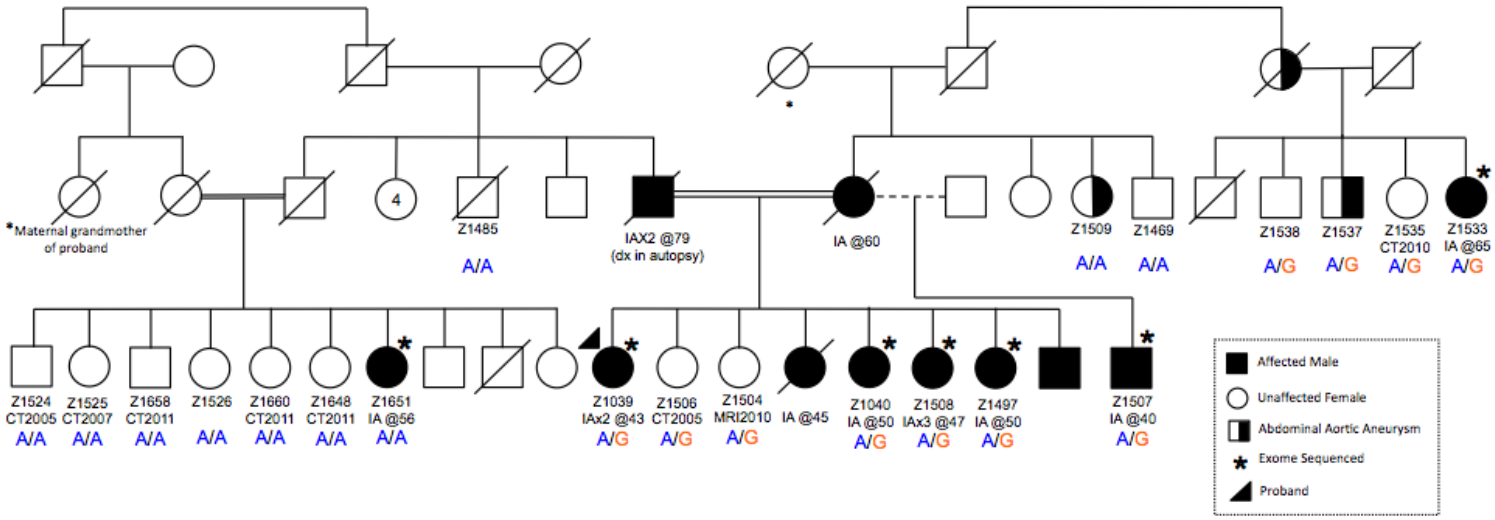


Figure 15. Segregation of *GIGYF2* c.3494A>G in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. For unaffected individuals, the date of their last known negative CT scan is indicated where available.

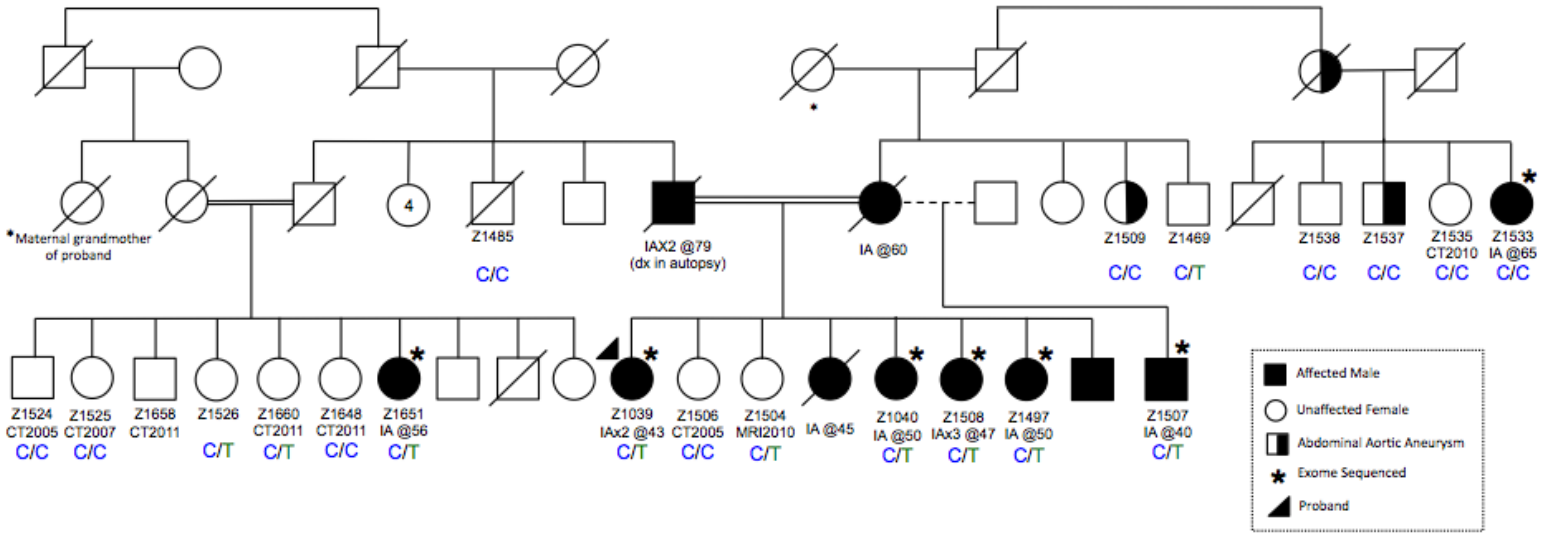
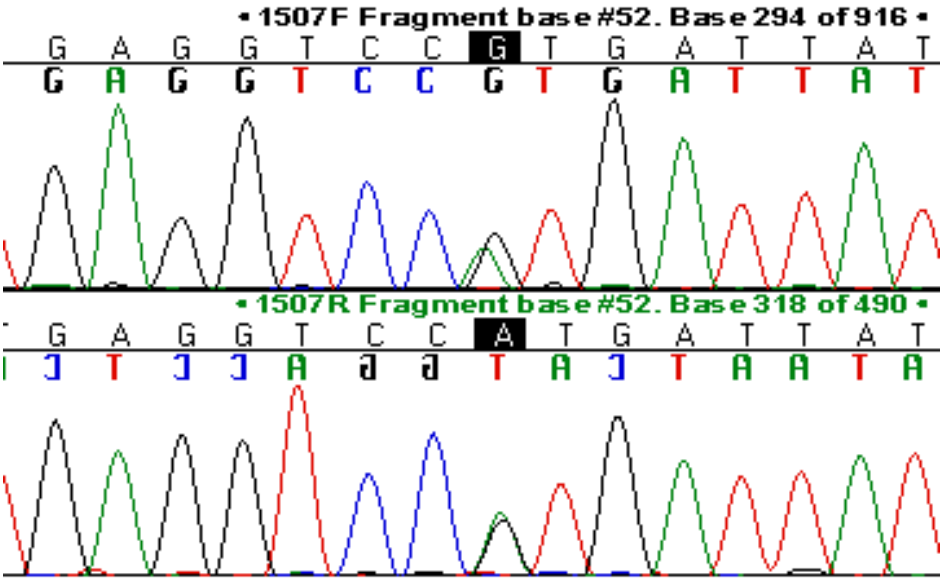


Figure 16. Segregation of *RP1L1*, c.202C>T in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. For unaffected individuals, the date of their last known negative CT scan is indicated where available.

(A)



(B)

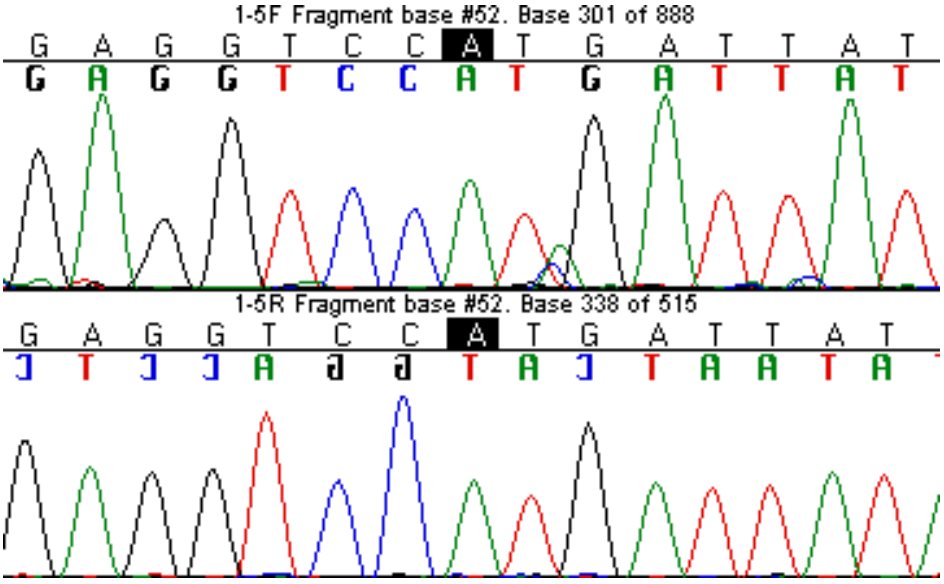


Figure 17. Chromatogram of Sanger sequencing results for *GIGYF2* c.3494A>G.

The top image (A) shows confirmation of the heterozygous variant in an affected member of R1352, in both the forward and reverse sequences. The bottom image (B) shows the absence of a variant at this position in an unaffected control sample.

3.5.3 Homozygous ZFPM1 Variant: Family R1352

Sanger sequencing of the *ZFPM1* c.1334_1339delCTCCGG variant was unsuccessful, as PCR ventures did not yield any product. Based on previous experience, it is possible that this deletion was a false positive, as it was shared by all exomes in our cohort. As well, the *ZFPM1* locus was only covered by 118 alleles in the ExAC browser. On the ExAC website, a warning is issued at the top of some entries, to indicate that a site is covered by less than 80% of the individuals in the database. This warning states that these entries may be low quality sites. Though none of the 14 variants in Table 18 is a clear functional candidate for IA, these lists show that different methods of thinking about exome data and filtering in the context of complex disease should be explored.

3.5.4 High Impact Variants: Family R1256

Sanger sequencing confirmed that the *CCDC3* and *OCIAD1* high impact variants were present in 6 of 7 exomes (Appendix G; Table 21). *CCDC3* encodes the coiled-coil domain containing 3 protein. Azad et al. (2014) determined that *CCDC3* was expressed in adipose tissues and endothelial cells, including those in the aortic wall. They were able to determine that *CCDC3* is involved in the repression of TNF-α induced VCAM-1 (vascular cell adhesion molecule 1) expression. TNF-α activity is believed to activate nuclear factor kB (NF-kB) signaling, and induce an inflammatory response in endothelial cells. Therefore, the secretory protein CCDC3 has a possible role in the regulation of inflammation in this pathway. *OCIAD1* encodes the ovarian carcinoma immunoreactive antigen domain-containing protein 1. This protein has wide expression throughout the body, and one of its isoforms is strictly expressed in the brain and nervous system (Luo,

Soosaipillai, & Diamandis, 2001). Functional annotation for this gene is limited, though it has been reported as over-expressed in ovarian cancer tissues through the use of a mouse model (Sengupta, Michener, Escobar, Belinson, & Ganapathi, 2008).

Additional affected members of R1256 were then sequenced, to see how these variants segregate with the disease throughout the entire affected family. *CCDC3* c.425delA was absent from 5 affected family members, and *OCIAD1* c.-6+1G>A was absent from 2 individuals in total (Table 21). The transmission of these variants is shown on the R1256 pedigree in Appendix G. The absence of these variants from multiple affected family members meant that they were not prioritized for further study at this time.

Table 21. Sanger sequencing results for filtered high impact variants from family R1256.

Gene	Variant Details	# of Sequenced Affected Individuals with Variant[Δ]	Affected Individuals with Variant	Affected Individuals without Variant
OCIAD1	c.-6+1G>A Splice Site Donor	6/8	Z929, Z1405, Z1406, Z1448, Z1459, Z1501	Z1390, Z1441
CCDC3	c.425delA; p.Tyr267ThrfsTer21 Frameshift	5/10	Z1390, Z1405, Z1406, Z1441, Z1448	Z929, Z1013, Z1459, Z1501, Z1471

[Δ]WES results were consistent with results obtained by Sanger sequencing.

3.5.5 Moderate Impact Variants: Family R1256

Sanger sequencing of the 15 novel, shared moderate impact variants revealed that the *MUC16* p.G13530S and *HSPBP1* p.G25_G26insAAD variants were false-positive calls, as discovered in R1352. The *MUC4* c.2989G>A variant was also classified as a false-positive, as it was not detected in exomes Z1405, Z1406 or Z1441. The sequencing chromatograms for Z929 and Z1448 were messy, and re-sequencing yielded similar results. A cluster of three variants (p.D2389N; p.D2389G; and p.A2390P) in *MUC4* were not validated, due to the highly polymorphic nature of this region. Acceptable primers could not be designed for this site with the Primer3 program, to meet the minimal accepted standards for successful use in PCR. Regardless, this area of *MUC4* contains a number of deletions, and many repetitive sequences, which would cause non-specific binding of any primers. Sanger sequencing of the *DSPP*, *POU3F1* and *MAML3* variants was also unsuccessful. Despite the use of various thermocycler protocols, and different attempts at primer optimization, PCR did not yield any product for these sites. None of these variants were confirmed by NextGene®, and thus they may be false-positive calls as well.

Validation of the 6 remaining variants was successful, and all 6 had been confirmed through NextGene®. Following validation, additional affected relatives were sequenced for each variant. *DOPEY1* c.6902C>T, *CCDC3* c.217C>G, and *KNDC1* c.2686G>T were all confirmed in 7/10 affected individuals in R1256 (Table 22). *MTG1* c.611C>T was only confirmed in 3/10 individuals, which may mean that read depth or overall sequencing quality was relatively low for exomes Z929 and Z1448 at this site of

the genome. An additional affected family member, Z1522, was only successfully sequenced for 2 variants, due to low DNA sample quality. This increased the pool of affected family members to 11. Thus, *ZPBP2* c.622A>T was confirmed in 8/11 affected family members. Finally, *SPDYE4* c.103C>T was validated in 10/11 affected members of R1256. This variant had the highest degree of segregation with IA, compared to the other 5 candidates. As a result, a cohort of unaffected relatives was Sanger sequenced for the *SPDYE4* c.103C>T site (Figure 18). Pedigrees depicting the segregation of the other R1256 validated variants are located in Appendix G. A chart outlining the gene function and predictive scores for these 5 variants is also located in Appendix H.

SPDYE4 c.103C>T was present in 8/14 unaffected relatives from R1256. As stated previously, it is possible that some of these family members may develop IA in their lifetime, or that incomplete penetrance is a factor. The dates at which these individuals received their last CT scan or MRI is indicated on the pedigree, and some individuals have not been screened in almost 10 years. It is possible that follow-up appointments could change the overall look of this pedigree, and result in the diagnosis and treatment of additional affected individuals. The *SPDYE4* gene encodes the Speedy protein E, and was classified by the gene ontology term "regulation of protein kinase activity". A PubMed search for this gene revealed no results, as it has not been previously connected to any human disease.

With regards to predictive scores, *SPDYE4* c.103C>T has a GERP score of 2.65, which is just below the cut-off for significant evolutionary conservation. It was predicted as benign by Polyphen2, but damaging by SIFT (0.02). The disagreement between these scores does not provide us with a clear indication of this variant's relevance to IA

pathogenesis. Further functional analysis would be necessary to better understand the role of *SPDYE4* in familial IA.

Consequently, Sanger sequencing of the *SPDYE4* c.103C>T variant was completed in 100 randomly selected control samples from the NFCCR cohort. The variant was not present in any of these individuals, as demonstrated by the example chromatogram in Figure 19. The absence of *SPDYE4* c.103C>T from NL controls suggests that this variant is not found in the local population at an increased frequency compared to global datasets. This variant is a promising top candidate in family R1256, based on my stringent filtering criteria. It is also possible that an additional candidate may be present among the other moderate impact variants that are highlighted in Appendix D. In the coming stages of this study, in silico predictive programs may be used to prioritize these variants, along with functional annotation from DAVID and other avenues.

Table 22. Sanger sequencing results for filtered moderate impact variants from family R1256.

Gene	Variant Details	# of Sequenced Affected Individuals with Variant[Δ]	Affected Individuals with Variant	Affected Individuals without Variant
DOPEY1	c.6902C>T; p.A2301V Missense	7/10	Z929, Z1390, Z1405, Z1406, Z1441, Z1448, Z1459	Z1013, Z1471, Z1501
CCDC3	c.217C>G; p.L73V Missense	7/10	Z929, Z1390, Z1405, Z1406, Z1441, Z1448, Z1471	Z1013, Z1459, Z1501
KNDC1	c.2686G>T; p.A896S Missense	7/10	Z929, Z1013, Z1405, Z1406, Z1441, Z1448, Z1471	Z1390, Z1459, Z1501
MTG1	c.611C>T; p.P204L Missense	3/10	Z929, Z1013, Z1448	Z1390, Z1405, Z1406, Z1441, Z1459, Z1471, Z1501
SPDYE4	c.103C>T; p.P35S Missense	10/11	Z929, Z1013, Z1390, Z1405, Z1406, Z1441, Z1448, Z1459, Z1471, Z1501	Z1522
ZPBP2	c.622A>T; p.T208S Missense	8/11	Z929, Z1013, Z1390, Z1405, Z1406, Z1441, Z1448, Z1501	Z1459, Z1471, Z1522

[Δ]WES results were consistent with results obtained by Sanger sequencing.

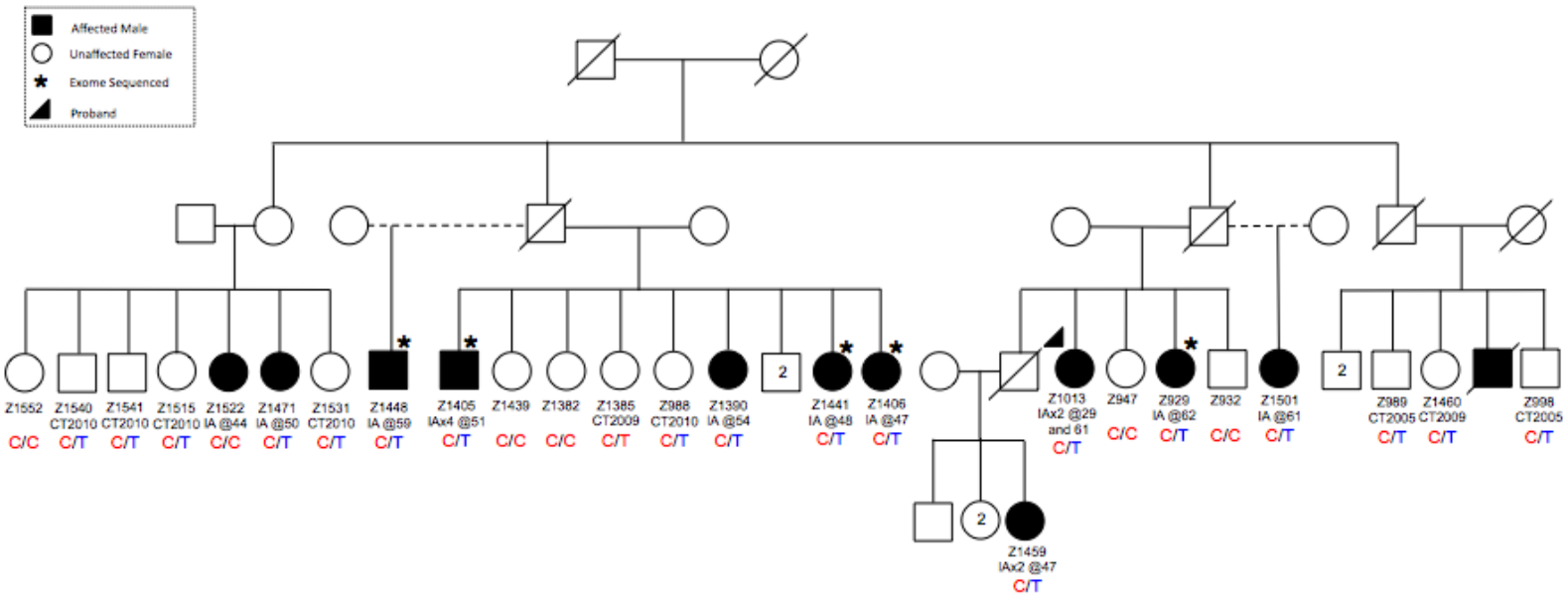
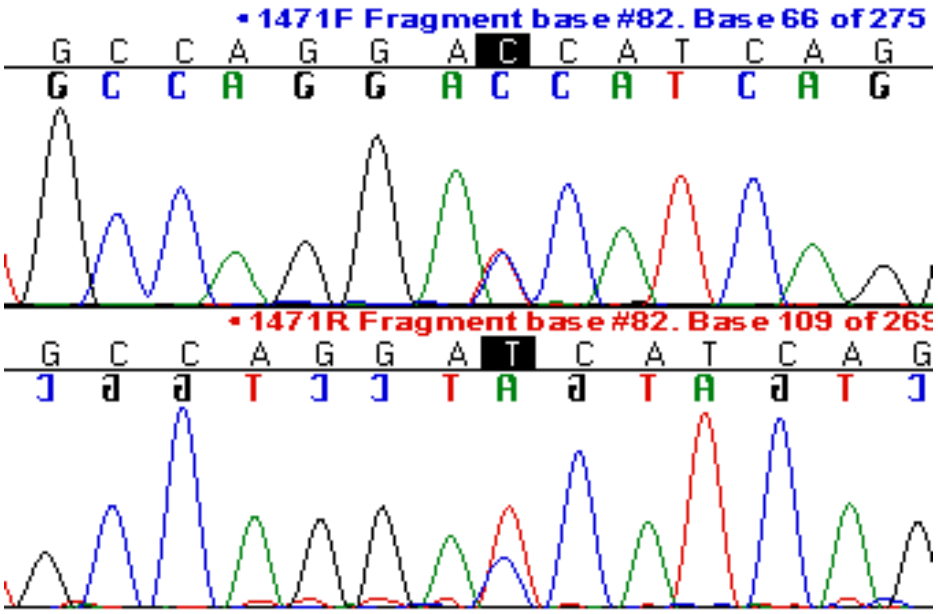


Figure 18. Segregation of *SPDYE4 c.103C>T* in family R1256.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *SPDYE4* c.103C>T variant is present in 10/11 sequenced affected individuals. For unaffected individuals with the variant, the date of their last known negative CT scan is indicated. It is likely that there is decreased penetrance of IA in this family.

(A)



(B)

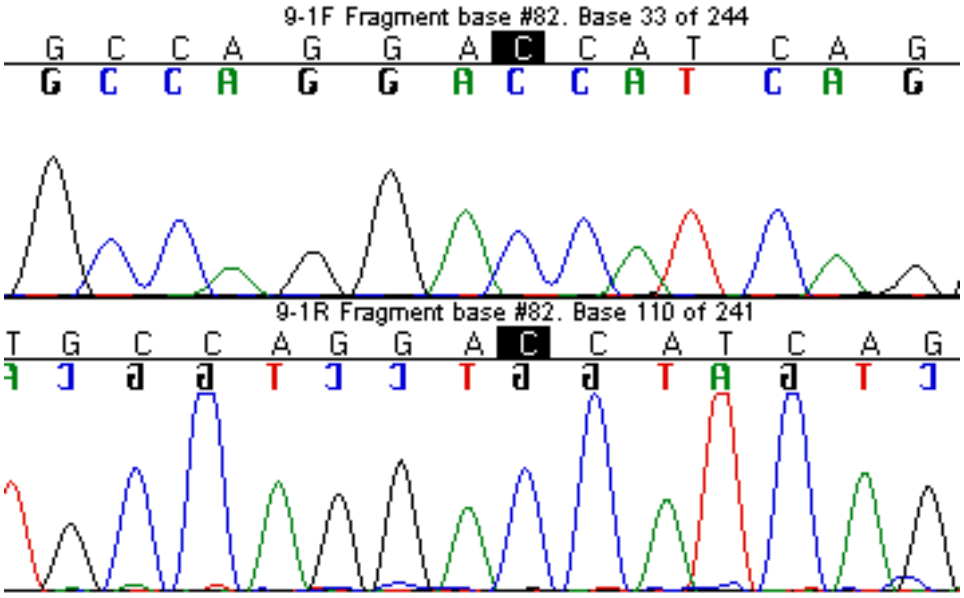


Figure 19. Chromatogram of Sanger sequencing results for *SPDYE4* c.103C>T.

The top image (A) shows confirmation of the heterozygous variant in an affected member of R1256, in both the forward and reverse sequences. Image (B) shows the absence of the variant in a population control individual who is homozygous for the reference allele.

3.5.6 Comparison to Farlow et al., 2015 Study

In reference to section 2.7, I compared my lists of filtered variants to those summarized in a recent study by Farlow et al. (2015). They identified 68 candidate variants from an exome sequencing study of 7 international families affected by non-syndromic IA. IA cases were diagnosed by surgical reports, autopsy, or non-invasive imaging. A group of 3 neurologists reviewed the medical records to ensure that all criteria were met for inclusion of these individuals in the study. None of these 68 variants overlapped with my filtered moderate and high impact lists. However, the gene *TRPA1* appeared in both their list of candidates, and my list of filtered moderate impact variants for family R1256. Farlow et al. (2015) reported a missense variant c.2059A>T in *TRPA1*, which encodes transient receptor potential cation channel, subfamily A, member 1 protein. This variant was detected in only 1 of their multiplex families, and segregated with IA incidence. *TRPA1* c.2059A>T also appears to be novel, as it has an allele frequency of 0% in 1000 Genomes and the NHLBI Exome Variant Server.

In my study, the missense variant c.1309G>A was detected in 4/5 exomes, as it was absent from sample Z1441. This missense change is predicted to be rare or novel, as it has not been previously reported in dbSNP or other exome variant databases including the ExAC Browser. As this gene was the only commonality between the Familial Intracranial Aneurysm study and our cohort, I completed validation via Sanger sequencing to confirm the presence of the variant. The variant was indeed present in 4/5 family members, and thus additional affected members were tested for segregation. Following segregation analysis, *TRPA1* c.1309G>A was found in 6/10 affected family

members that were successfully sequenced (Figure 20). The incomplete segregation of this variant resulted in its elimination from further analysis at this time.

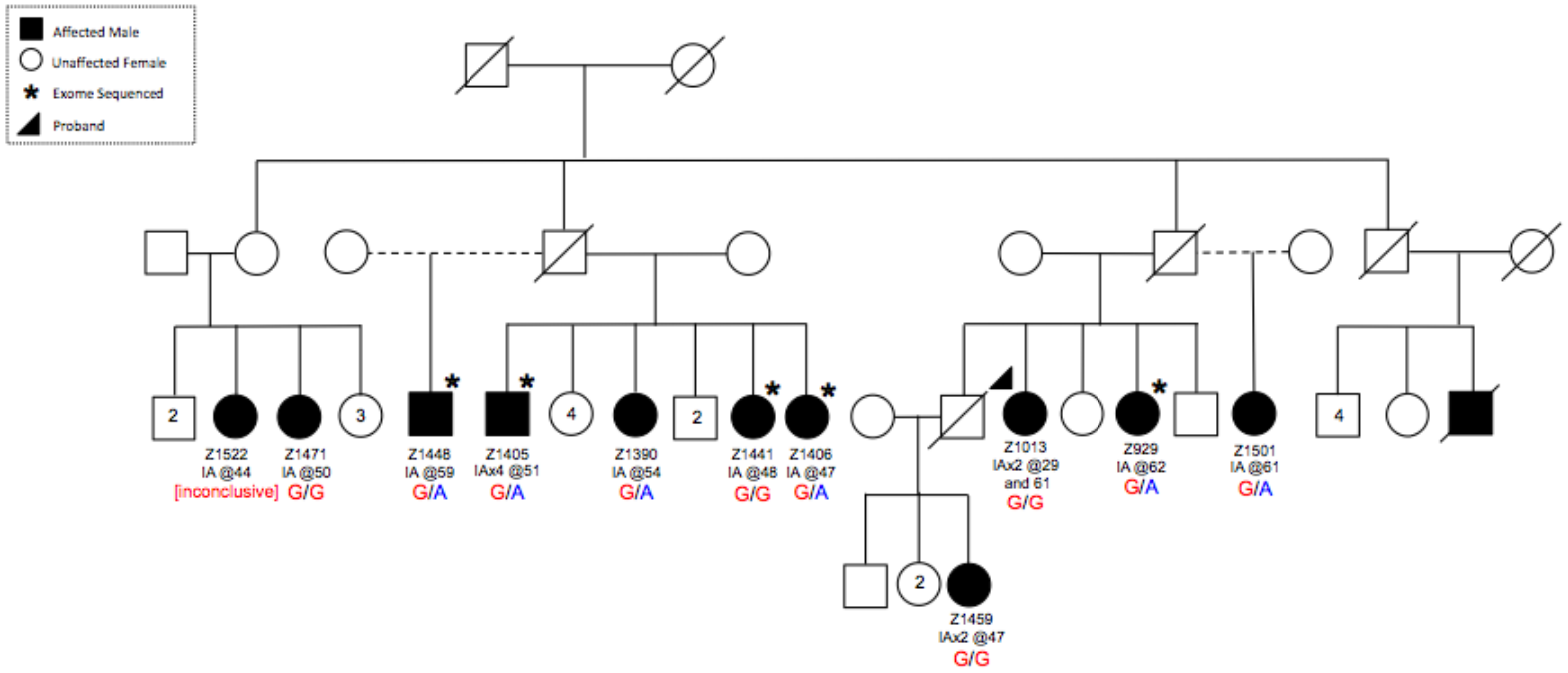


Figure 20. Segregation of *TRPA1* c.1309G>A in affected family members of family R1256.

DNA extraction was completed for individuals with a Z identification number. IA affection status including the patient age at diagnosis and number of aneurysms is indicated. Sequencing results were inconclusive for individual Z1522, due to poor quality DNA.

3.6 Excluded Moderate Impact Variants in Family R1256

Following the main collection of results, it was determined that further exploration of the excluded moderate impact variants in family R1256 was a necessary endeavor. The initial list of 66 filtered, moderate impact variants in R1256 was previously reduced to novel variants that were found in 5/5 exomes. Of the remaining 51 variants (Appendix D), several could be prioritized based on MAF alone. Of the 13 that were detected in 5/5 exomes, and previously reported in dbSNP, a variant in the *MYO18A* gene stands out. The c.3704C>T variant, *rs777985641*, was called as heterozygous in 2 exomes, and homozygous in 3 exomes. This variant has an extremely low MAF of 0.000828% in the ExAC browser, and was only detected in 1/120,720 sequenced alleles. It was determined that Sanger sequencing was worth pursuing to validate this variant. Homozygosity of this variant was confirmed in individuals Z1405, Z1441, and Z1448. Family members Z1406 and Z929 were also confirmed to be heterozygous at this site. Additional affected relatives were also successfully sequenced. Three other relatives (Z1013, Z1390 and Z1501) were heterozygous for *MYO18A* c.3704C>T, and three did not have the variant (Z1459, Z1471 and Z1522). The confirmation of this substitution in 8/11 affected relatives is consistent with the results obtained for *ZPBP2* c.622C>T.

Of the 38 variants that were detected in 4/5 exomes, several were similarly rare in the ExAC browser. In family R1256, it was previously reported that 3 markers in a 22 cM region on chromosome 14 were suggestive of linkage, and the Chr 14q23-31 region has been linked to familial IA in other publications (Ozturk et al., 2006; Mineharu et al., 2008). In the list of 38, there were 2 variants on Chromosome 14 that had been previously

reported in dbSNP (Appendix D). A missense variant in *OR11H4*, c.55G>A, was identified, which is located at chromosomal position 14q11.2. This variant has a MAF of 0.1199% in the ExAC browser. Secondly, a missense variant in *LTBP2*, c.2657C>A, passed the initial filtering strategy. This variant has a MAF of 0.07206% in ExAC, and had been previously detected in 70/97,140 alleles. It has been predicted as benign with Polyphen2, and alternatively, damaging with the SIFT algorithm. *LTBP2* is located at position 14q24.3, which is close to our linked region of interest. This gene encodes the latent transforming growth factor beta binding protein 2, which is an extracellular matrix protein that plays a role in cell structure and adhesion. Based on the Online Mendelian Inheritance in Man (OMIM) database, *LTBP2* has been connected to various eye disorders including congenital forms of glaucoma, but has not been associated with aneurysm formation or other vascular anomalies. Though this variant did not pass our established filtering criteria for family R1256, its potential involvement in IA will be explored in detail in subsequent analyses.

3.7 Digenic Inheritance in Family R1352

The validation of 4 filtered variants in R1352, that were each found in 6/7 exomes, led to the consideration of alternate modes of inheritance. The possibility of digenic inheritance was explored in the context of these 4 candidate variants (*C4orf6* c.1A>G, *GIGYF2* c.3494A>G, *ATP1A4* c.1798C>T, and *RP1L1* c.202C>T). Digenic inheritance is defined as a case where "variant genotypes at two loci explain the phenotypes of some patients and their unaffected (or mildly affected) relatives more clearly than the genotypes at one locus alone" (Schaffer, 2013). To better depict the co-

existence of these genetic variations in family R1352, their dual segregation was shown on the condensed pedigree (Figures 21-26). Six different combinations of these 4 variants were depicted on the pedigree, to show the various possibilities of digenic inheritance. The co-segregation of *C4orf6* c.1A>G and *GIGYF2* c.3494A>G is shown in Figure 21, which demonstrates an interesting pattern. It appears that each variant is being transmitted from a different side of the pedigree, culminating in the presence of both variants in the 4 affected siblings: Z1039, Z1040, Z1497 and Z1508, and the half-sibling Z1507. Individual Z1651 has the *C4orf6* variant, while Z1533 has the *GIGYF2* variant. Tested unaffected relatives have 1 of these variants, or neither of them.

A similar pattern was seen for *GIGYF2* c.3494A>G and *RP1L1* c.202C>T, though unaffected member Z1504 had both variants (Figure 22). In the combinations of *C4orf6* c.1A>G and *ATP1A4* c.1798C>T (Figure 23), and *ATP1A4* c.1798C>T and *RP1L1* c.202C>T (Figure 24), the 5 siblings also share both variants. The main difference is that Z1507 is homozygous for *ATP1A4* c.1798C>T. In each case, none of the unaffected relatives have both variants.

Finally, the pair of *ATP1A4* c.1798C>T and *GIGYF2* c.3494A>G does not appear to be indicative of digenic inheritance (Figure 25). In this scenario, individual Z1651 has neither of the 2 variants, though the remaining 6 affected individuals have both. Three of the unaffected relatives (Z1506, Z1535 and Z1538) have both variants as well. Similarly, the *C4orf6* c.1A>G and *RP1L1* c.202C>T pair does not clearly fit the digenic inheritance model (Figure 26). Individual Z1533 does not have either of these variants, and 1 unaffected relative, Z1660 has both. The consideration of digenic inheritance in the context of familial IA will be explored in the Discussion.

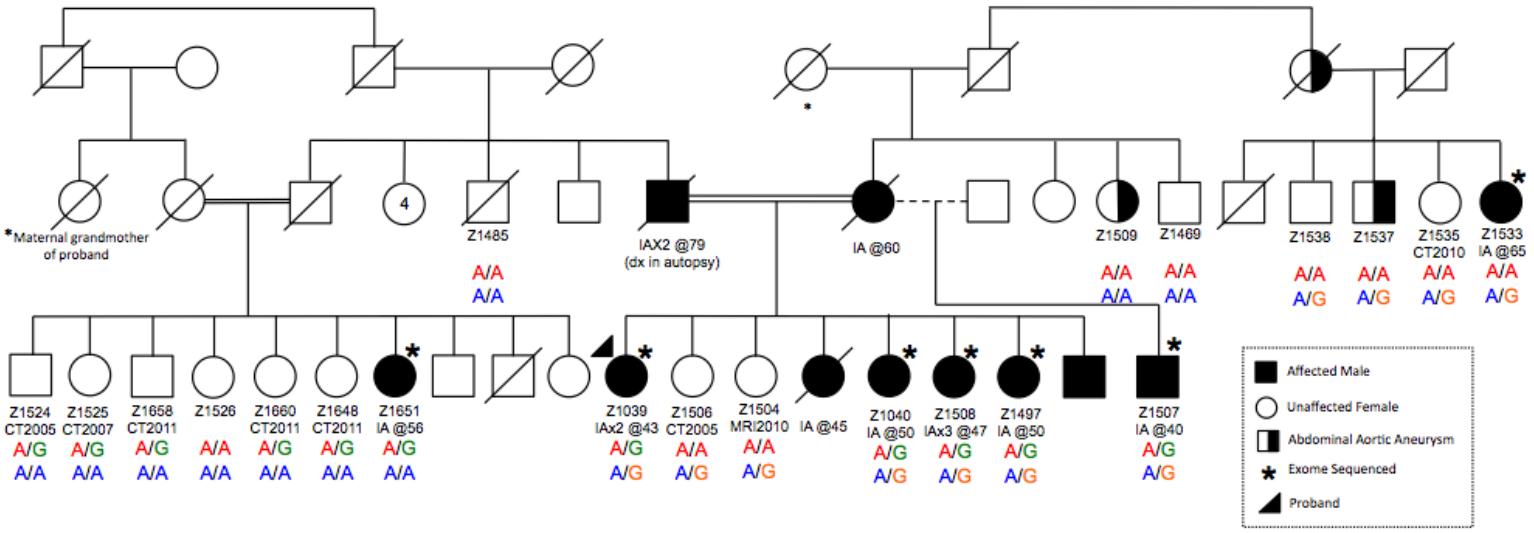


Figure 21. Segregation of *C4orf6* c.1A>G and *GIGYF2* c.3494A>G in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *C4orf6* c.1A>G (top row) and *GIGYF2* c.3494A>G (bottom row) variants are both present in five affected siblings, which may indicate that both variants are implicated in IA development or disease severity. For unaffected individuals, the date of their last known negative CT scan is indicated.

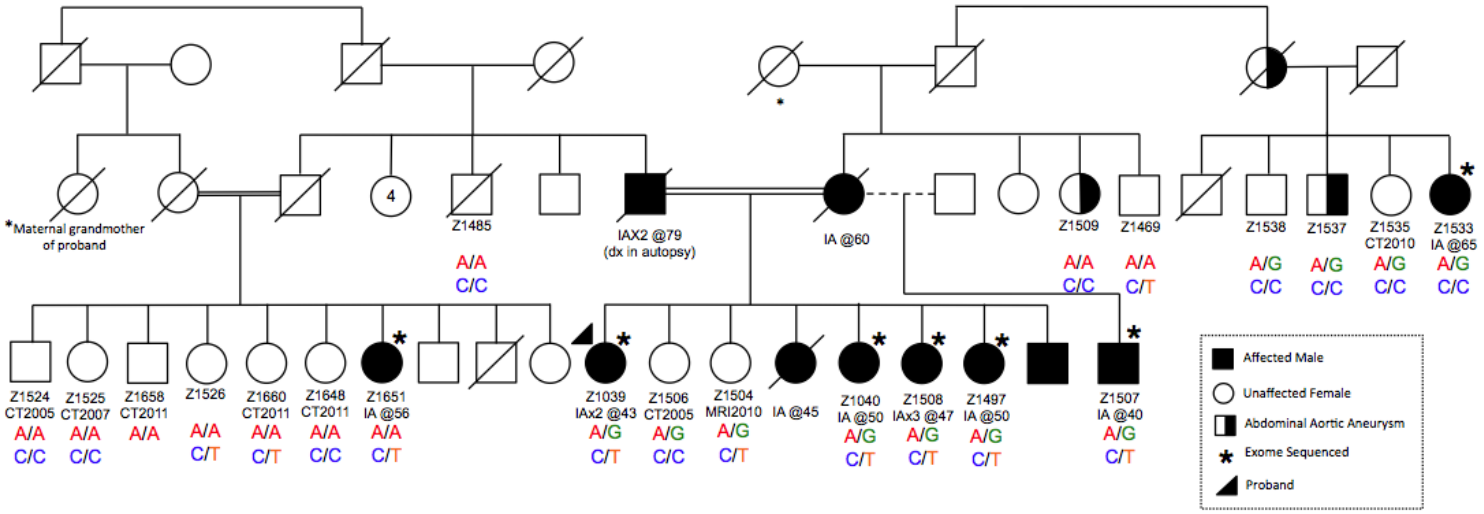


Figure 22. Segregation of *GIGYF2* c.3494A>G and *RP1L1* c.202C>T in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *GIGYF2* c.3494A>G (top row) and *RP1L1* c.202C>T (bottom row) variants are both present in five affected siblings, which may indicate that both variants are implicated in IA development or disease severity. For unaffected individuals, the date of their last known negative CT scan is indicated.

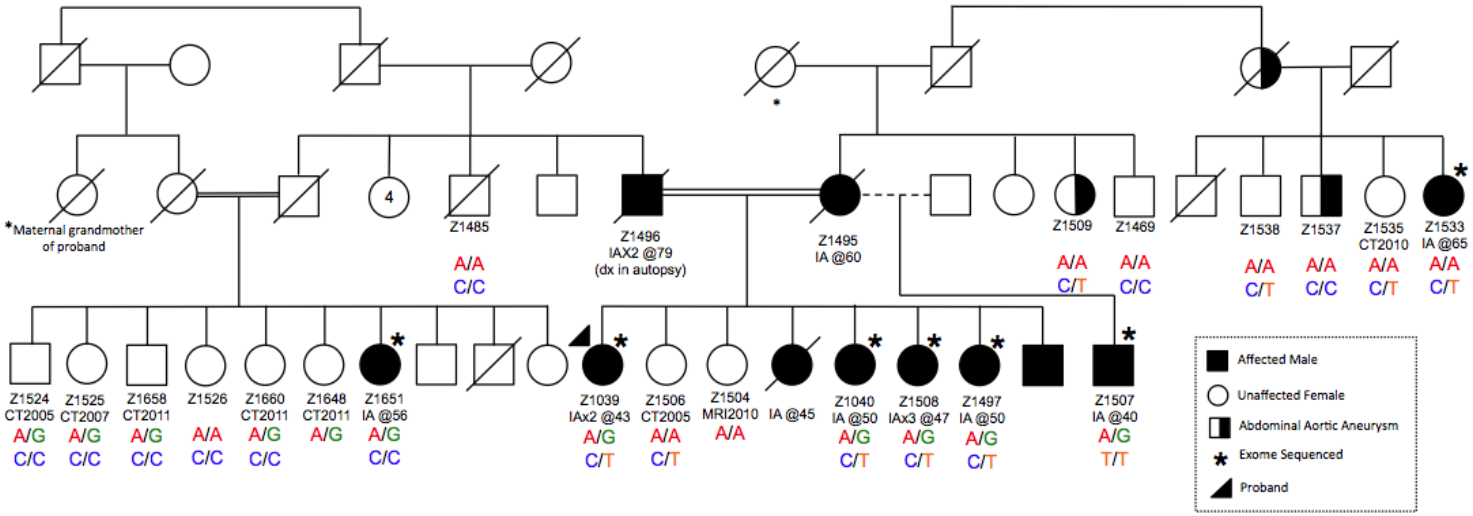


Figure 23. Segregation of *C4orf6* c.1A>G and *ATP1A4* c.1798C>T in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *C4orf6* c.1A>G (top row) and *ATP1A4* c.1798C>T (bottom row) variants are both present in five affected siblings, which may indicate that both variants are implicated in IA development or disease severity. For unaffected individuals, the date of their last known negative CT scan is indicated.

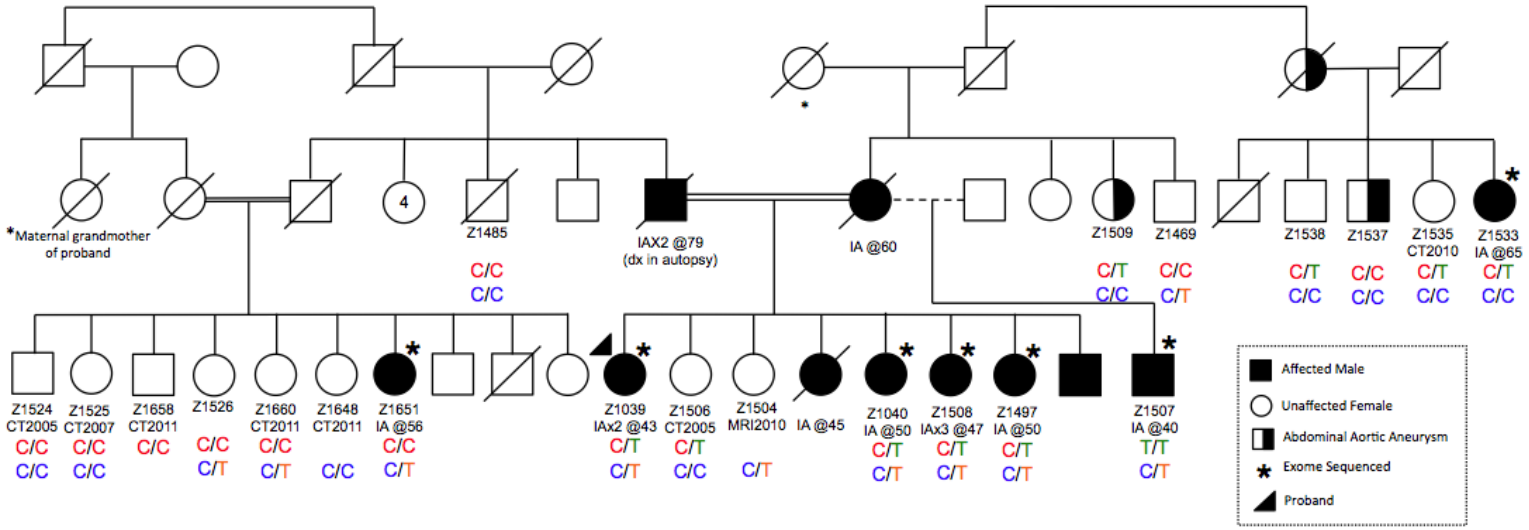


Figure 24. Segregation of *ATP1A4* c.1798C>T and *RP1L1* c.202C>T in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *ATP1A4* c.1798C>T (top row) and *RP1L1* c.202C>T variants are both present in five affected siblings, which may indicate that both variants are implicated in IA development or disease severity. For unaffected individuals, the date of their last known negative CT scan is indicated.

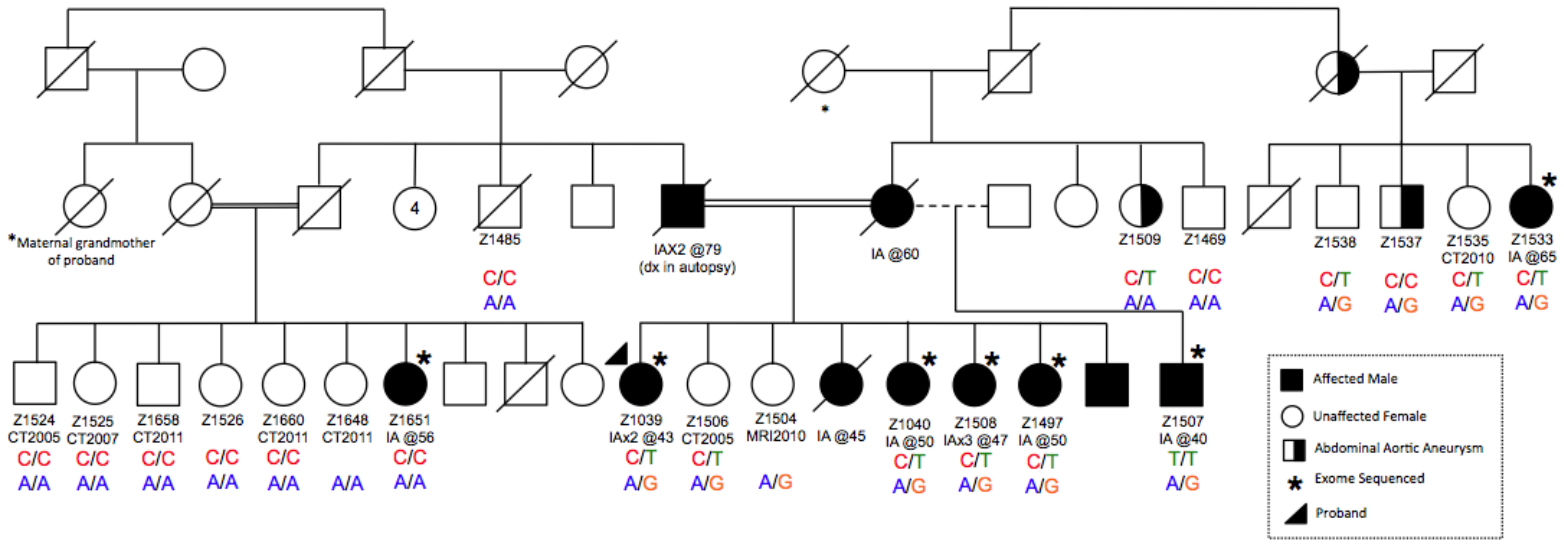


Figure 25. Segregation of *ATP1A4* c.1798C>T and *GIGYF2* c.3494A>G in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *ATP1A4* c.1798C>T (top row) and *GIGYF2* c.3494A>G variants are both present in five affected siblings, but are absent from individual Z1651. For unaffected individuals, the date of their last known negative CT scan is indicated.

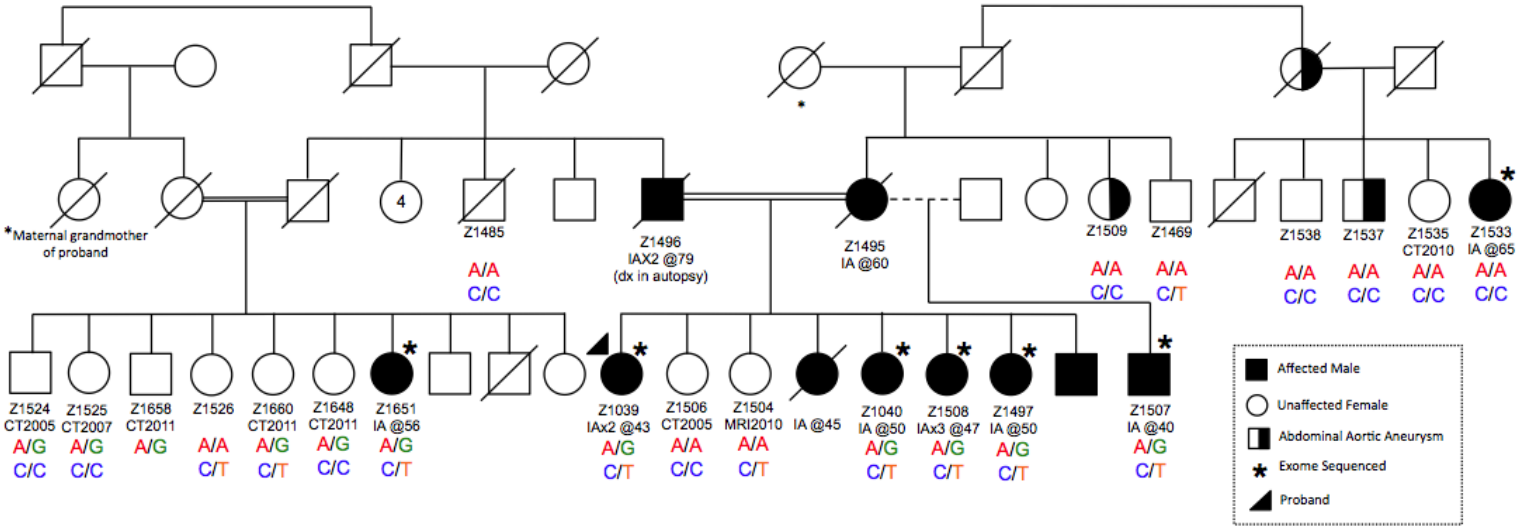


Figure 26. Segregation of *C4orf6* c.1A>G and *RP1L1* c.202C>*T* in family R1352.

DNA extraction was completed for individuals with a Z identification number. IA affection status is indicated. The *C4orf6* c.1A>G (top row) and *RP1L1* c.202C>T variants are both present in five affected siblings, but are absent from individual Z1533. For unaffected individuals, the date of their last known negative CT scan is indicated.

4. Discussion

4.1 IA Mode of Inheritance

It is widely acknowledged that there is a genetic component to the aggregation of IA cases in families. Presently, several statistical associations have been identified, though no causal variants for IA predisposition have been confirmed. Through the study of two multiplex families from Newfoundland and Labrador, it has become evident that inheritance of IA is a complex genetic phenomenon. At the onset of this study, it was hypothesized that there were one or more strongly penetrant variants in our cohort that were contributing to IA predisposition, and that these variants would be rare or novel. It was predicted that these families were exhibiting either autosomal dominant or recessive inheritance of an IA causative variant, based on strong family history and pedigree characteristics, in concert with environmental risk factors. A distinct lack of homozygous variants may indicate that IA is not an autosomal recessive disorder in these Newfoundland families, and that this interpretation of the pedigrees should no longer be considered. If these families exhibited autosomal recessive inheritance, then all affected individuals (with the exception of phenocopies) would have a homozygous variant in a single gene or compound heterozygous variants. However, there is also the possibility that somehow homozygous variants were not called with accuracy through the MUGQIC bioinformatics pipeline. Further inquiry into the absence of shared, rare homozygous variants is necessary.

Our results revealed incomplete segregation of several variants of unknown functional consequence, and an overall lack of shared filtered variants between the two

families. As well, none of the filtered variants overlapped with significant candidate genes or GWAS loci from previous IA association studies. As most large-scale association studies have focused on sporadic IA, it can be hypothesized that different genetic factors may be associated with familial IA development.

Based on our knowledge of modifiable risk factors such a smoking, hypertension and alcohol use in these families, it is quite possible that one or more affected individuals may be phenocopies in our study. IA can occur sporadically as a consequence of lifestyle and environment, and manifest with the same level of severity as familial IA. The presence of phenocopies would explain the absence of a shared candidate variant among all affected individuals within a single family. In this study, I chose to examine variants found in at least 4/5 exomes from R1256, and 6/7 from R1352. This low phenocopy estimation was selected based on the prediction that most members of the family would in fact share a genetic risk factor. In previous family-based WES studies of IA, Farlow et al. (2015) and Yan et al. (2015) did not account for phenocopies, and only kept variants that were shared by all affected relatives with the phenotype. Feng et al. (2011) discussed the importance of accounting for phenocopies when studying complex disease in families, given that sporadic cases occur frequently in the general population. Further examination of the clinical data, and a less strict estimation of phenocopy percentage could be used in future studies of the NL cohort.

Incomplete penetrance may also be an issue to consider as we move forward with this study. We chose to not include an unaffected individual in our exome cohort in response to this possibility. Therefore, our segregation analysis in unaffected family members should be treated with caution, as variant carriers may not develop IA. To cover

all possibilities of disease pathogenesis, alternate modes of inheritance could also be explored. It is possible that IA exhibits oligogenic or polygenic inheritance, and that our focus on monogenic risk factors is not fitting for this disease model. As our knowledge of the genetic etiology of IA grows, more questions continue to arise. The in-depth exploration of top candidate variants in our cohort, ideally through functional studies, is necessary to determine their relevance as we move forward with alternate study designs and new hypotheses.

4.2 Predicted Pathogenicity of Top Candidate Variants

Functionally, the top candidate genes in this study, *C4orf6*, *GIGYF2* and *SPDYE4*, do not have an obvious role to play in the pathophysiology of IA. As a result of our continually growing knowledge of the intricate human genome, these genes may have functions in vascular biology that are not yet known. To further assess the predicted pathogenicity of these variants, statistical methods can be used to determine the likelihood that these variants are involved in IA predisposition.

Moller, Clarke and Maehle (2011) developed a set of equations called the simplified method for segregation analysis (SISA), which can be used to statistically interpret variant segregation and penetrance in a family. This method is based on the value "n" that equals the number of informative meioses in a family. The value "n" is denoted as the number of affected variant carriers minus one. This value can be used to calculate the probability of co-segregation of the phenotype and a genetic variant by chance, $(1/2)^n$. For *SPDYE4* c.103C>T, "n" equals 9, and the probability of co-segregation by chance is 0.001953125, or 0.195%. This means that there is a high

probability that the segregation of this variant with the disease is unlikely to be due to chance. However, this calculation does not include obligate carriers, and should just be used as a supportive piece of evidence, rather than a definitive clue toward pathogenicity. This calculation was also performed for *C4orf6* c.1A>G, where "n" equals 5 (number of known affected carriers minus one). In this case, the probability of co-segregation by chance is equal to 0.03125, or 3.125%. This is the same probability for *GIGYF2*, c.3494A>G, as well as the *ATP1A4* and *RP1L1* variants.

4.3 Digenic Inheritance in Family R1352

The presence of 4 variants of interest in R1352, each with an equal probability of cosegregation by chance, led to the consideration of digenic inheritance in this family. Schaffer (2013) stated that digenic inheritance encompasses cases where both loci are responsible for determining affected status, or where the combination of both loci leads to a "substantial change in severity, or a substantial change in age of onset". Within this definition, one locus may be the primary variant, or both variants may be equally responsible for the phenotype (Schaffer, 2013). The limited scope of studying monogenic inheritance alone may not be sufficient to explain the features of genetically heterogeneous diseases. The introduction of WES and other high throughput methods has allowed researchers to uncover more than one candidate variant that segregates in a single family.

The first case of digenic inheritance was described by Kajiwara et al. (1994), in their report on retinitis pigmentosa. In three unrelated families, only individuals with both a missense mutation in *PRPH2* and a null mutation in *ROM1* exhibited retinitis

pigmentosa. The two genes, *PRPH2* and *ROM1*, are located on different chromosomes, and encode proteins that are known to interact. Since this time, digenic inheritance has been observed in a multitude of heritable diseases, including several adult-onset disorders with varied modes of inheritance. Tang et al. (2006) reported on a family with early-onset Parkinson's disease, exhibiting digenic inheritance. Previously, autosomal recessive inheritance of this disease, via homozygous mutations in the *PINK1* and *DJ-1* genes had been described. However, the mechanism by which heterozygous variants in these genes led to the phenotype was unclear. The study by Tang et al. (2006) focused on a Chinese family, with two sisters that developed Parkinson's disease in their late 20s. Both siblings shared heterozygous variants in *PINK1* and *DJ-1*, and two unaffected relatives had either the *PINK1* or *DJ-1* variant alone. Another one of their unaffected siblings had both variants, which the authors attributed to decreased penetrance in this early onset form of the disease.

In the R1352 Results, it was determined that the segregation of *C4orf6* c.1A>G and *GIGYF2* c.3494A>G was of particular interest. In this scenario, the 4 siblings (Z1039, Z1040, Z1497, Z1508) and 1 half-sibling (Z1507) share both variants. Individual Z1533 has the *GIGYF2* variant alone, while Z1651 has the *C4orf6* start-lost variant. Schaffer (2013) emphasized that strong supportive evidence for digenic inheritance includes the comparison of phenotypic data between affected family members. In our study, we were fortunate to have access to a wealth of clinical information from each family. The age range of diagnosis for the 5 affected relatives that share both variants was 40-50 years. The 2 individuals that have only 1 of the variants, Z1651 and Z1533, were diagnosed at age 56 and 65, respectively. It is possible that the combination of both

loci could cause a slightly earlier age of onset, though the age of diagnosis might have also been influenced by the year that they were recruited to the IA study. For example, the proband, Z1039, was diagnosed at age 43, in 2005, whereas individual Z1533 was diagnosed by CT scan in 2010. As well, patient diagnosis either occurred following clinical presentation (i.e. complications or rupture) or after clinical screening, which must be taken into account when assessing age of diagnosis.

IA location or size does not appear to be correlated with the presence of both variants versus one. Z1651 had an aneurysm measuring 1.5-2 mm, and Z1533 had a much larger 9.4x5.6 mm aneurysm. Across the entire family, there was variance in IA location (Table 6), and all individuals with both variants do not share a common site or size in the Circle of Willis.

In order to explore protein-protein interactions between *C4orf6*, *GIGYF2*, *RP1L1*, and *ATP1A4*, the STRING v.10 database was queried. The "Search Tool for the Retrieval of Interacting Genes/Proteins" or STRING, is an online database of known and predicted interactions between genes and proteins, that have been integrated from multiple sources for ease of use (Szklarczyk et al., 2015). Upon entering the names of specific proteins, STRING generates a map, with coloured lines connecting interacting proteins. For the 4 genes of interest in R1352 there were no known interactions detected. Therefore, this database does not provide any supporting evidence for the digenic inheritance hypothesis. However, digenic inheritance is not solely dependent on protein-protein interaction, and there are a myriad of analyses that could be performed. Replication of these variants in additional IA families from NL would contribute to the plausibility of this mode of inheritance.

4.4 Implications of Low Impact Variants

To assess the potential role of synonymous variation in IA, I decided to apply the same filtering stategy, as outlined in Figure 11, to the low impact variant lists from families R1352 and R1256. My study design allowed an in-depth analysis of the moderate and high impact SNVs and small INDELs present in the exome and narrow exon/intron boundaries. Low impact variants are less commonly implicated in heritable disease, though several cases exist. For example, a synonymous variant, c.313C>T, in the *IRGM* gene has been associated with Crohn's disease susceptibility, and is a predicted causal variant for this complex immune disease (Brest et al., 2011). In addition to statistical association, synonymous mutations have also been classified as the direct cause of heritable disease. A synonymous mutation in exon 3 of the *FGFR2* gene has been reported as the cause of an autosomal dominant bone disease, Crouzon syndrome, in multiple families (Del Gatto & Breathnach, 1995).

There is a growing body of evidence that supports the potential effect of synonymous variation on the human genome. Sauna & Kimchi-Sarfaty (2011) have emphasized that these variants can cause aberrant mRNA splicing, and also impact the stability of mRNA molecules. Our inability to demonstrate the mechanisms by which synonymous variants can cause different diseases is currently one of the roadblocks that prevents many research groups from including low impact variants in their analyses. As our knowledge of the human genome and its complexities increases, more synonymous changes may be implicated in heritable disease.

In family R1352, 2 of the filtered synonymous variants, *APBB2* c.231G>A and *COL6A3* c.702C>T were of interest functionally. The *APBB2* gene encodes amyloid beta A4 precursor protein-binding family B member 2 protein. This gene has been associated with late-onset Alzheimer disease (Y. Li et al., 2005). Besides beta-amyloid binding, other known protein functions include extracellular matrix organization and regulation of apoptosis, which may be of interest in IA pathophysiology. The *COL6A3* gene is responsible for encoding the alpha 3 chain of type 6 collagen. Variants in this gene have been connected to Mendelian inheritance of Bethlem myopathy 1 (Baker et al., 2007), Dystonia 27 (Zech et al., 2015), and Ullrich congenital muscular dystrophy 1 (Baker et al., 2005). Collagen is an essential structural component of the vascular wall, though this gene has not been connected to any aneurysm phenotypes. In our NextGene® results, there was a rare heterozygous missense variant in the same gene, that was found in 6/7 exomes from this same family. The *COL6A3* c.5610C>T variant was not called by the MUGQIC bioinformatics pipeline, and thus was not validated by Sanger sequencing. Interestingly, this variant was only absent from individual Z1651. This same family member was also the only 1 of 7 that did not have the *COL6A3* c.702C>T synonymous variant. Individual Z1651 could possibly be a phenocopy, and this gene could be explored further in upcoming research plans.

In family R1256, 2 novel synonymous changes were detected in the *TNRC18* and *FAM83G* genes, and were called in all 5 exomes. *TNRC18* encodes the trinucleotide containing 18 protein, which may be involved in chromatin binding and silencing. *FAM83G* is also known as family with sequence similarity 83, member G. This gene is found in a complex with the *SMAD1* gene. SMAD1 is a transforming growth factor beta

signaling protein, and has been linked to various diseases, including pulmonary arterial hypertension (Nasim et al., 2011). However, this disease-gene connection has not been confirmed. Previously reported variants in *OTOGL*, *PCDH9* and *GPR179* also passed the filtering strategy, and were detected in all five exomes. *OTOGL* has specific expression in the inner ear, while *PCDH9* is involved in the mediation of cell adhesion in the nervous system. *GPR179* is involved in the vision pathway, and has been associated with a congential form of night blindness (Audo et al., 2012). Sanger sequencing was not used to validate any of these rare, synonymous variants that passed our filtering methodology. At this time, there are insufficient bioinformatics tools and functional assays available to assess the predicted pathogenicity of synonymous variation in the genome. Therefore, it would be difficult to establish a link between any of these variants and IA predisposition, based primarily on segregation. As well, none of these synonymous changes are found in previously reported IA-associated genes, or genes with a clear functional relevance to this disease. However, these additional analyses could be helpful when studying additional IA families from the Newfoundland and Labrador population. It is possible that one or more of these rare, synonymous variants segregates with IA in another family as well.

4.5 Strengths and Limitations of Study

This study design had several strengths, including the selection of WES technology for disease gene discovery. The emergence of WES has expanded our ability to visualize the unique variation present in the coding portion of the human genome. As the cost and time associated with WES has declined, its accessibility and use in disease-gene discovery has rapidly increased. The use of a WES-based study design allowed me

to visualize all of the coding variants in our 12-patient cohort in a high-throughput manner, which is a definite strength in comparison to traditional sequencing methods. Additionally, access to a wealth of clinical information and the recruitment of multiplex families was a major strength of our research. Established diagnoses and strong family histories allowed us to build a filtering strategy around the exploration of shared variants in multiple family members.

Consequently, many of the limitations of this study also stemmed from the particular nature of WES technology, and the uncertainty surrounding the mode of inheritance of IA. The major challenge presented by WES is the shear abundance of data generated from a single exome. Accurately calling and annotating the thousands of variants from a single individual is a task that requires immense bioinformatics support. Deciding the best way to prioritize the thousands of variants in an exome, and determining which ones are likely instrumental in human disease is an ongoing challenge.

Each step of my filtering strategy was carefully planned, based on consultation of other family-based research studies that have utilized WES. As none of the top candidate variants in my study segregated completely with IA, it is possible that the filtering steps were too stringent for the characteristics of this disease. Adjustment of the MAF cut-off to 5%, for example, would allow us to consider more common variants, which may be involved in IA predisposition in conjunction with other risk factors. However, this dataset would have been much larger, and issues with variant prioritization would still be present. Gaps in the available data such as gene function and expression in bodily tissues is a definite limitation to our ability to interpret variants with accuracy (Figure 27).

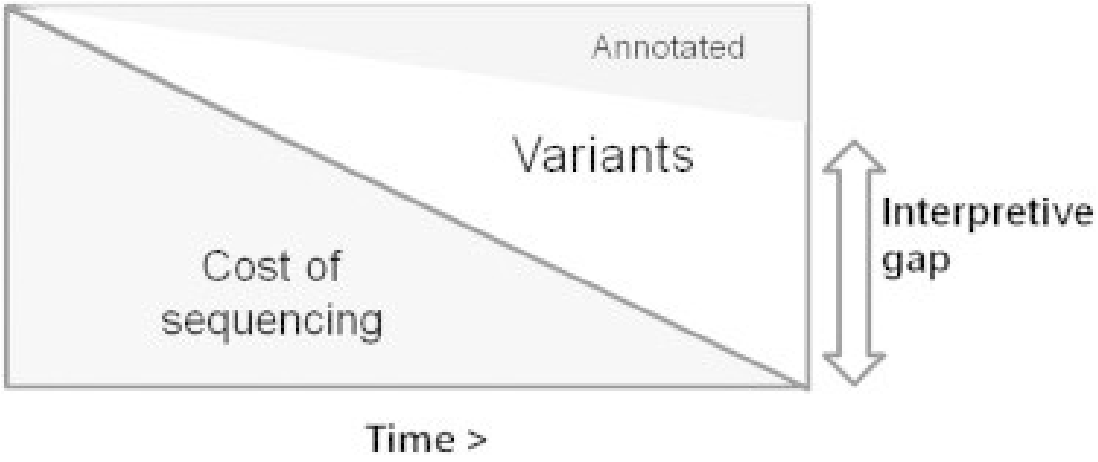


Figure 27. The increasing gap between the accessibility of genomic data and our ability to interpret variants and their clinical implications. Reprinted from Cutting (2014) with copyright permission.

Technical limitations must also be considered in WES studies. Failed exome capture and low coverage in certain regions of the genome are both possible complications. For example, areas of the genome with a high GC content may be more difficult to capture with accuracy (Schwartzentruber, 2012). DNA sample quality can also impact overall coverage. As well, due to quality control methods, such as filtering out areas of low read depth, certain regions of the exome may be unrepresented. As methods of WES and complimentary data analysis become more powerful, the identification of candidate variants for IA within these regions may be possible. Variant calling errors and software glitches can also occur, which are an unavoidable consequence of computational analysis. We outsourced our WES and bioinformatics support, as we did not have the technology and infrastructure to perform these experiments at the time. Therefore, it was not possible for us to oversee each step of the analyses.

Another limitation of a WES-based study is that this technology only captures 1% of the entire human genome. Though 85% of the mutations identified in Mendelian diseases are located in the exome, this study design ignores the non-coding portion of the genome, which may have an unknown role to play in IA inheritance (Ng et al., 2009). Ultimately, the unknown mode of inheritance of IA is a major limitation to our ability to identify causal variants in our cohort. As well, WES only allows the detection of SNPs and small INDELs. In the context of complex disease, Wang et al. (2013) discusses the need to also evaluate low frequency variants, copy number variants, and also other structural changes.

Finally, there were several faults within the MUGQIC pipeline and NextGene®, which limited our ability to accurately annotate all variants. The major fault lies within the dbNSFP database, which was used in both the MUGQIC and NextGene® analyses. This database has an abundance of missing scores that may need to be imputed manually for certain SNPs, and this database must be re-downloaded and installed within NextGene® periodically to maintain updates. The incorporation of ExAC Browser and NHLBI Exome Variant Server allele frequencies would be a welcome addition to the NextGene® track manager as well. The utility of NextGene® for our laboratory's purposes is still undetermined. This software is still a relatively new development, and has not been referenced in many published works. As newer versions of NextGene® are released, and we gather larger amounts of exome data, its desktop convenience may be of interest for our data analysis needs.

4.6 Future Directions

In the immediate future, there are plans to complete a copy number variant (CNV) analysis for the 12 individuals that were selected for WES in this study. As WES only covers SNVs and small INDELs, it is possible that a larger chromosomal abnormality may be present in these families. To identify CNVs in these samples, a genome-wide SNP chip assay will be used, in conjunction with analytical software. The unknown genetic etiology of IA opens up a wide variety of possible study designs for our research team. As suggested earlier, the low, moderate and high impact lists for these families can be re-filtered in a variety of ways. It is possible that a more common genetic factor is responsible for IA development, in conjunction with environmental agents. Therefore, the

MAF cut-off could be extended to the 2-5% range. To compensate for the additional variants, a different filter could be included such as Polyphen2 scores, or the presence of a functional keyword such as "vascular" in gene descriptions. There are endless opportunities for re-design of the filtering strategy, which may be adjusted in the coming years as research concerning best practices for exome data analysis and disease gene discovery undergoes advancement.

It would also be beneficial to connect with potential collaborators from other institutions that have conducted WES of familial IA cases. It is possible that there are shared variants between our project and another independent study. The comparison of filtered high, moderate, and low impact lists, and discussion about IA mode of inheritance in general, with other IA researchers could be a future direction.

Though my research has focused specifically on families R1256 and R1352, future work will involve the other families in the NL cohort that have a high incidence of IA. Six additional families have 4 or more affected members, which can be analyzed through WES. Comparisons could then be drawn between multiple families, to look for shared genetic risk factors for IA in this population. Ultimately, our future pursuits will lead to the functional analysis of any variants that have strong evidence for causation. A zebrafish model is proposed to demonstrate the effects of these genetic variations in the cerebral vasculature and overall body. Zebrafish are a highly useful model organism for this project, as they have rapid development and their embryos are optically clear, in comparison to a mouse model (Walcott & Peterson, 2014) Thus, the vascular system of the zebrafish can be viewed in real-time through a microscope, to investigate the formation of arterial lesions, and view alterations in hemodynamics and overall vessel

structure (Walcott and Peterson, 2014). Gene knock-down through the use of morpholino oligonucleotides in zebrafish would allow us to model the effects of deleterious variants, in a cost-effective and high throughput manner (Phillips & Westerfield, 2014). This is a practical model for the investigation of a disease of unknown genetic etiology.

4.7 Conclusion

This study marks the first investigation into genetic risk factors for familial IA in the Newfoundland and Labrador population, which is known to have an increased prevalence of several heritable diseases. Through the use of WES technology, the exomes of 12 affected individuals from two families – R1352 and R1256, were analyzed. Filtering of this data resulted in the identification of several candidate variants, though none of the variants segregated completely with IA incidence. In R1352, the high impact variant *C4orf6* c.1A>G was identified as a top candidate, and the moderate impact variant *GIGYF2* c.3494A>G was also highlighted as a variant of interest. In R1256, *SPDYE4* c.103C>T was the only variant to segregate in the majority of affected family members. Each of these three variants was absent from 100 population controls from Newfoundland and Labrador. Functional annotation and replication of these variants in other affected individuals would be necessary to generate more evidence for pathogenicity. Further genetic research in this province may culminate in better outcomes for familial IA patients, and result in the ability to diagnose IA through genetic testing methods, leading to more effective treatment and preventative measures. The results generated from this study have stimulated multiple new research questions and hypotheses, which will be exciting to pursue in the years to come.

References

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248-249.

Akagawa, H., Tajima, A., Sakamoto, Y., Krischek, B., Yoneyama, T., Kasuya, H., . . . Inoue, I. (2006). A haplotype spanning two genes, ELN and LIMK1, decreases their transcripts and confers susceptibility to intracranial aneurysms. *Human Molecular Genetics, 15*(10), 1722-1734.

Akahori, M., Tsunoda, K., Miyake, Y., Fukuda, Y., Ishiura, H., Tsuji, S., . . . Iwata, T. (2010). Dominant mutations in RP1L1 are responsible for occult macular dystrophy. *American Journal of Human Genetics, 87*(3), 424-429.

Alg, V. S., Sofat, R., Houlden, H., & Werring, D. J. (2013). Genetic risk factors for intracranial aneurysms: A meta-analysis in more than 116,000 individuals. *Neurology, 80*(23), 2154-2165.

Ali, M. S., Starke, R. M., Jabbour, P. M., Tjoumakaris, S. I., Gonzalez, L. F., Rosenwasser, R. H., . . . Dumont, A. S. (2013). TNF-alpha induces phenotypic modulation in cerebral vascular smooth muscle cells: Implications for cerebral aneurysm pathology. *Journal of Cerebral Blood Flow and Metabolism : Official Journal of the International Society of Cerebral Blood Flow and Metabolism, 33*(10), 1564-1573.

Aoki, T., Nishimura, M., Matsuoka, T., Yamamoto, K., Furuyashiki, T., Kataoka, H., . . . Narumiya, S. (2011). PGE(2) -EP(2) signalling in endothelium is activated by

haemodynamic stress and induces cerebral aneurysm through an amplifying loop via NF-kappaB. *British Journal of Pharmacology, 163*(6), 1237-1249.

Audo, I., Bujakowska, K., Orhan, E., Poloschek, C. M., Defoort-Dhellemmes, S., Drumare, I., . . . Zeitz, C. (2012). Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness. *American Journal of Human Genetics, 90*(2), 321-330.

Baker, N. L., Morgelin, M., Pace, R. A., Peat, R. A., Adams, N. E., Gardner, R. J., . . . Lamande, S. R. (2007). Molecular consequences of dominant bethlem myopathy collagen VI mutations. *Annals of Neurology, 62*(4), 390-405.

Baker, N. L., Morgelin, M., Peat, R., Goemans, N., North, K. N., Bateman, J. F., & Lamande, S. R. (2005). Dominant collagen VI mutations are a common cause of ullrich congenital muscular dystrophy. *Human Molecular Genetics, 14*(2), 279-293.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011). Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics, 12*(11), 745-755.

Bederson, J. B., Awad, I. A., Wiebers, D. O., Piepgras, D., Jr, H.,E.C., Brott, T., . . . Caroselli, C. (2000). Recommendations for the management of patients with unruptured intracranial aneurysms: A statement for healthcare professionals from the stroke council of the american heart association. *Stroke; a Journal of Cerebral Circulation, 31*(11), 2742-2750.

Bilguvar, K., Yasuno, K., Niemela, M., Ruigrok, Y. M., von Und Zu Fraunberg, M., van Duijn, C. M., . . . Gunel, M. (2008). Susceptibility loci for intracranial aneurysm in european and japanese populations. *Nature Genetics, 40*(12), 1472-1477.

Boileau, C., Guo, D. C., Hanna, N., Regalado, E. S., Detaint, D., Gong, L., . . . Milewicz, D. M. (2012). TGFB2 mutations cause familial thoracic aortic aneurysms and

dissections associated with mild systemic features of marfan syndrome. *Nature Genetics, 44*(8), 916-921.

Brest, P., Lapaquette, P., Souidi, M., Lebrigand, K., Cesaro, A., Vouret-Craviari, V., . . . Hofman, P. (2011). A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in crohn's disease. *Nature Genetics, 43*(3), 242-245.

Broderick, J. P., Brown, R. D.,Jr, Sauerbeck, L., Hornung, R., Huston, J.,3rd, Woo, D., . . . FIA Study Investigators. (2009). Greater rupture risk for familial as compared to sporadic unruptured intracranial aneurysms. *Stroke; a Journal of Cerebral Circulation, 40*(6), 1952-1957.

Broderick, J. P., Sauerbeck, L. R., Foroud, T., Huston, J.,3rd, Pankratz, N., Meissner, I., & Brown, R. D.,Jr. (2005). The familial intracranial aneurysm (FIA) study protocol. *BMC Medical Genetics, 6*, 17.

Brown, R. D.,Jr, & Broderick, J. P. (2014). Unruptured intracranial aneurysms: Epidemiology, natural history, management options, and familial screening. *The Lancet Neurology, 13*(4), 393-404.

Chalouhi, N., Ali, M. S., Jabbour, P. M., Tjoumakaris, S. I., Gonzalez, L. F., Rosenwasser, R. H., . . . Dumont, A. S. (2012). Biology of intracranial aneurysms: Role of inflammation. *Journal of Cerebral Blood Flow and Metabolism : Official Journal of the International Society of Cerebral Blood Flow and Metabolism, 32*(9), 1659-1676.

Chalouhi, N., Hoh, B. L., & Hasan, D. (2013). Review of cerebral aneurysm formation, growth, and rupture. *Stroke; a Journal of Cerebral Circulation, 44*(12), 3613-3622.

Chapman, A. B., Rubinstein, D., Hughes, R., Stears, J. C., Earnest, M. P., Johnson, A. M., . . . Kaehny, W. D. (1992). Intracranial aneurysms in autosomal dominant polycystic kidney disease. *The New England Journal of Medicine, 327*(13), 916-920.

Chen, Z., Ma, J., Cen, Y., Liu, Y., & You, C. (2013). The angiotensin converting enzyme insertion/deletion polymorphism and intracranial aneurysm: A meta-analysis of case-control studies. *Neurology India, 61*(3), 293-299.

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research, 19*(9), 1553-1561.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics, 3*, 35.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly, 6*(2), 80-92.

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research, 15*(7), 901-913.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research, 43*, D662-9.

Cutting, G. R. (2014). Annotating DNA variants is the next major goal for human genetics. *American Journal of Human Genetics, 94*(1), 5-10.

de Oliveira Manoel, A. L., Turkel-Parrella, D., Duggal, A., Murphy, A., McCredie, V., & Marotta, T. R. (2015). Managing aneurysmal subarachnoid hemorrhage: It takes a team. *Cleveland Clinic Journal of Medicine, 82*(3), 177-192.

Del Gatto, F., & Breathnach, R. (1995). A crouzon syndrome synonymous mutation activates a 5' splice site within the IIIc exon of the FGFR2 gene. *Genomics, 27*(3), 558-559.

Exome Aggregation Consortium. (2015). ExAC, cambridge, MA. Retrieved from http://exac.broadinstitute.org.

Exome Variant Server. (2015). NHLBI exome sequencing project (ESP), seattle, WA. Retrieved from http://evs.gs.washington.edu/EVS/.

Faivre, L., Collod-Beroud, G., Callewaert, B., Child, A., Loeys, B. L., Binquet, C., . . . Jondeau, G. (2009). Pathogenic FBN1 mutations in 146 adults not meeting clinical diagnostic criteria for marfan syndrome: Further delineation of type 1 fibrillinopathies and focus on patients with an isolated major criterion. *American Journal of Medical Genetics.Part A, 149A*(5), 854-860.

Farlow, J. L., Lin, H., Sauerbeck, L., Lai, D., Koller, D. L., Pugh, E., . . . FIA Study Investigators. (2015). Lessons learned from whole exome sequencing in multiplex families affected by a complex genetic disorder, intracranial aneurysm. *PloS One, 10*(3), e0121104.

Farnham, J. M., Camp, N. J., Neuhausen, S. L., Tsuruda, J., Parker, D., MacDonald, J., & Cannon-Albright, L. A. (2004). Confirmation of chromosome 7q11 locus for predisposition to intracranial aneurysm. *Human Genetics, 114*(3), 250-255.

Feng, B. J., Tavtigian, S. V., Southey, M. C., & Goldgar, D. E. (2011). Design considerations for massively parallel sequencing studies of complex human disease. *PloS One, 6*(8), e23221.

Foroud, T., & FIA Study Investigators. (2013). Whole exome sequencing of intracranial aneurysm. *Stroke; a Journal of Cerebral Circulation, 44*(6 Suppl 1), S26-8.

Foroud, T., Koller, D. L., Lai, D., Sauerbeck, L., Anderson, C., Ko, N., . . . FIA Study Investigators. (2012). Genome-wide association study of intracranial aneurysms confirms role of anril and SOX17 in disease risk. *Stroke; a Journal of Cerebral Circulation, 43*(11), 2846-2852.

Foroud, T., Sauerbeck, L., Brown, R., Anderson, C., Woo, D., Kleindorfer, D., . . . Familial Intracranial Aneurysm Study Investigators. (2009). Genome screen in familial intracranial aneurysm. *BMC Medical Genetics, 10*, 3-2350-10-3.

Foroud, T., Sauerbeck, L., Brown, R., Anderson, C., Woo, D., Kleindorfer, D., . . . FIA Study Investigators. (2008). Genome screen to detect linkage to intracranial aneurysm susceptibility genes: The familial intracranial aneurysm (FIA) study. *Stroke; a Journal of Cerebral Circulation, 39*(5), 1434-1440.

Fox, J. L. (1982). Familial intracranial aneurysms. *Journal of Neurosurgery, 57*(3), 416-417.

Frosen, J. (2014). Smooth muscle cells and the formation, degeneration, and rupture of saccular intracranial aneurysm wall--a review of current pathophysiological knowledge. *Translational Stroke Research, 5*(3), 347-356.

Frosen, J., Tulamo, R., Paetau, A., Laaksamo, E., Korja, M., Laakso, A., . . . Hernesniemi, J. (2012). Saccular intracranial aneurysm: Pathology and mechanisms. *Acta Neuropathologica, 123*(6), 773-786.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics (Oxford, England), 25*(12), i54-62.

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.

Handley, C. J., Samiric, T., & Ilic, M. Z. (2006). Structure, metabolism, and tissue roles of chondroitin sulfate proteoglycans. *Advances in Pharmacology (San Diego, Calif.), 53*, 219-232.

Harrod, C. G., Batjer, H. H., & Bendok, B. R. (2006). Deficiencies in estrogen-mediated regulation of cerebrovascular homeostasis may contribute to an increased risk of cerebral aneurysm pathogenesis and rupture in menopausal and postmenopausal women. *Medical Hypotheses, 66*(4), 736-756.

Helgadottir, A., Thorleifsson, G., Magnusson, K. P., Gretarsdottir, S., Steinthorsdottir, V., Manolescu, A., . . . Stefansson, K. (2008). The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genetics, 40*(2), 217-224.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., . . . Stefansson, K. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science (New York, N.Y.), 316*(5830), 1491-1493.

Hua, T., Zhang, D., Zhao, Y. L., Wang, S., & Zhao, J. Z. (2008). Correlation of COL3A1 gene with type III collagen stability in intracranial aneurysm. *Zhonghua Yi Xue Za Zhi, 88*(7), 445-448.

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research, 37*(1), 1-13.

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57.

Hussain, I., Duffis, E. J., Gandhi, C. D., & Prestigiacomo, C. J. (2013). Genome-wide association studies of intracranial aneurysms: An update. *Stroke; a Journal of Cerebral Circulation, 44*(9), 2670-2675.

Inoue, K., Mineharu, Y., Inoue, S., Yamada, S., Matsuda, F., Nozaki, K., . . . Koizumi, A. (2006). Search on chromosome 17 centromere reveals TNFRSF13B as a susceptibility gene for intracranial aneurysm: A preliminary study. *Circulation, 113*(16), 2002-2010.

International Study of Unruptured Intracranial Aneurysms Investigators. (1998). Unruptured intracranial aneurysms--risk of rupture and risks of surgical intervention. *The New England Journal of Medicine, 339*(24), 1725-1733.

Kajiwara, K., Berson, E. L., & Dryja, T. P. (1994). Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science (New York, N.Y.), 264*(5165), 1604-1608.

Krischek, B., Tajima, A., Akagawa, H., Narita, A., Ruigrok, Y., Rinkel, G., . . . Inoue, I. (2010). Association of the jun dimerization protein 2 gene with intracranial aneurysms in japanese and korean cohorts as compared to a dutch cohort. *Neuroscience, 169*(1), 339-343.

Krishnamurthi, R. V., Moran, A. E., Forouzanfar, M. H., Bennett, D. A., Mensah, G. A., Lawes, C. M., . . . Global Burden of Diseases, Injuries, and Risk Factors 2010 Study Stroke Expert Group. (2014). The global burden of hemorrhagic stroke: A summary of findings from the GBD 2010 study. *Global Heart, 9*(1), 101-106.

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols, 4*(7), 1073-1081.

Lautier, C., Goldwurm, S., Durr, A., Giovannone, B., Tsiaras, W. G., Pezzoli, G., . . . Smith, R. J. (2008). Mutations in the GIGYF2 (TNRC15) gene at the PARK11 locus in familial parkinson disease. *American Journal of Human Genetics, 82*(4), 822-833.

Leung, H. K., Lam, Y., Cheng, K. M., Chan, C. M., & Cheung, Y. L. (2011). Intracranial aneurysms in twins: Case report and review of the literature. *Hong Kong Medical Journal = Xianggang Yi Xue Za Zhi / Hong Kong Academy of Medicine, 17*(2), 151-154.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England), 27*(21), 2987-2993.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754-1760.

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.

Li, Y., Hollingworth, P., Moore, P., Foy, C., Archer, N., Powell, J., . . . Grupe, A. (2005). Genetic association of the APP binding protein 2 gene (APBB2) with late onset alzheimer disease. *Human Mutation, 25*(3), 270-277.

Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation, 34*(9), E2393-402.

Loeys, B. L., Schwarze, U., Holm, T., Callewaert, B. L., Thomas, G. H., Pannu, H., . . . Dietz, H. C. (2006). Aneurysm syndromes caused by mutations in the TGF-beta receptor. *The New England Journal of Medicine, 355*(8), 788-798.

Low, S. K., Takahashi, A., Cha, P. C., Zembutsu, H., Kamatani, N., Kubo, M., & Nakamura, Y. (2012). Genome-wide association study for intracranial aneurysm in the japanese population identifies three candidate susceptible loci and a functional genetic variant at EDNRA. *Human Molecular Genetics, 21*(9), 2102-2110.

Low, S. K., Zembutsu, H., Takahashi, A., Kamatani, N., Cha, P. C., Hosono, N., . . . Nakamura, Y. (2011). Impact of LIMK1, MMP2 and TNF-alpha variations for intracranial aneurysm in japanese population. *Journal of Human Genetics, 56*(3), 211-216.

Luo, L. Y., Soosaipillai, A., & Diamandis, E. P. (2001). Molecular cloning of a novel human gene on chromosome 4p11 by immunoscreening of an ovarian carcinoma cDNA library. *Biochemical and Biophysical Research Communications, 280*(1), 401-406.

Mackey, J., Brown, R. D., Sauerbeck, L., Hornung, R., Moomaw, C. J., Koller, D. L., . . . Broderick, J. P. (2015). Affected twins in the familial intracranial aneurysm study. *Cerebrovascular Diseases (Basel, Switzerland), 39*(2), 82-86.

Magrane, M., & Consortium, U. (2011). UniProt knowledgebase: A hub of integrated protein data. *Database : The Journal of Biological Databases and Curation, 2011*, bar009.

Maroun, F. B., Murray, G. P., Jacob, J. C., Mangan, M. A., & Faridi, M. (1986). Familial intracranial aneurysms: Report of three families. *Surgical Neurology, 25*(1), 85-88.

Matsui, T., Kanai-Azuma, M., Hara, K., Matoba, S., Hiramatsu, R., Kawakami, H., . . . Kanai, Y. (2006). Redundant roles of Sox17 and Sox18 in postnatal angiogenesis in mice. *Journal of Cell Science, 119*(Pt 17), 3513-3526.

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.

Merner, N. D., Hodgkinson, K. A., Haywood, A. F., 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.

Mineharu, Y., Inoue, K., Inoue, S., Kikuchi, K., Ohishi, H., Nozaki, K., . . . Koizumi, A. (2008). Association analyses confirming a susceptibility locus for intracranial aneurysm at chromosome 14q23. *Journal of Human Genetics, 53*(4), 325-332.

Mineharu, Y., Inoue, K., Inoue, S., Yamada, S., Nozaki, K., Hashimoto, N., & Koizumi, A. (2007). Model-based linkage analyses confirm chromosome 19q13.3 as a susceptibility locus for intracranial aneurysm. *Stroke; a Journal of Cerebral Circulation, 38*(4), 1174-1178.

Moller, P., Clark, N., & Maehle, L. (2011). A SImplified method for segregation analysis (SISA) to determine penetrance and expression of a genetic variant in a family. *Human Mutation, 32*(5), 568-571.

Moriwaki, T., Takagi, Y., Sadamasa, N., Aoki, T., Nozaki, K., & Hashimoto, N. (2006). Impaired progression of cerebral aneurysms in interleukin-1beta-deficient mice. *Stroke; a Journal of Cerebral Circulation, 37*(3), 900-905.

MUGQIC Bioinformatics. (2014). *DNAseq report*. Provided by McGill University and Genome Quebec Innovation Centre.

Mukherjee, D., & Patil, C. G. (2011). Epidemiology and the global burden of stroke. *World Neurosurgery, 76*(6 Suppl), S85-90.

Nahed, B. V., Seker, A., Guclu, B., Ozturk, A. K., Finberg, K., Hawkins, A. A., . . . Gunel, M. (2005). Mapping a mendelian form of intracranial aneurysm to 1p34.3-p36.13. *American Journal of Human Genetics, 76*(1), 172-179.

Nasim, M. T., Ogo, T., Ahmed, M., Randall, R., Chowdhury, H. M., Snape, K. M., . . . Machado, R. D. (2011). Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. *Human Mutation, 32*(12), 1385-1389.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., . . . Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature, 461*(7261), 272-276.

NordNordWest. (2009). *Location map of Newfoundland and Labrador, Canada*. Licensed by Creative Commons (CC BY-SA 3.0). doi:http://en.wikipedia.org/wiki/template:location_map_Canada_Newfoundland_and_Labrador.

Olson, J. M., Vongpunsawad, S., Kuivaniemi, H., Ronkainen, A., Hernesniemi, J., Ryynanen, M., . . . Tromp, G. (2002). Search for intracranial aneurysm susceptibility gene(s) using finnish families. *BMC Medical Genetics, 3*, 7.

Onda, H., Kasuya, H., Yoneyama, T., Takakura, K., Hori, T., Takeda, J., . . . Inoue, I. (2001). Genomewide-linkage and haplotype-association studies map intracranial aneurysm to chromosome 7q11. *American Journal of Human Genetics, 69*(4), 804-819.

Ozturk, A. K., Nahed, B. V., Bydon, M., Bilguvar, K., Goksu, E., Bademci, G., . . . Gunel, M. (2006). Molecular genetic analysis of two large kindreds with intracranial aneurysms demonstrates linkage to 11q24-25 and 14q23-31. *Stroke; a Journal of Cerebral Circulation, 37*(4), 1021-1027.

Phillips, J. B., & Westerfield, M. (2014). Zebrafish models in translational research: Tipping the scales toward advancements in human health. *Disease Models & Mechanisms, 7*(7), 739-743.

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*, R167-72.

Regalado, E. S., Guo, D. C., Villamizar, C., Avidan, N., Gilchrist, D., McGillivray, B., . . . Milewicz, D. M. (2011). Exome sequencing identifies SMAD3 mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms. *Circulation Research, 109*(6), 680-686.

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research, 39*(17), e118.

Rossetti, S., & Harris, P. C. (2013). The genetics of vascular complications in autosomal dominant polycystic kidney disease (ADPKD). *Current Hypertension Reviews, 9*(1), 37-43.

Ruigrok, Y. M., Rinkel, G. J., van't Slot, R., Wolfs, M., Tang, S., & Wijmenga, C. (2006). Evidence in favor of the contribution of genes involved in the maintenance of the extracellular matrix of the arterial wall to the development of intracranial aneurysms. *Human Molecular Genetics, 15*(22), 3361-3368.

Ruigrok, Y. M., Rinkel, G. J., & Wijmenga, C. (2005). Genetics of intracranial aneurysms. *The Lancet Neurology, 4*(3), 179-189.

Ruigrok, Y. M., Wijmenga, C., Rinkel, G. J., van't Slot, R., Baas, F., Wolfs, M., . . . Roos, Y. B. (2008). Genomewide linkage in a large dutch family with intracranial aneurysms: Replication of 2 loci for intracranial aneurysms to chromosome 1p36.11-p36.13 and Xp22.2-p22.32. *Stroke; a Journal of Cerebral Circulation, 39*(4), 1096-1102.

Santiago-Sim, T., Mathew-Joseph, S., Pannu, H., Milewicz, D. M., Seidman, C. E., Seidman, J. G., & Kim, D. H. (2009). Sequencing of TGF-beta pathway genes in familial cases of intracranial aneurysm. *Stroke; a Journal of Cerebral Circulation, 40*(5), 1604-1611.

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews.Genetics, 12*(10), 683-691.

Schaffer, A. A. (2013). Digenic inheritance in medical genetics. *Journal of Medical Genetics, 50*(10), 641-652.

Schievink, W. I., Michels, V. V., & Piepgras, D. G. (1994). Neurovascular manifestations of heritable connective tissue disorders. A review. *Stroke; a Journal of Cerebral Circulation, 25*(4), 889-903.

Schwartzentruber, J. (2012). McGill University and Genome Quebec tutorials. Retrieved from http://gqinnovationcenter.com/services/bioinformatics/tutorials/tutorials.aspx?l=e.

Schwarz, J. M., Rodelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods, 7*(8), 575-576.

Semmler, A., Linnebank, M., Krex, D., Gotz, A., Moskau, S., Ziegler, A., & Simon, M. (2008). Polymorphisms of homocysteine metabolism are associated with intracranial aneurysms. *Cerebrovascular Diseases (Basel, Switzerland), 26*(4), 425-429.

Sengupta, S., Michener, C. M., Escobar, P., Belinson, J., & Ganapathi, R. (2008). Ovarian cancer immuno-reactive antigen domain containing 1 (*OCIAD1*), a key player in ovarian cancer cell adhesion. *Gynecologic Oncology, 109*(2), 226-233.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., . . . Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Human Mutation, 34*(1), 57-65.

Siepel, A., & Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution, 21*(3), 468-488.

Slowik, A., Borratynska, A., Turaj, W., Pera, J., Dziedzic, T., Figlewicz, D. A., . . . Szczudlik, A. (2005). Alpha1-antichymotrypsin gene (SERPINA3) A/T polymorphism as a risk factor for aneurysmal subarachnoid hemorrhage. *Stroke; a Journal of Cerebral Circulation, 36*(4), 737-740.

Smigielski, E. M., Sirotkin, K., Ward, M., & Sherry, S. T. (2000). dbSNP: A database of single nucleotide polymorphisms. *Nucleic Acids Research, 28*(1), 352-355.

Starke, R. M., Chalouhi, N., Ding, D., & Hasan, D. M. (2015). Potential role of aspirin in the prevention of aneurysmal subarachnoid hemorrhage. *Cerebrovascular Diseases (Basel, Switzerland), 39*(5-6), 332-342.

Statistics Canada. (2011). Mortality: Summary list of causes 2008. doi:Cat no. 84F0209X.

Sun, H., Zhang, D., & Zhao, J. (2008). The interleukin-6 gene -572G>C promoter polymorphism is related to intracranial aneurysms in chinese han nationality. *Neuroscience Letters, 440*(1), 1-3.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . von Mering, C. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research, 43*, D447-52.

Tang, B., Xiong, H., Sun, P., Zhang, Y., Wang, D., Hu, Z., . . . Zhang, Z. (2006). Association of PINK1 and DJ-1 confers digenic inheritance of early-onset parkinson's disease. *Human Molecular Genetics, 15*(11), 1816-1825.

Tromp, G., Weinsheimer, S., Ronkainen, A., & Kuivaniemi, H. (2014). Molecular basis and genetic predisposition to intracranial aneurysm. *Annals of Medicine, 46*(8), 597-606.

van der Voet, M., Olson, J. M., Kuivaniemi, H., Dudek, D. M., Skunca, M., Ronkainen, A., . . . Tromp, G. (2004). Intracranial aneurysms in finnish families: Confirmation of linkage and refinement of the interval to chromosome 19q13.3. *American Journal of Human Genetics, 74*(3), 564-571.

Verlaan, D. J., Dube, M. P., St-Onge, J., Noreau, A., Roussel, J., Satge, N., . . . Rouleau, G. A. (2006). A new locus for autosomal dominant intracranial aneurysm, ANIB4, maps to chromosome 5p15.2-14.3. *Journal of Medical Genetics, 43*(6), e31.

Vlak, M. H., Algra, A., Brandenburg, R., & Rinkel, G. J. (2011). Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: A systematic review and meta-analysis. *The Lancet Neurology, 10*(7), 626-636.

Walcott, B. P., & Peterson, R. T. (2014). Zebrafish models of cerebrovascular disease. *Journal of Cerebral Blood Flow and Metabolism : Official Journal of the*

International Society of Cerebral Blood Flow and Metabolism, 34(4), 571-577. doi:10.1038/jcbfm.2014.27 [doi]

Wang, 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.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164.

Wang, Y., Emeto, T. I., Lee, J., Marshman, L., Moran, C., Seto, S. W., & Golledge, J. (2014). Mouse models of intracranial aneurysm. *Brain Pathology (Zurich, Switzerland),*

Wang, Z., Liu, X., Yang, B. Z., & Gelernter, J. (2013). The role and challenges of exome sequencing in studies of human diseases. *Frontiers in Genetics, 4*, 160.

Weinsheimer, S., Goddard, K. A., Parrado, A. R., Lu, Q., Sinha, M., Lebedeva, E. R., . . . Tromp, G. (2007). Association of kallikrein gene polymorphisms with intracranial aneurysms. *Stroke; a Journal of Cerebral Circulation, 38*(10), 2670-2676.

Williams, L. N., & Brown, R. D.,Jr. (2013). Management of unruptured intracranial aneurysms. *Neurology Clinical Practice, 3*(2), 99-108.

Wills, S., Ronkainen, A., van der Voet, M., Kuivaniemi, H., Helin, K., Leinonen, E., . . . Tromp, G. (2003). Familial intracranial aneurysms: An analysis of 346 multiplex finnish families. *Stroke; a Journal of Cerebral Circulation, 34*(6), 1370-1374.

Yamada, S., Utsunomiya, M., Inoue, K., Nozaki, K., Inoue, S., Takenaka, K., . . . Koizumi, A. (2004). Genome-wide scan for japanese familial intracranial

aneurysms: Linkage to several chromosomal regions. *Circulation, 110*(24), 3727-3733.

Yan, J., Hitomi, T., Takenaka, K., Kato, M., Kobayashi, H., Okuda, H., . . . Koizumi, A. (2015). Genetic study of intracranial aneurysms. *Stroke; a Journal of Cerebral Circulation, 46*(3), 620-626. doi:10.1161/STROKEAHA.114.007286 [doi]

Yasuno, K., Bakircioglu, M., Low, S. K., Bilguvar, K., Gaal, E., Ruigrok, Y. M., . . . Gunel, M. (2011). Common variant near the endothelin receptor type A (EDNRA) gene is associated with intracranial aneurysm risk. *Proceedings of the National Academy of Sciences of the United States of America, 108*(49), 19707-19712.

Yasuno, K., Bilguvar, K., Bijlenga, P., Low, S. K., Krischek, B., Auburger, G., . . . Gunel, M. (2010). Genome-wide association study of intracranial aneurysm identifies three new risk loci. *Nature Genetics, 42*(5), 420-425.

Young, T. L., Penney, L., Woods, M. O., Parfrey, P. S., Green, J. S., Hefferton, D., & Davidson, W. S. (1999). A fifth locus for bardet-biedl syndrome maps to chromosome 2q31. *American Journal of Human Genetics, 64*(3), 900-904.

Zech, M., Lam, D. D., Francescatto, L., Schormair, B., Salminen, A. V., Jochim, A., . . . Winkelmann, J. (2015). Recessive mutations in the alpha3 (VI) collagen gene COL6A3 cause early-onset isolated dystonia. *American Journal of Human Genetics, 96*(6), 883-893.

Zhang, J., & Claterbuck, R. E. (2008). Molecular genetics of human intracranial aneurysms. *International Journal of Stroke : Official Journal of the International Stroke Society, 3*(4), 272-287.

Zimprich, A., Benet-Pages, A., Struhal, W., Graf, E., Eck, S. H., Offman, M. N., . . . Strom, T. M. (2011). A mutation in VPS35, encoding a subunit of the retromer

complex, causes late-onset parkinson disease. *American Journal of Human Genetics, 89*(1), 168-175.

Appendices

Appendix A: Promega Wizard® Genomic DNA Extraction

Materials:

1) Blood collected in EDTA tubes
2) Laminar flow biosafety containment cabinet
3) Centrifuge
4) Vortex
5) 50 ml sterile centrifuge tubes
6) 1.5 ml sterile microcentrifuge tubes
7) Promega Wizard® Genomic Purification Kit Cat #PRA-1620
8) Isopropanol
9) 70% Ethanol
10) Disposable 15 ml pipettes
11) Motorized pipette

Protocol for 12-16 ml of whole blood:

1) Add 30 ml of Cell Lysis Solution to a 50 ml sterile centrifuge tube.
2) Gently rock the tubes of blood until mixed thoroughly. Add 12-16 ml of whole blood to the centrifuge tube containing the Cell Lysis Solution. Invert the tube 5-6 times.
3) Incubate the mixture at room temperature for 10 minutes (invert 5-6 times halfway through incubation).
4) Centrifuge the mixture at 2000 g for 10 minutes.
5) Remove and discard as much supernatant as possible without disturbing the pellet at the bottom of the tube.
6) Vortex the tube, add 10 ml of Nuclei Lysis Solution, and vortex again for 20 seconds.
7) Add 3.3 ml of Protein Precipitation Solution. Using a motorized pipette, mix the solution 5-6 times.
8) Centrifuge at 2000 g for 10 minutes.
9) Add 10 ml of Isopropanol to a new sterile 50 ml centrifuge tube.
10) Gently pour the supernatant into the centrifuge tube containing the Isopropanol.
11) Centrifuge at 2000 g for 2 minutes.
12) Wash with 70% ethanol and repeat step 11.
13) Let air dry.
14) Add 400 µl of Rehydration Buffer, let sit overnight.
15) Mix sample briefly and quantify concentration.
16) DNA sample can be stored at 4ºC or frozen for an extended period.

Appendix B: Primer Sequences and PCR Protocols

Primer Name	Primer Sequence (5'- 3')	Product Size	PCR Protocol Used
ATP1A4_P600S_F	CTGGGGTGAGAAATCAAGGA	382 bp	Standard
ATP1A4_P600S_R	AGGAAACAGAAATCCGCTCA		(62°/30 cycles)
C4ORF6_start_F	GCCAGCCTCCTACCTCAAAT	364 bp	Standard
C4ORF6_start_R	GAGCACCTTCCGACTCACTC		(62°/30 cycles)
CCDC3_L73V_F	GGCCGAGACCATCGTGTA	216 bp	Standard
CCDC3_L73V_R	GAGTGGCACGAGAAGTAGCC		(62°/30 cycles)
CCDC3_FS_F	GGAGCGAGTGAAGAAGGTCA	236 bp	Standard
CCDC3_FS_R	GCTGCATGTACGAAACCTCA		(60°/30 cycles)
DOPEY1_A2301V_F	CAAACATTCCACCCGCCTTG	590 bp	Standard
DOPEY1_A2301V_R	CCGATAACTCTGCTGTGCCA		(68°/30 cycles)
GIGYF2_H1165R_F	TCCCCAAATTACTTCAGCCTT	529 bp	Standard
GIGYF2_H1165R_R	GGTACCGCATACACCACACT		(62°/30 cycles)
HSPBP1_G26GAAD_F	GACTTTGGGGAAGGGTCCTG	511 bp	Touchdown-A
HSPBP1_G26GAAD_R	CTCCCCAAGTCACACTTCCC		(T1: 64.5°, 10 cycles, T2: 60°, 25 cycles)
KNDC1_A896S_F	GGGGAACGTGATGACCAGAG	520 bp	Standard
KNDC1_A896S_R	TGCCCGTTGACCACCTTAAA		(67.9°/35 cycles)
MAML3_F	CAGTCCCCTCAAACTCCAAA	394 bp	Standard
MAML3_R	AGGCTTGGGGGTACATCATT		(62°/30 cycles)
MTG1_P204L_F	TGGGAGCAGAAGACAAGCTG	570 bp	Standard
MTG1_P204L_R	CTGGTCCGTCAAATGGGGAA		(66°/30 cycles)
MUC4_D997N_F	TGTCACCTTCAGGGTCTGGT	850 bp	N/A
MUC4_D997N_R	GCGGAAGGGATGGTTACA		
MUC16_G13530S_F	GAGCAGTGGGGTTTCTCTCC	525 bp	Standard
MUC16_G13530S_R	GAGGTGGTGGGAACAGGAAG		(66°/30 cycles)
MYO18A_F	GGCCATTCGCTGTGTACAGA	395 bp	Standard
MYO18A_R	GGCATGTCCCCAATAGCAGA		(62°/30 cycles)
OCIAD1_F	TCAGTCTGTAACGGCAGGTG	186 bp	Standard
OCIAD1_R	CCATAACGGCATCCTTCCTA		(60°/30 cycles)
POU3F1_A28-_F	ACCACCGCGCAGTACCTG	585 bp	N/A
POU3F1_A28-_R	CTCGTGGCCATCCTCGTG		
RP1L1_L68F_F	TGGAGTGGAGCACATTTGGG	512 bp	Standard
RP1L1_L68F_R	GAGCAGTGGGGTTTCTCTCC		(68°/30 cycles)
SPDYE4_P35S_F	ATTATGGCCAGTGGTCAAGC	311 bp	Standard
SPDYE4_P35S_R	TCATTGCTCCCCAGACTTTC		(62°/30 cycles)
TRPA1_F	AATGGATGAAGACAACGATGG	242 bp	Standard
TRPA1_R	ACGCCCATAACTTGGAAAAA		(62°/30 cycles)
ZFPM1_F	CCGTTTCAGCCTTCGCTCTA	523 bp	N/A
ZFPM1_R	ACGTACTGCGGAAGGAACAG		
ZPBP2_T208S_F	TGCACATGGAATTCAGCACT	297 bp	Touchdown-A
ZPBP2_T208S_R	CTTGAGCCCAGGAGTTTGAG		(T1: 67°, 15 cycles, T2: 60°, 19 cycles)

Appendix C: Thermocycler Protocols

1. ExoSap

Step	Temperature	Duration
1	37 °C	30 minutes
2	80 °C	15 minutes
3	4 °C	Hold

2. ABISeq

Step	Temperature	Duration
1	96 °C	1 minutes
2	96 °C	10 seconds
3	50 °C	5 seconds
4	60 °C	4 minutes
5	Return to Step 2 (34 cycles)	
6	4 °C	Hold

3. Denature

Step	Temperature	Duration
1	95 °C	2 minutes
2	4 °C	Hold

4. Standard (PCR)

Step	Temperature	Duration
1	95 °C	2 minutes
2	95 °C	30 seconds
3	Annealing Temp. (varies)	30 seconds
4	72 °C	1 minute
5	Return to Step 2 (# cycles varies)	
6	72 °C	10 minutes
7	4 °C	Hold

5. Touchdown-A (PCR)

Step	Temperature	Duration
1	95 °C	2 minutes
2	95 °C	20 seconds
3	Annealing Temp. 1 (varies)	45 seconds
4	72 °C	45 seconds
5	Return to Step 2 (10-15 cycles, decreasing at 0.5 °C increments)	
6	95 °C	20 seconds
7	Annealing Temp. 2 (varies)	45 seconds
8	72 °C	45 seconds
9	Return to Step 5 (15-20 cycles)	
10	72 °C	5 minutes
11	4 °C	Hold

Appendix D: Additional Moderate Impact Variants in Family R1256

1. Variants that passed filtering criteria, were detected in 5/5 exomes, and were previously reported in dbSNP:

Gene	Chr. Position	Status in dbSNP (*rs#*)	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
KIAA0040	1q25.1	*rs150137790*	c.217_225delAAGAAGAAG p.K73_K75del Inframe Del	3 / 2 / 0	0.12% (1/848)	0.4964% (83/16720)
LSG1	3q29	*rs114485048*	c.1039C>T p.R347W Missense	5 / 0 / 0	0.94% (8/848)	0.1805% (222/122962)
SIM1	6q16.3	*rs145479047*	c.1082C>T p.T361I Missense	5 / 0 / 0	0% (0/848)	0.2145% (263/122600)
ZMIZ2	7p13	*rs189007540*	c.1640G>A p.S547N Missense	5 / 0 / 0	0% (0/848)	0.09167% (112/122174)
PSPH	7p11.2	*rs75395437*	c.268G>A p.G90S Missense	5 / 0 / 0	0% (0/848)	0.514% (608/118296)
PSPH	7p11.2	*rs73343757*	c.249A>C p.Q83H Missense	5 / 0 / 0	0% (0/848)	0.8515% (997/117092)
ANKRD30A	10p11.21	*rs202149101*	c.1232G>T p.R411M Missense	5 / 0 / 0	0% (0/848)	0.06915% (83/120022)
GPR123	10q26.3	*rs144212964*	c.1283G>A p.R428Q Missense	5 / 0 / 0	0.24% (2/848)	0.2666% (45/16882)
ZFPM1	16q24.2	*rs149145771*	c.1334_1339delCTCTGG p.L446_A447del Inframe Del	0 / 5 / 0	0.35% (3/848)	0% (0/118)
PHF12	17q11.2	*rs200985028*	c.1473C>G p.H491Q Missense	4 / 1 / 0	0% (0/848)	0.1009% (124/122950)
MYO18A	17q11.2	*rs777985641*	c.3704C>T p.P1235L Missense	2 / 3 / 0	0.12% (1/848)	0.000828% (1/120720)
KANSL1	17q21.31	*rs74867664*	c.2698G>A p.G900R Missense	5 / 0 / 0	0.24% (2/848)	0.1078% (132/122452)
DGCR6	22q11.21	*rs146390355*	c.392G>A p.R131H Missense	5 / 0 / 0	0.71% (6/848)	0.3033% (355/117050)

2. Variants that passed filtering criteria, were detected in 4/5 exomes, and were previously reported in dbSNP:

Gene	Chr. Position	Status in dbSNP (*rs#)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
KIAA0319L	1p34.3	*rs144951042*	c.700A>G p.T234A Missense	4 / 0 / 1	0% (0/848)	0.0838% (103/122912)
DYSF	2p13.2	*rs771062534*	c.1609G>A p.V537M Missense	4 / 0 / 1	0% (0/848)	0.006109% (5/81840)
FAHD2A	2q11.1	*rs200437887*	c.281C>T p.S94L Missense	4 / 0 / 1	0% (0/848)	0.06028% (71/117782)
RNF149	2q11.2	*rs143827530*	c.416A>G p.N139S Missense	4 / 0 / 1	0% (0/848)	0.1776% (199/112024)
FANCD2	3p25.3	*rs35625434*	c.983G>A p.R328Q Missense	4 / 0 / 1	0.35% (3/848)	0.3032% (480/122080)
SLC15A2	3q13.33	*rs748018463*	c.1301A>G p.H434R Missense	4 / 0 / 1	0.12% (1/848)	0.005767% (7/121370)
MUC4	3q29	*rs779985296*	c.7189G>C p.D2397H Missense	4 / 0 / 1	0% (0/848)	0.04447% (11/24736)
ABLIM2	4p16.1	*rs370025292*	c.481G>A p.V161I Missense	4 / 0 / 1	0% (0/848)	0.004107% (5/121740)
FOXP4	6p21.1	*rs41273784*	c.97G>A p.G33R Missense	4 / 0 / 1	0.59% (5/848)	0.3616% (437/120854)
GTF2E2	8p12	*rs2978277*	c.548A>G p.K183R Missense	4 / 0 / 1	0.71% (6/848)	0.3661% (449/122660)
WRN	8p12	*rs78488552*	c.3785C>G p.T1262R Missense	4 / 0 / 1	0.59% (5/848)	0.2753% (338/122784)
BRINP1	9q33.1	*rs142894245*	c.1046C>T p.T349M Missense	4 / 0 / 1	0.24% (2/848)	0.2099% (258/122902)
ANTXRL	10q11.22	*rs148029033*	c.509A>G p.N170S Missense	4 / 0 / 1	0.47% (4/848)	0.07254% (12/16542)

AGAP7	10q11.22	*rs201609209*	c.362T>C p.L121S Missense	3 / 1 / 1	0.59% (5/848)	0.3333% (219/65700)
MYOF	10q23.33	*rs201634420*	c.4534C>T p.R1512W Missense	4 / 0 / 1	0.83% (7/848)	0.2305% (262/113648)
LRRC56	11p15.5	*rs138291757*	c.655G>A p.V219M Missense	4 / 0 / 1	0.59% (5/848)	0.6479% (105/16206)
LDHA	11p15.1	*rs34305721*	c.434G>C p.G145A Missense	4 / 0 / 1	0.59% (5/848)	0.2521% (310/122960)
IFITM5	11p15.5	*rs747064580*	c.268G>A p.A90T Missense	4 / 0 / 1	0% (0/848)	0.004152% (5/120416)
CPT1A	11q13.3	*rs80356779*	c.1436C>T p.P479L Missense	4 / 0 / 1	0.47% (4/848)	0.001627% (2/122898)
C11orf30	11q13.5	*rs184345272*	c.2861T>G p.L954R Missense	3 / 1 / 1	0.12% (1/848)	0.1952% (237/121386)
MOGAT2	11q13.5	*rs373540522*	c.299C>A p.S100Y Missense	3 / 1 / 1	0% (0/848)	0.01155% (14/121212)
GSG1	12p13.1	*rs148537880*	c.391G>A p.E131K Missense	4 / 0 / 1	0% (0/848)	0.01546% (19/122894)
OR11H4	14q11.2	*rs142720326*	c.55G>A p.V19M Missense	4 / 0 / 1	0.24% (2/848)	0.1199% (147/122594)
LTBP2	14q24.3	*rs201591982*	c.2657C>A p.T886K Missense	4 / 0 / 1	0.24% (2/848)	0.07206% (70/97140)
NIPA1	15q11.2	*rs549007670*	c.40_48delGCGGCGGCC p.A14_A16del Inframe Deletion	4 / 0 / 1	0% (0/848)	0.05244% (2/3814)
MBP	18q23	*rs149668522*	c.194C>T p.P65L Missense	4 / 0 / 1	0% (0/848)	0.2069% (76/36732)
ZNF816	19q13.41	*rs61740548*	c.1768G>C p.E590Q Missense	3 / 1 / 1	0% (0/848)	0.4302% (522/121346)

3. Variants that passed filtering criteria, were detected in 4/5 exomes, and were not previously reported in dbSNP:

Gene	Chr. Position	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC Controls (/848)	MAF in ExAC Browser (/alleles)
AIM1L	1p36.11	c.28C>G p.R10G Missense	4 / 0 / 1	0% (0/848)	0.004735% (1/21118)
C2orf62	2q35	c.222G>C p.E65Q Missense	4 / 0 / 1	0% (0/848)	N/A
ADPRH	3q13.33	c.91G>A p.E31K Missense	4 / 0 / 1	0% (0/848)	N/A
MUC4	3q29	c.7183A>T p.T2395S Missense	4 / 0 / 1	0% (0/848)	0.003966% (1/25216)
MUC4	3q29	c.3053C>G p.S1018C Missense	4 / 0 / 1	0.12% (1/848)	0.003766% (1/26552)
MGST2	4q31.1	c.44C>T p.S15L Missense	4 / 0 / 1	0% (0/848)	N/A
MCC	5q22.2	c.60_61insTCGGCGGCA p.G20_G21insSAA Inframe Insertion	4 / 0 / 1	0% (0/848)	N/A
LAMA4	6q21	c.1546G>A p.V516M Missense	4 / 0 / 1	0% (0/848)	N/A
TRPA1	8q21.11	c.1309G>A p.V437M Missense	4 / 0 / 1	0% (0/848)	N/A
CCDC178	18q12.1	c.764C>G p.A255G Missense	4 / 0 / 1	0.12% (1/848)	N/A
KIAA1210	Xq24	c.263C>A p.P88Q Missense	1 / 3 / 1	0% (0/848)	0.002018% (1/49550)

Abbreviations: Het = heterozygous, Hom = homozygous

Appendix E: Additional Low Impact Variants in Family R1256

Variants that passed filtering criteria, and were detected in 4/5 exomes:

Gene	Chr. Position	Status in dbSNP (*rs #)*	Variant Details	# Het / # Hom for Minor Allele / # Hom for Reference Allele	MAF in MUGQIC controls (/848)	MAF in ExAC Browser (/alleles)
POU3F1	1p34.3	*rs201037684*	c.1272A>T p.A424	4 / 0 / 1	0.47% (4/848)	0% (0/974)
B3GALNT2	1q42.3	Unreported	c.1290G>A p.S430	4 / 0 / 1	0.12% (1/848)	N/A
FAM110C	2p25.3	*rs761388273*	c.168G>A p.R56	3 / 1 / 1	0% (0/848)	N/A
SNRNP200	2q11.2	*rs139731897*	c.5664C>T p.H1888	4 / 0 / 1	0.71% (6/848)	0.2891% (351/121410)
LONRF2	2q11.2	*rs148009215*	c.1437C>T p.H479	4 / 0 / 1	0.24% (2/848)	0.1071% (130/121412)
MRPS9	2q12.1	*rs149463519*	c.711G>A p.E237	4 / 0 / 1	0.24% (2/848)	0.1254% (152/121232)
CHCHD5	2q14.1	*rs146873532*	c.72G>A p.A24	4 / 0 / 1	0% (0/848)	0.003301% (4/121184)
POTEF	2q21.1	*rs201958629*	c.2172C>T p.D724	4 / 0 / 1	0.12% (1/848)	0.02214% (13/58718)
LANCL1	2q34	*rs147186536*	c.843G>A p.G281	4 / 0 / 1	0% (0/848)	0.001664% (2/120184)
IL17RC	3p25.3	*rs181990653*	c.1716C>G p.G572	4 / 0 / 1	0.35% (3/848)	0.5044% (468/92792)
MUC4	3q29	*rs74612617*	c.5820T>C p.T1940	4 / 0 / 1	0.59% (5/848)	0.9803% (234/23870)
DDX60L	4q32.3	*rs200379104*	c.1410G>A p.P470	3 / 1 / 1	0.24% (2/848)	0.006123% (6/97984)
LRRC69	8q21.3	*rs138518526*	c.738C>T p.N246	4 / 0 / 1	0% (0/848)	0.07696% (17/22088)
UCMA	10p13	Unreported	c.204C>T p.S68	4 / 0 / 1	0% (0/848)	N/A
WDFY4	10q11.23	Unreported	c.4068G>A p.G1356	4 / 0 / 1	0.47% (4/848)	N/A
ZSWIM8	10q22.2	*rs201945010*	c.1164C>T p.S388	4 / 0 / 1	0% (0/848)	0.0185% (22/118912)
MUC5B	11p15.5	*rs75760167*	c.3372T>C p.C1124	4 / 0 / 1	0% (0/848)	N/A
		rs74763753	c.3390T>C p.A1130	4 / 0 / 1	0% (0/848)	0.004034% (1/24790)
		rs79585387	c.3411C>T p.H1137	4 / 0 / 1	0% (0/848)	N/A

CTSC	11q14.2	*rs181685520*	c.-45C>G (Start-gained)	4 / 0 / 1	0.24% (2/848)	0.3819% (55/14402)
SIK2	11q23.1	*rs200427353*	c.1386C>T p.A462	4 / 0 / 1	0.12% (1/848)	0.02728% (33/120976)
SIK3	11q23.3	*rs560511616*	c.1629G>A p.Q543	4 / 0 / 1	0.12% (1/848)	0.02797% (25/89388)
PRB4	12p13.2	*rs146939904*	c.408C>A p.G136	4 / 0 / 1	59% (5/848)	0.3851% (464/120478)
HNRNPA1	12q13.13	*rs536130883*	c.462C>T p.D154	4 / 0 / 1	0.12% (1/848)	0.006657% (8/120182)
PAPLN	14q24.2	*rs145397376*	c.2718C>T p.D906	4 / 0 / 1	0.12% (1/848)	0.6774% (815/120312)
MEF2A	15q26.3	*rs367780642*	c.1071G>A p.P357	4 / 0 / 1	0.47% (4/848)	0.1052% (21/19956)
PKD1L2	16q23.2	Unreported	c.45C>T p.D15	4 / 0 / 1	0% (0/848)	N/A
NLRP1	17p13.2	Unreported	c.3510G>A p.V1170	4 / 0 / 1	0% (0/848)	N/A
MUC16	19p13.2	*rs200235837*	c.39564C>T p.G13188	4 / 0 / 1	0% (0/848)	N/A
ZSCAN22	19q13.43	*rs756431968*	c.162C>T p.H54	4 / 0 / 1	0% (0/848)	N/A

Abbreviations: Het = heterozygous, Hom = homozygous

Appendix F: IA-Related Keywords and Abbreviations

Phenotype:

- intracranial aneurysm (ICA; IA)
- cerebral aneurysm
- subarachnoid hemorrhage (SAH)
- cerebral hemorrhage
- stroke
- hemorrhagic stroke
- cerebrovascular disease
- hypertension
- Ehlers-Danlos syndrome
- polycystic kidney disease (PCKD)
- aortic aneurysm
- abdominal aortic aneurysm (AAA)

Physiology:

- artery
- blood vessel
- brain
- blood pressure
- vascular smooth muscle cells (VSMC)
- vascular endothelial cells
- extracellular matrix (ECM)

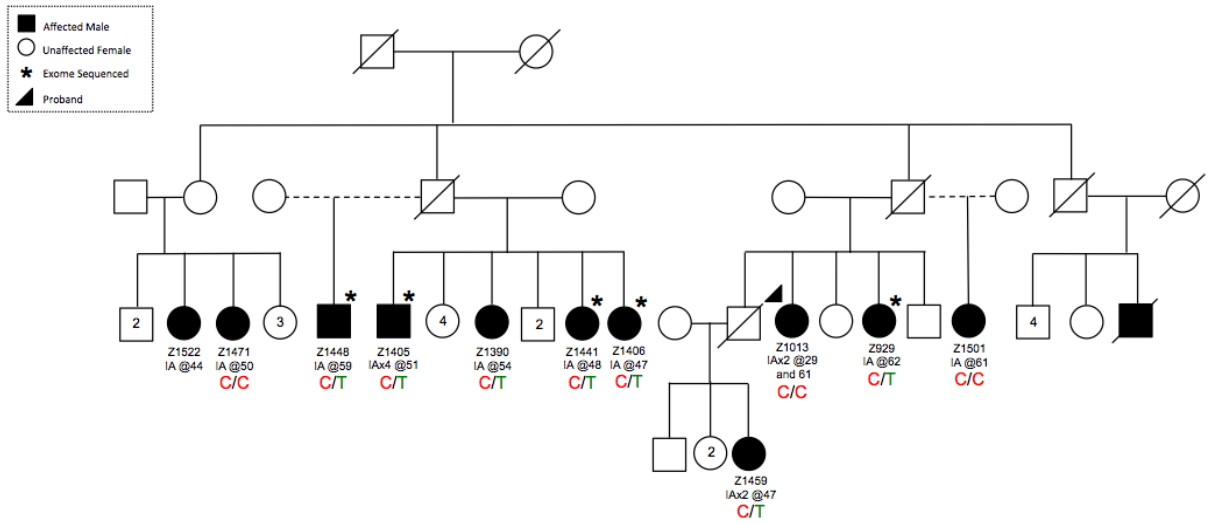
Pathways & Proteins:

- inflammation
- hemodynamic stress
- vasculature development
- vascular process in circulatory system
- endothelin
- collagen
- elastin
- fibrillin

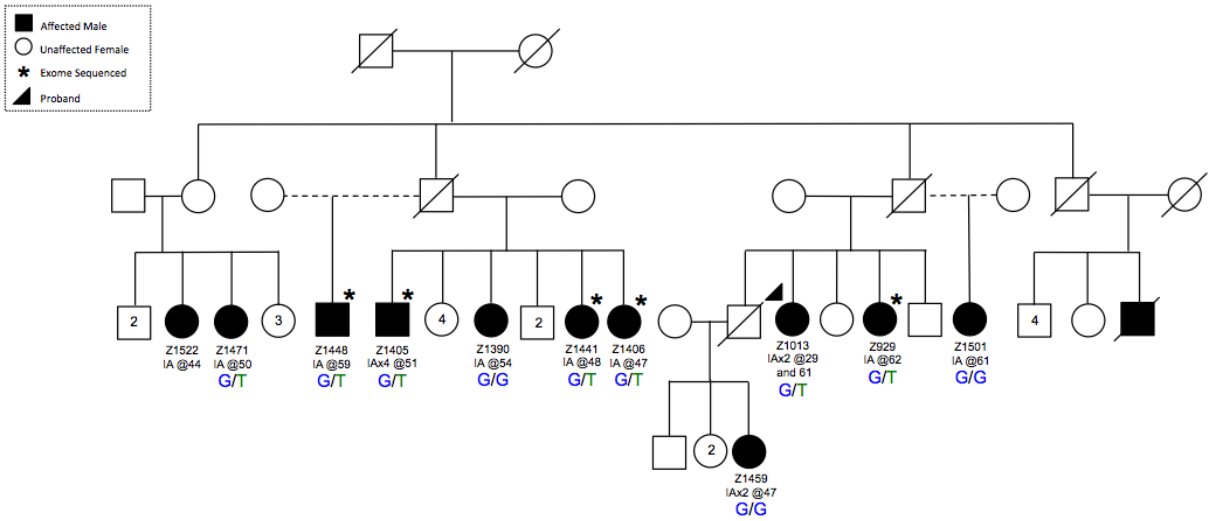
Appendix G: Additional Pedigrees for Family R1256

Validated Moderate Impact Variants:

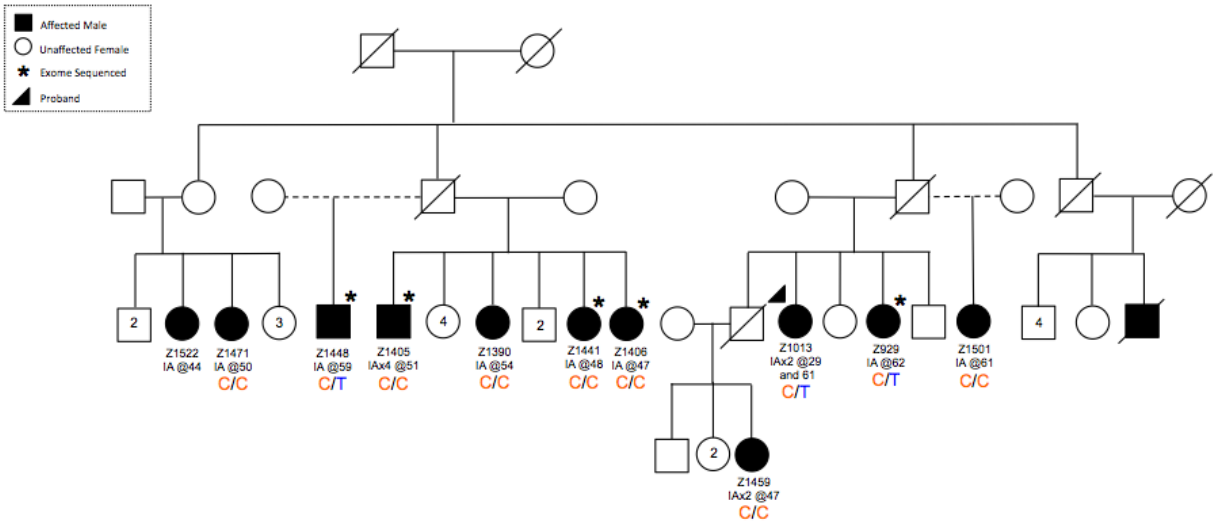
(A) *DOPEY1* c.6902C>T



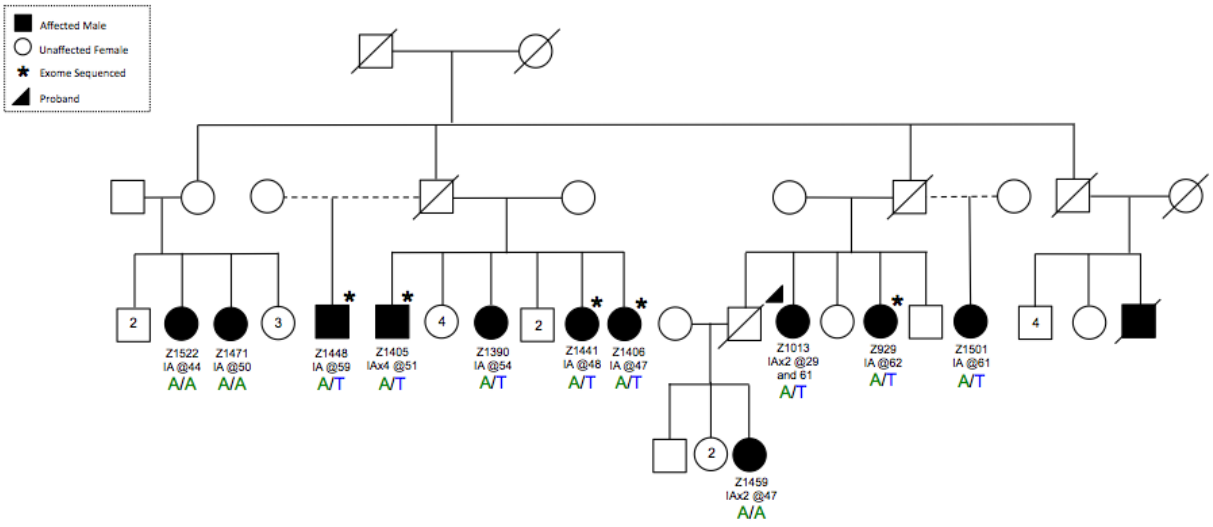
(B) KNDC1, c.2686G>T



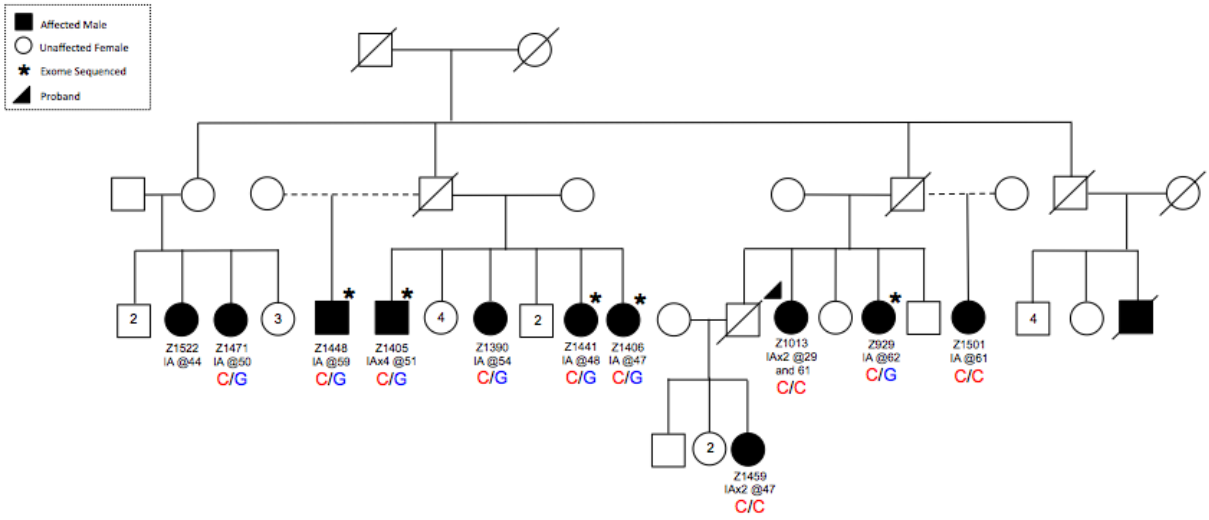
(C) *MTG1*, c.611C>T



(D) *ZPBP2*, c.622A>T

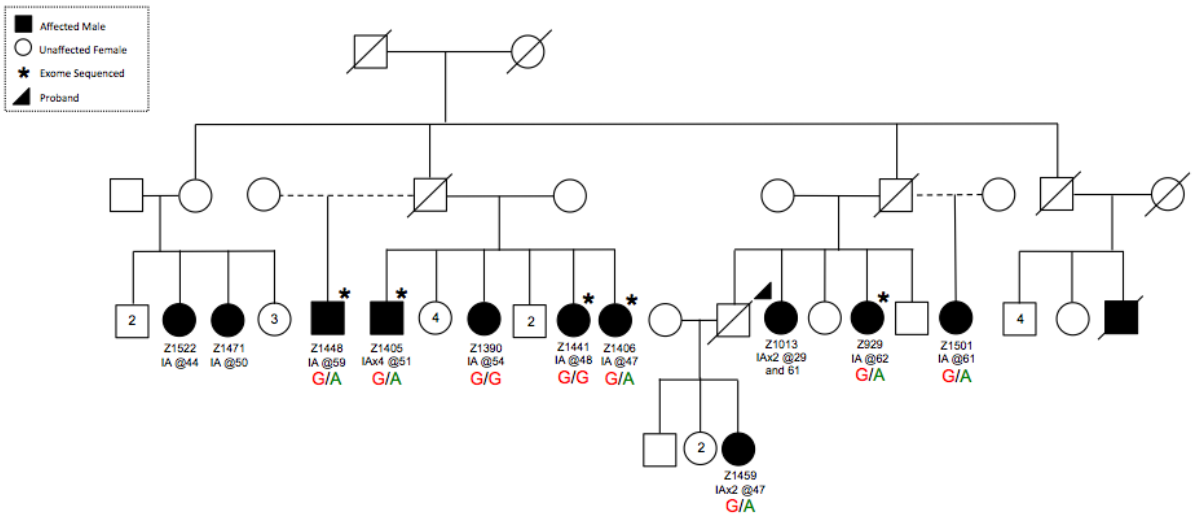


(E) *CCDC3*, c.217C>G

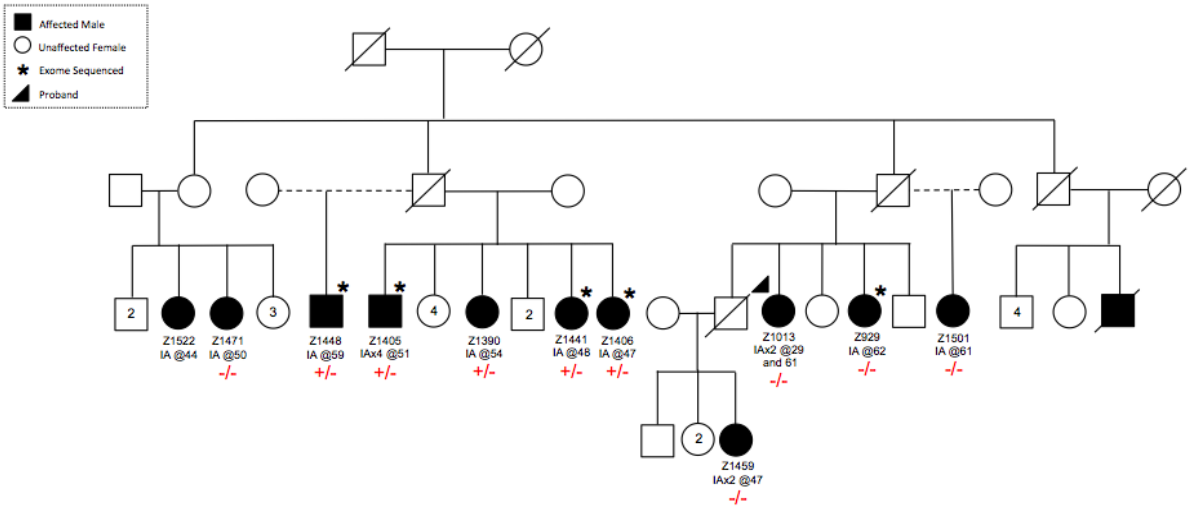


Validated High Impact Variants:

(F) *OCIAD1*, c.-6+1G>A



(G) *CCDC3*, c.425delA



The symbol +/+ is used to indicate individuals that are heterozygous for the deletion, and -/- is used to indicate the absence of the variant.

Appendix H: Functional Annotation of Moderate Impact Candidates in Family R1256

Gene	Variant Details	GERP	Polyphen 2	SIFT	Gene Description
DOPEY1	c.6902C>T; p.A2301V	5.77	D	1	Dopey family member 1; may be involved in protein transport between golgi and endosomes
KNDC1	c.2686G>T; p.A896S	3.83	D	0.06	Kinase non-catalytic C-lobe domain containing 1; is a guanine nucleotide exchange factor, has high expression in the brain
MTG1	c.611C>T; p.P204L	5.59	D	0	Mitochondrial ribosome-associated GTPase 1; involved in GTP binding, which is involved in signal transduction processes
ZPBP2	c.622A>T; p.T208S	5.67	B	0.35	Zona pellucida-binding protein 2; has expression in testis and brain medulla
CCDC3	c.217C>G; p.L73V	4.33	D	0.06	Coiled-coil domain containing 3; expressed in adipose tissue and endothelial cells

GERP scores above 3.00 are considered to be predictive of evolutionary conservation. For Polyphen2 scores, D=damaging, and B=benign. For SIFT scores, values below or equal to 0.05 are considered to be damaging, while values from 0.05 to 1 are benign or "tolerated" changes. Gene functions and pathway annotations are collected from UniProt, OMIM, DAVID and GeneCards.