WHOLE EXOME SEQUENCING FOR THE IDENTIFICATION OF NOVEL SUSCEPTIBILITY GENES RELATED TO FAMILIAL PULMONARY FIBROSIS IN A NEWFOUNDLAND COHORT

by

Robyn Byrne

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

Discipline of Genetics, Faculty of Medicine

Memorial University of Newfoundland

May 2015

St. John's Newfoundland

Abstract

Idiopathic pulmonary fibrosis (IPF) is a multifactorial, interstitial lung disease (ILD) which leads to the scarring and fibrosis of the alveolar interstitium. In the province of Newfoundland and Labrador, the prevalence of familial pulmonary fibrosis (FPF), in which two or more first degree relatives are affected, is high and consistent with strong genetic components segregating in this population. The study uses next generation sequencing to identify novel susceptibility genes for idiopathic pulmonary fibrosis. DNA samples from 24 patients from 14 different FPF families were analysed using whole exome sequencing. Of the 14 families sequenced, two families were selected for further analysis, R0942 and R1136. Using a filtering strategy that annotated genetic variants based on prevalence in variant databases and predicted phenotypic outcome using bioinformatics programs, a list of candidate gene variants was created. Furthermore, these variants were filtered based on functional gene annotation. Of interest were rare variants found in the genes *CD109* and telomeric repeat-binding factor 1(*TERF1)* in families R1136 and R0942, respectively, that passed filtering criteria. The variants in *CD109* (c.1474C>T; p.R492X) and *TERF1*, (c.311G>T; p.S104I) are thought to be involved in the regulation of the telomerase protein complex, whose reduced activity has been implicated in the development of IPF. Although neither variant completely segregated with the disease, several *in silico* programs support their pathogenicity and the variants appear to be rare in the general population. Functional assays of these variants will be required to accurately determine their phenotypic effects.

Acknowledgment

I would like to first and foremost thank my supervisor, Dr. Michael Woods, for his support and guidance throughout this project. I have learned a wealth of knowledge surrounding research techniques and data analysis which will be useful in my future career. I would also like that thank my supervisory committee, Drs. Bridget Fernandez and Roger Green, for their continuous support and input with this project. Special thanks to Ms. Barbara Noble and Dr. Fernandez for their help in ascertaining patients and providing the clinical data for this project. Furthermore, I would like to thank all the patients enrolled in the study, without whom we would be unable to conduct research. I would like to acknowledge CIHR and RDC for their financial support in funding this project. A big thank you to Krista Mahoney, Daniel Evans and Amy Powell, for their assistance with this study, the editing process of my thesis and with the general upkeep of the lab. Thank you to Ms. Deborah Quinlan for her help throughout the past two years of my program, as well as all members of the Genetics Department whom have helped me both directly and indirectly. Thank you to Ms. Jennifer Maclean and all staff in the Provincial Medical Genetics Program for allowing me to volunteer in the clinic this year. My experience in a clinical setting has not only prepared me for my future profession as a genetic counsellor, but provided me with clinical insight which was applied to the current project. Lastly, and most importantly, I would like to thank my parents for providing me with many opportunities throughout my life to pursue my academic endeavours. Their commitment to my education has given me the confidence and determination to accomplish the many goals I have set out before me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Adenine (A)

Alveolar Epithelial Cell (AEC)

American Thoracic Society / European Respiratory Society (ATS/ERS)

Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 (ARVCD)

Burrows-Wheeler Aligner (BWA)

Centimorgan (cM)

Cytokine Induced Apoptosis Inhibitor 1 (CIAPIN1)

Cytosine (C)

Chronic Obstructive Pulmonary Disease (COPD

Database for Annotation, Visualization and Integrated Discovery (DAVID)

Desmoplakin (DSP)

Deoxyribonucleic Acid (DNA)

Dyskeratosis Congenita (DKC)

Dyskeratosis Congenital-1 (DKC1)

Exonuclease (EXO)

Extracellular Matrix (ECM)

Familial Interstitial Pneumonia (FIP)

Familial Pulmonary Fibrosis (FPF)

Fibroblast Growth Factor Receptor 4 (FGFR4)

Finding of Rare Disease Genes (FORGE)

Genome Analysis Toolkit (GATK)

Genome Wide Association Study (GWAS)

Gene Ontology (GO)

Genomic DNA (gDNA)

Genomic Evolutionary Rating Profile (GERP)

Glycosylphosphatidylinositol **(**GPI**)**

Guanine (G)

Guanine/ Cytosine (GC)

Heterogeneity LOD scores (HLOD)

High Resolution Computed Topography (HRCT)

Human Leukocyte Antigen (HLA)

Idiopathic Interstitial Pneumonia (IIP)

Idiopathic Pulmonary Fibrosis (IPF)

Insertion/Deletion (INDEL)

Interleukin 32 (IL32)

Interstitial Lung Diseases (ILD)

Kilobase (Kb)

Length of Read (L)

Logarithm of Odds (LOD)

Memorial University of Newfoundland (MUN)

Mucin 3 A (MUC3A)

Mucin 5 B (MUC5B)

Megabase (Mb)

Minor Allele Frequency (MAF)

National Heart, Lung and Blood Institute (NHLBI)

Newfoundland Colorectal Cancer Research (NFCCR)

Next Generation Sequencing (NGS)

NOP10 ribonucleoprotein (NOP10)

Nucleotide Triphosphates (dNTPs)

Number of Reads at a Given Locus (N)

Original Length of Genome/ Exome (O)

Probability (P)

Polymerase Chain Reaction (PCR)

Pulmonary Function Test (PFT)

Phred Quality Score (Q)

Quality Control (QC)

Shrimp Alkaline Phosphatase (SAP)

Single Nucleotide Polymorphism (SNP)

Single Nucleotide Polymorphism Database (dbSNP)

Sorting Intolerant from Tolerant (SIFT)

Surfactant Protein A1 (SFTPA1)

Surfactant Protein A2 (SFTPA2)

Surfactant Protein C (SFTPC)

Telomerase Protein Component 1 (TEP1)

Telomerase RNA component (TERC)

Telomerase Reverse Transcriptase (*TERT)*

Telomere Maintenance Interacting Protein 1 (TTI1)

Telomeric Repeat-Binding Factor 1 (TERF1)

Thymine (T)

Transforming Growth Factor- Beta (TGF-β)

Transforming Growth Factor- Beta Receptor (TGFβR)

United States of America (USA)

Usual Interstitial Pneumonia (UIP)

Variant Call Format (VCF)

Whole Genome Sequencing (WGS)

List of Appendices

1. Introduction

1.1 Interstitial Lung Disease

1.1.1 Classifications of Interstitial Lung Diseases

A large proportion of debilitating pulmonary disorders falls under the classification of interstitial lung disease (ILD), a subset of pulmonary diseases that affects the alveolar interstitium, including the tissues surrounding the air sacs in the lung. Many of the over 200 subtypes of ILDs, with varying etiologies (Steele et al., 2013), share common symptoms and phenotypes. Because of overlapping clinical findings in many of these pulmonary diseases, a combination of histological, radiological and pulmonary examinations are required to properly classify ILDs. As shown in Figure 1, there are four main classifications of ILD defined by the American Thoracic Society and European Respiratory Society (ATS/ERS) including: 1) ILDs of a known cause, typically associated with environmental exposures, such as silicosis; 2) idiopathic interstitial pneumonias (IIPs); 3) granulomatous ILDs, including sarcoidosis; and 4) rare ILDs with distinct clinical findings. IIPs are one of the most common forms of ILDs, and can be further categorized into major IIPs, unclassified IIPs and rare IIPs. Major IIPs are additionally sub-categorized into idiopathic pulmonary fibrosis (IPF), which compromises nearly 71% of all IIP cases (Raghu et al., 2011), idiopathic nonspecific interstitial pneumonia, respiratory bronchiolitis ILDs, desquamative interstitial pneumonia, cryptogenic organizing pneumonia and unclassified major IIPs (Steele et al., 2013). Furthermore, families with two or more first degree relatives with IPF are

classified as having familial pulmonary fibrosis (FPF) (Lee et al., 2005), while individuals with no affected first degree relative are termed sporadic.

1.1.2 Idiopathic Pulmonary Fibrosis

IPF is an adult onset lung disease characterized by the scarring and fibrosis of the alveolar interstitium, with histological findings similar to usual interstitial pneumonia (UIP). Symptoms of IPF typically appear between 50-70 years of age in the form of dyspnea and non-productive coughing due to aggregates of interstitial infiltrates. As IPF progresses, inadequate arterial oxygen diffusion leads to hypoxia, respiratory insufficiency and ultimately death. Although a variety of drugs have been tested in clinical trials, including pirfenidone, N-acetylcysteine, etanercept, and bostentan, success in preventing the progression of IPF in patients has been limited (Garber, 2013). The only treatment demonstrated to prolong survival is lung transplantation, without which, death typically occurs between three and five years post-diagnosis (ATS/ERS, 2013). In sporadic IPF cases, the underlying cytology is predicted multifactorial. Contributing factors are believed to include exposure to environmental irritants, aberrant wound healing and genetic predisposition. The current estimated prevalence of IPF the general population ranges between 2-29 per 100,000, with a 14-27.9 in 100,000 prevalence in the United States of America (USA) (Nalysnyk et al., 2012). Newfoundland is shown to have similar prevalence of IPF compared to other populations (13.22 per 100,000; Raghu et al., 2006); however, the proportion of familial cases is estimated to be higher than in more admixed populations at approximately 36% (Fernandez et al., 2012).

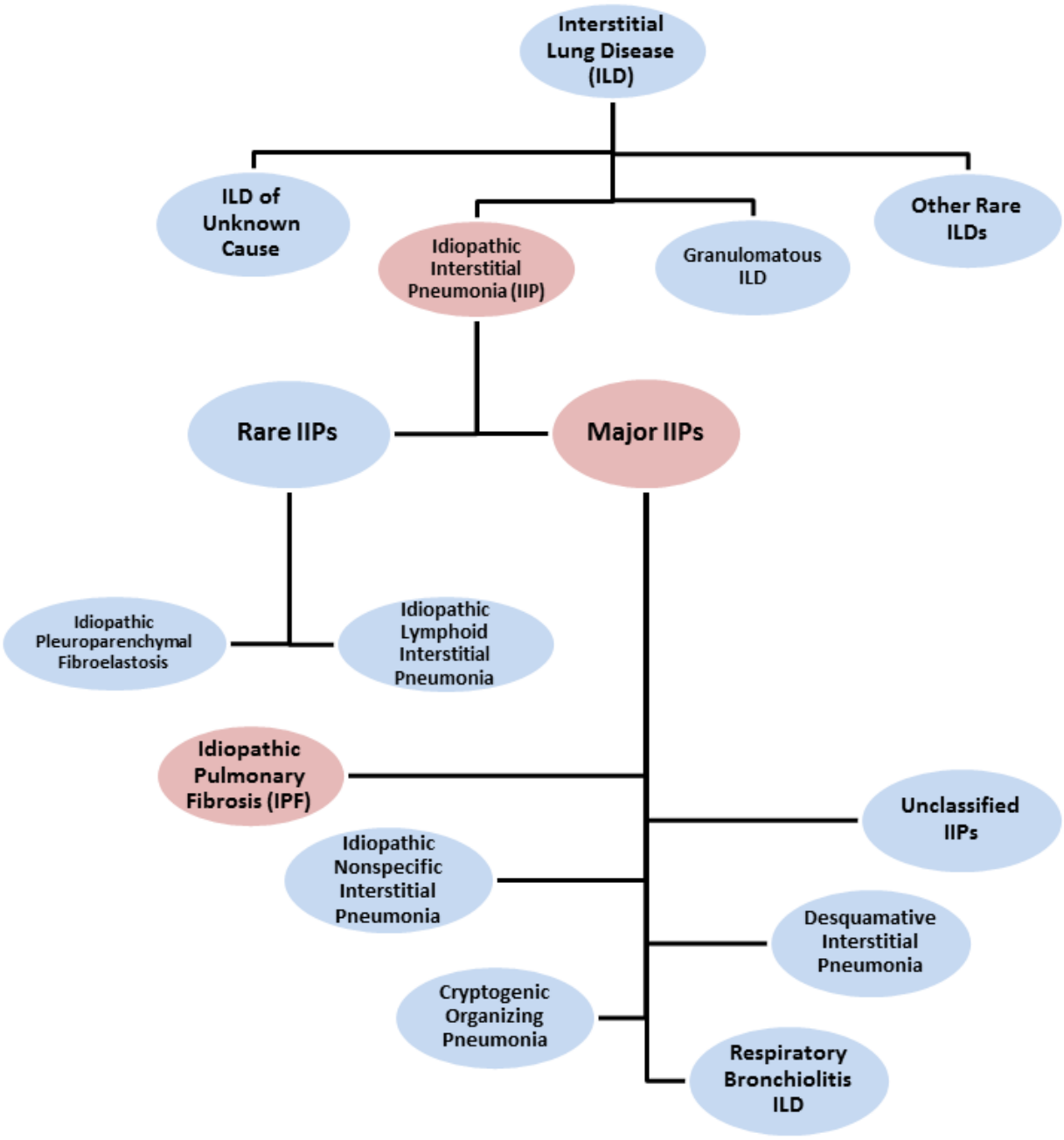


Figure 1: Classification Hierarchy of Interstitial Lung Diseases

The classification of ILDs includes over 200 subtypes. A recent update by the ATS/ERS expands the major IIPs to include over six sub-categories of IIPs, including IPF. Subtypes highlighted in pink represent IPF classification. Modified from ATS/ERS (2013).

1.1.3 Pulmonary Fibrosis in Newfoundland

The province of Newfoundland and Labrador, situated on the eastern coast of Canada, was initially settled by immigrants of Irish and English descent in the mid-18th century (Figure 2). Over time, small isolated communities were settled, along the rugged coast and little immigration out of the province. Over time, an increase in the prevalence of genetic disorders in Newfoundland families has resulted, including ones due to founder mutations. (Rahman et al., 2003) Examples of such disorders which are enriched in the NL population include: Bardet-Biedl Syndrome (Young et al., 1999), hemophilia A (Xie et al., 2002), arrhythmogenic right ventricular cardiomyopathy type 5 (ARVCD) (Merner et al., 2008), Lynch syndrome (Stuckless et al., 2013) and IPF (Fernandez et al., 2012).

Of the IPF cases reported in the Newfoundland population, 36% are familial (Fernandez et al., 2012), compared to an estimated 3% in Finnish population (Hodgson et al., 2002), and 0.5-2.2% in the United Kingdom (Marshall et al., 2000). As of January 2014, 146 patients have been enrolled in the IPF study at Memorial University of Newfoundland (MUN), including 68 sporadic and 79 familial patients who were clinically diagnosed through the Provincial Medical Genetics Program.

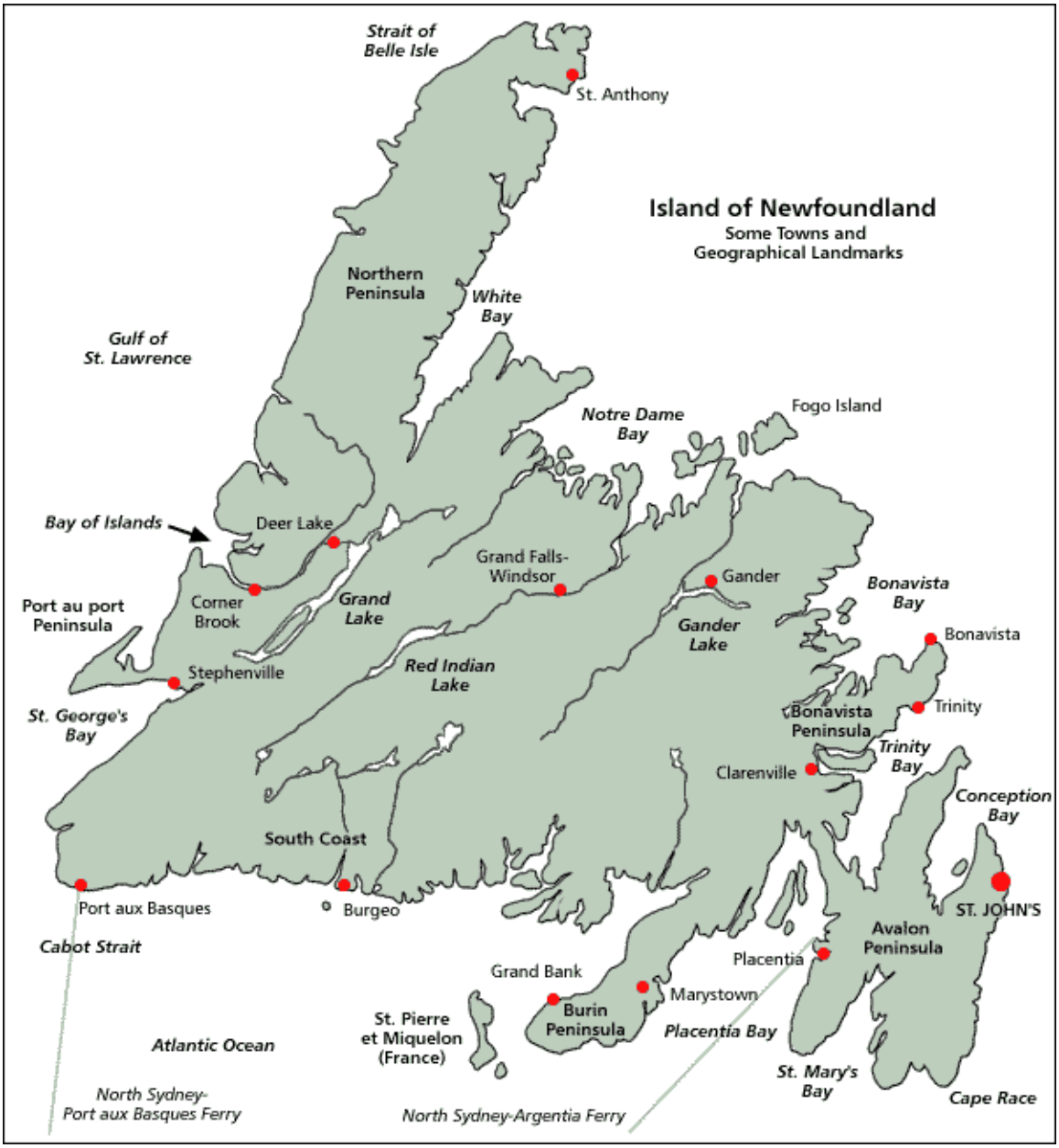


Figure 2: Major Settlements along the Coast of Newfoundland. Reproduced with permission by Wijayawardhana, 1999. Copyright the Newfoundland and Labrador Heritage Web Site.

Early colonization of coastal communities in Newfoundland has contributed to the isolation of many families and communities and is predicted to be a factor contributing to the development of multiple founder effects in the Newfoundland population.

1.1.4 Diagnosis of Pulmonary Fibrosis

In 2002, the American Thoracic Society and the European Respiratory Society (ATS/ERS) compiled detailed, international diagnostic criteria for IIPs in an attempt to minimize variable and confusing diagnoses within the ILD classifications (ATS/ERS, 2002). With specific regards to IPF, a surgical lung biopsy (showing "usual interstitial pneumonia" histology) is the gold standard clinical test. However, due to the invasive nature of this testing procedure, this is not always an option. Additional assessments, including pulmonary function tests (PFT) and high resolution computed tomography (HRCT) of the chest, allow the disorder to be diagnosed non-invasively (as outlined in Table 1). Recently, the ATS/ERS published updated diagnostic criteria for ILDs based on research spanning the previous ten years (ATS/ERS, 2013). Currently, the molecular etiology of ILDs are better understood and are diagnosed more often through analysis of PFTs and HRCTs, without the need for a surgical lung biopsy, as characterized in Table 2. In terms of IPF development, the natural progression is quite variable, with some patients displaying prolonged periods of stability, while others move rapidly to acute exacerbation. The ATS/ERS 2013 update is to be used in conjunction with the original 2002 classification for IIPs. In general, there is still much research needed to better understand the underlying molecular mechanisms of ILDs, as some patients remain diagnostic challenges due to mixed patterns of lung histology.

Table 1: Criteria for Diagnosis of Idiopathic Pulmonary Fibrosis in Absence of Surgical Lung Biopsy by the American Thoracic and European Respiratory Societies

Major Criteria
Exclusion of other known causes of ILD such as certain drug toxicities, environmental exposures, and connective tissue diseases
Abnormal pulmonary function studies that include evidence of restriction (reduced VC, often with an increased FEV_1/FVC ratio) and impaired gas exchange [increased $P(a–a)O_2$, decreased Pa_{O2} with rest or exercise or decreased DL_{CO}]
Bibasilar reticular abnormalities with minimal ground glass opacities on HRCT scans
Transbronchial lung biopsy or BAL showing no features to support an alternative diagnosis
Minor Criteria
Age > 50 yr
Insidious onset of otherwise unexplained dyspnea on exertion
Duration of illness > 3 mo
Bibasilar, inspiratory crackles (dry or "Velcro"-type in quality)

Definition of abbreviations: BAL = bronchoalveolar lavage; DL_{CO} = diffusing capacity of the lung for CO; HRCT = high-resolution computerized tomography; ILD = interstitial lung disease; $P(a–a)O_2$ = alveolar–arterial pressure difference for O_2; VC = vital capacity.

[†]In the immunocompetent adult, the presence of all of the major diagnostic criteria as well as at least three of the four minor criteria increases the likelihood of a correct clinical diagnosis of IPF.

Table 2: Categorization of Major Idiopathic Interstitial Pneumonias

Category	Clinical–Radiologic–Pathologic Diagnoses	Associated Radiologic and/or Pathologic–Morphologic Patterns
Chronic fibrosing IIP	Idiopathic pulmonary fibrosis	Usual interstitial pneumonia
	Idiopathic nonspecific interstitial pneumonia	Nonspecific interstitial pneumonia
Smoking-related IIP*	Respiratory bronchiolitis-interstitial lung disease	Respiratory bronchiolitis
	Desquamative interstitial pneumonia	Desquamative interstitial pneumonia
Acute/subacute IIP	Cryptogenic organizing pneumonia	Organizing pneumonia
	Acute interstitial pneumonia	Diffuse alveolar damage

*Desquamative interstitial pneumonia can occasionally occur in non-smokers.

Reproduced from "An Official American Thoracic Society/ European Respiratory Society Statement: Update of the International Multidisciplinary Classification of the Idiopathic Interstitial Pneumonias" (2013). Reprinted with permission of the American Thoracic Society.

1.2 Pathophysiology of Pulmonary Fibrosis

Although considerable research on the pathophysiology surrounding the development of IPF has been conducted, which provides a broad understanding of disease etiology and development, much of the underlying molecular mechanisms are still poorly understood. It appears that the development of IPF is multifactorial, with both environmental and genetic factors contributing to the occurrence and progression of the disease. The lungs are constantly exposed to environmental assaults from inhalation of irritants and invading microorganisms, thereby increasing the risk of microscopic and macroscopic lung injury. Many studies have found that exposure to environmental inhalants, such as cigarette smoke (Baumgartner et al., 1997) and wood dust (Hubbard et al., 1996), is associated with an increased risk for developing on ILDs. As well, there is an association with certain occupations, such as hairdressing, stone cutting and farming (Baumgartner et al., 2000). Overall the literature suggests that persistent occupational exposures are a risk for ILDs in the general population. Additionally, FPF candidates may be ultrasensitive to such exposures due to abnormal wound healing in the lung.

1.2.1 Pathophysiology of Abnormal Wound Healing

In a healthy lung, the normal response to tissue injury from external factors occurs in three distinct phases: injury, inflammation and repair, as described in Figure 4 (Wilson et al., 2009). Briefly, as a result of inflammation and the recruitment of profibrotic chemokines in damaged epithelial cells, a series of signalling pathways occurs in an attempt to regenerate, replace and re-establish normal lung tissue architecture. The pathways leading to wound repair have not been fully elucidated, but it is clear that there

are many cell types, chemicals and pathways involved in normal wound restoration, allowing many opportunities for repair mechanisms to malfunction. One of the many hypotheses about IPF development is that it may be triggered by aberrant wound healing (ie: the loss of the lung's ability to repair itself in response to environmental exposures). Disease progression may also be mediated by malfunctions of proteins and pathways involved in normal (Figure 3) and abnormal wound healing (Figure 4).

1.2.2 Hypothesis of Wound Healing in Pulmonary Fibrosis

Multiple researchers have suggested that inflammatory stimuli may lead to tissue destruction and over-active wound healing in IPF (Strieter, 2002). Other groups have hypothesised that inflammation is not a root cause of IPF, but rather an uncontrolled wound healing may be a major factor (Gauldie, 2002). Treatment of IPF patients with corticosteroids and other anti-inflammatory drugs appears ineffective in reducing disease progression (Davies et al., 2003). It has also been shown that increased production of transforming growth factor-beta (TGF-β) from epithelial cells invokes a fibrotic cascade in the absence of inflammation, suggesting that inflammation may not be essential to IPF development (Xu et al., 2003). It is becoming evident that an exaggerated recruitment of stimulatory cytokines, cellular growth factors and chemokines may be central to the pathogenesis of IPF (Figure 4).

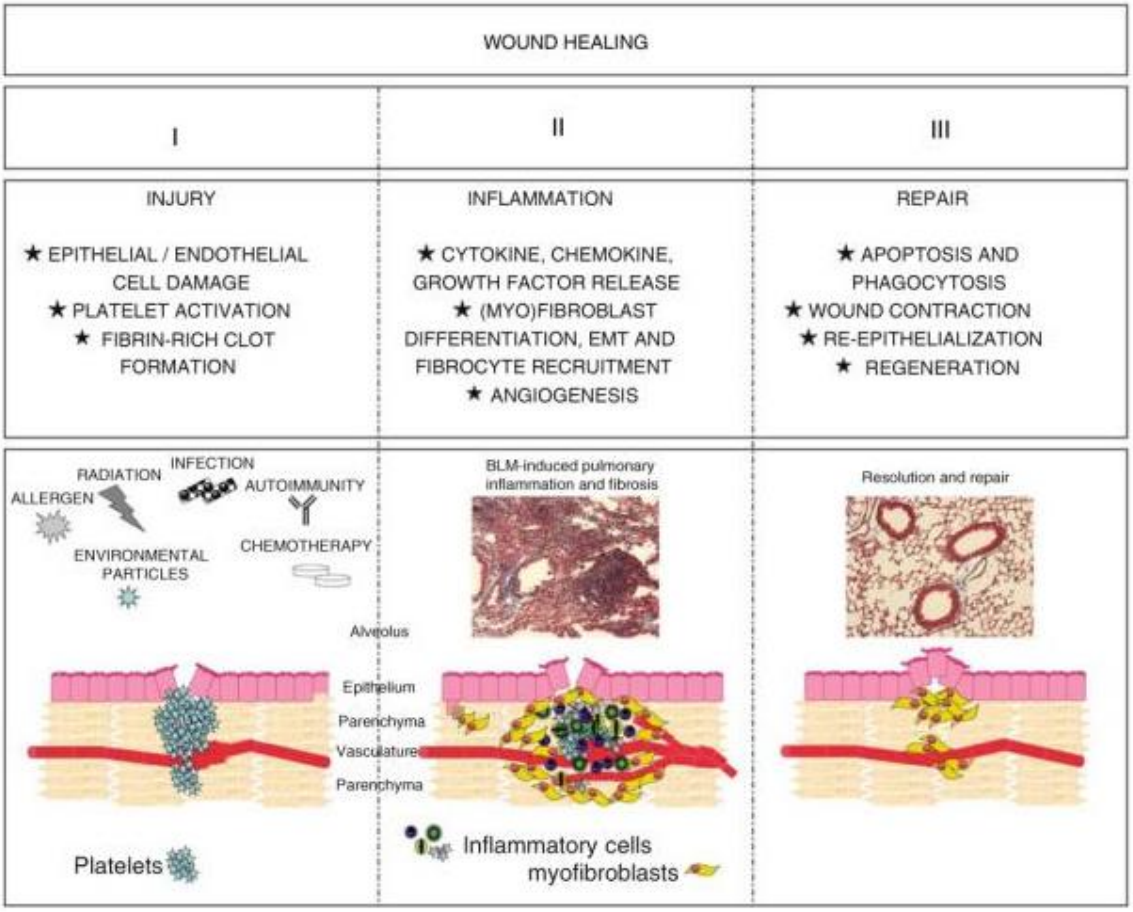


Figure 3: Phases of Normal Wound Healing. Reproduced from Wilson and Wynn, 2009 with copyright permission (Appendix E)

In the first phase of wound healing, injury to pulmonary tissue caused by environmental agents results in the formation of a provisional, fibrin-rich clot. In the second phase, circulating inflammatory cells are recruited to the site of injury, supplying growth factors and fibroblast-activating cytokines. During the last phase, the tissues attempt to repair the site of injury, resulting in contraction, fibroblast reduction and new epithelial and endothelia cells replace the provisional wound matrix.

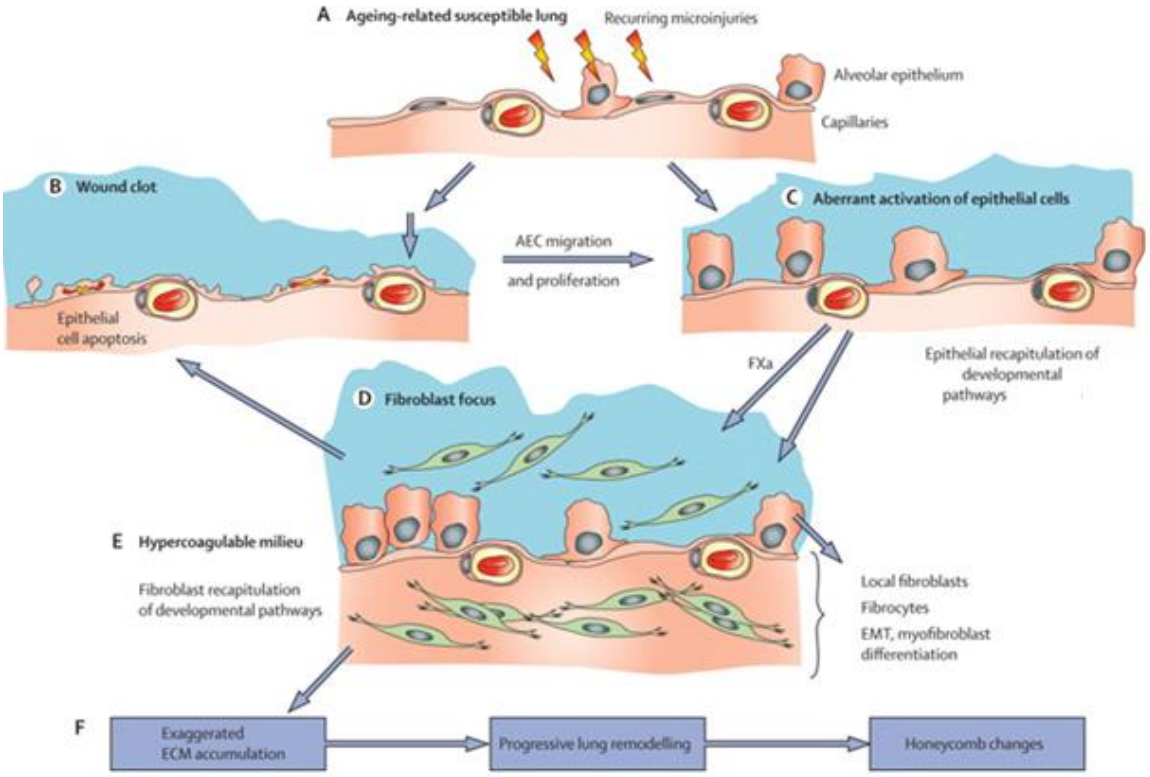


Figure 4: Predicted Pathogenesis of Idiopathic Pulmonary Fibrosis. Reprinted from King et al. (2011) with copyright permission (Appendix E)

A) Injuries from external stimuli, abnormal telomere shortening or genetic changes provoke type I and II epithelial cell death. B) Recruitment of extracellular matrix proteins (eg, fibronectin, fibrinogen) forms a wound clot. C) This process is followed by alveolar epithelial cell (AEC) migration and proliferation and the formation of fibrocytes from the production of several chemokines. D) This mixture of chemokines, fibrocytes and epithelial cells result in the formation of fibroblasts and myofibroblast foci E) which causes signalling pathways which prevents matrix degradation and enhances the fibrogenic response. F) The overall affect causes an increase in extracellular matrix (ECM) protein secretion, resulting in a progressive remodelling of the lung extracellular tissue.

The interaction of genetic and environmental factors during the development of IPF is largely unknown. In many FPF cases, it is believed that the lungs are already susceptible to lesions and/or poor wound repair mechanisms due to mutations in genes responsible for normal pulmonary function. These genetically vulnerable tissues may initially be irritated by environmental stimuli. Over time, recurrent injury and failure of tissue repair result in changes in the ECM of the lung tissue which may lead to the development of IPF. This may be due to the altered expression of inflammatory and stimulating growth molecules, including TGF-β (Leask et al., 2004).

1.2.3 Association of TGF- β with Pulmonary Fibrosis

TGF-β has long been affiliated with the development of IPF (Khalil et al., 1996; Tatler et al., 2012; Warburton et al., 2013), as well as with other diseases such as cardiac remodeling in heart disease (Rosenkranz et al., 2004) and specific cancers, such as uterine carcinosarcomas (Semczuk et al., 2013). As one of the primary growth factors in the body, the response elicited by TGF-β depends largely on the cell type. For example, TGF-β has been shown to arrest cell growth in epithelial cells (Howe et al., 1991), whereas addition of TGF-β to fibroblast cultures increases proliferation and differentiation into myofibroblasts (Bissell, 2001). TGF-β signalling plays a critical role in skin development, homeostasis and wound healing, and has been associated with a variety of wound repair and skin diseases such as psoriasis (Li et al., 2004) and scleroderma (Varga et al., 2009). As one of the most potent profibrotic chemokines, TGF-β has also been associated with IPF, yet the precise mechanism behind its action is

still largely unknown. It has been hypothesised that TGF-β signalling is related to the development of IPF through its downstream target genes.

Briefly, TGF-β signalling is mediated through serine/threonine kinase phosphorylation, as shown in Figure 5. TGF-β has multiple receptors, with the most common signalling pathway involving TGF-β receptors (TGFβR) types I and II. Initially, TGF-β binds to TGFβRII dimers on the cell surface. This results in recruitment of TGFβRI monomers, which form a complex with the TGF-β bound to TGFβRII. The formation of this complex results in the autophosphorylation of the TGFβR complex, resulting in the attraction and phosphorylation of the intracellular substrates Smad2 and Smad3. Phosphorylated Smad2 and Smad3 complex with Smad4 which then subsequently translocates into the nucleus. This Smad complex binds to transcription factor binding sites. Depending on the specific cell types, gene expression may be up or down regulated. While the majority of cells experience TGF-β signalling through a SMAD-mediated response, fibroblasts have been shown to undergo a different pathway. These pathways, independent of SMAD signalling, occur mainly through p21-activated kinase-2 (PAK2), which results in phosphatidylinositol 3-kinase signalling (Wilkes et al., 2005), as shown in Figure 6. Additionally, co-receptors for TGF-β have been found, including CD109 (Bizet et al., 2011), a negative regulator of TGF-β signalling. Due to the complexity of TGF-β signalling, more research is required in order to fully understand the signalling pathways involved in specific cell types as well as to identify TGF-β downstream targets.

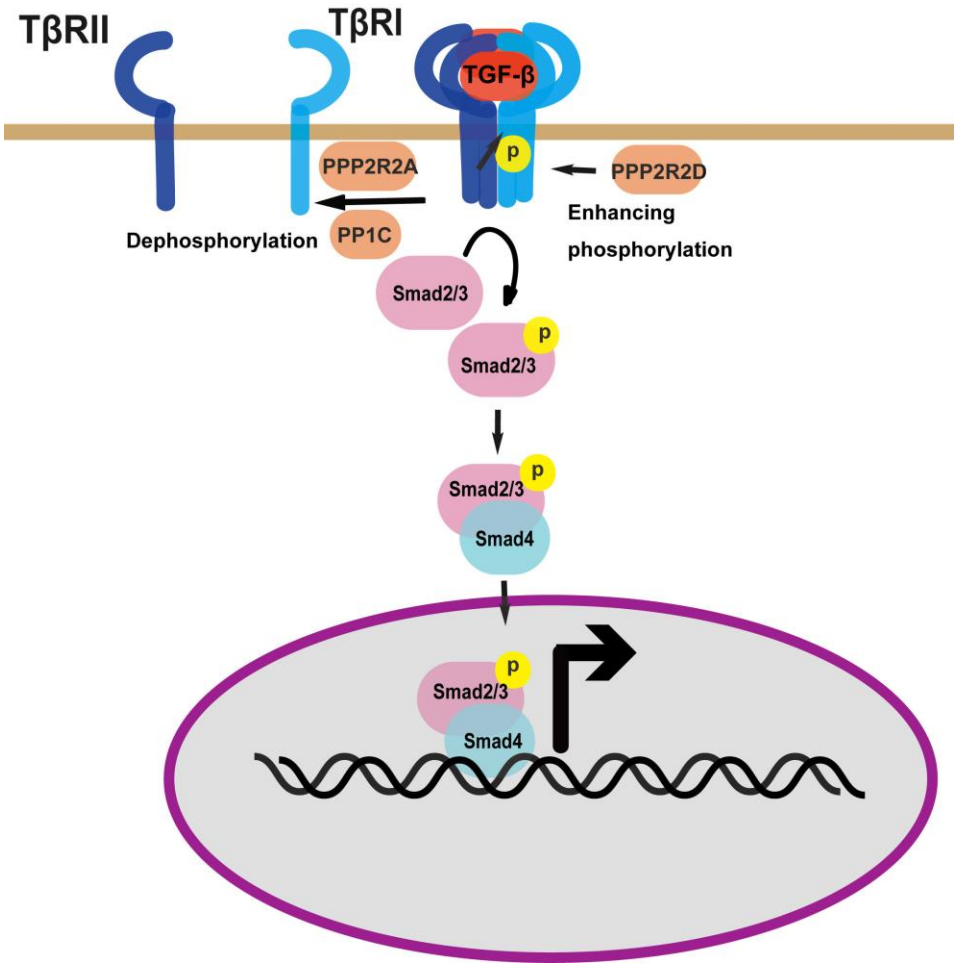


Figure 5: TGF-β Signalling Pathway. Reproduced with permission by Huang et al., 2012. Copyright Cell and Bioscience.

Phosphorylation of TGF-β receptors regulates their activity and subsequent downstream Smad-dependent signalling. Activation of the Smad2/3 4 complex via phosphorylation will permit this protein complex to enter the nucleus where it can bind to transcription factor binding sites and aid in gene regulation.

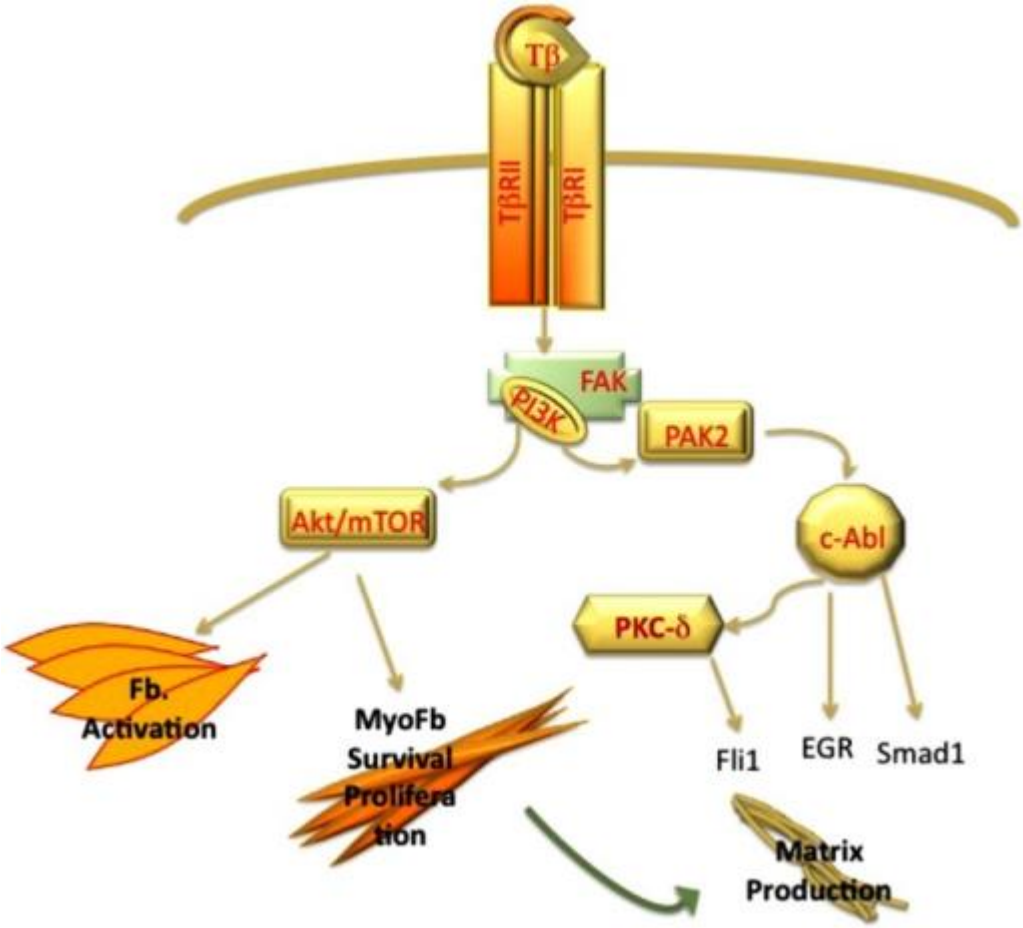


Figure 6:P13 Kinase Pathway Contributes to TGF-β Induced Fibrosis AKT and PAK2 Pathways. Reproduced with permission by Nakerakanti et al., 2012. Copyright by The Open Rheumatology Journal.

Activation of the TGF-β pathway via binding of TGF-β to its receptors results in the activation of P13K and Akt/mTOR pathways. These pathways are integral in the activation of fibroblast formation and myofibroblast proliferation, as is seen upon Akt/mTOR activation, as well as extracellular matrix production upon PI3K activation.

Studies have shown that one of the downstream targets for TGF-β signalling is telomerase reverse transcriptase (*TERT)* (Li et al., 2006; Lacerte et al., 2008), a known IPF susceptibility gene. The TERT enzyme, along with an RNA template entitled telomerase RNA component (*TERC*), belong to the telomerase protein complex which functions in maintaining chromosome ends through the addition of telomere nucleotide repeats. Li and colleagues first showed that TGF-β is capable of rapid repression of *TERT* gene expression in both normal and neoplastic cells. It was also shown that silencing *SMAD3* gene expression in human breast cancer cells resulted in disruption of the TGF-β repression of *TERT*. Similar findings were reported by Lacerte and colleagues in 2008. These discoveries may explain why many individuals who are affected with IPF have selectively shortened telomeres in circulating white blood cells compared with age-matched controls, even in the absence of *TERT* or *TERC* mutations (Alder et al., 2008; Cronkhite et al., 2008).

1.3 Genetic Etiologies of Pulmonary Fibrosis

There is a plethora of evidence to suggest that genetic factors underlie the development of IPF. The inheritance of IPF in many families is described in "Online Mendelian Inheritance of Man (OMIM, #178500) (OMIM, 2013) has been described in multiple members of the same family, as is seen in several Newfoundland families (Fernandez et al., 2012) and, in many pedigrees, is consistent with an autosomal dominant mode of inheritance (van Moorsel et al., 2010). Additionally, monozygotic twins raised in separate environments have displayed almost indistinguishable clinical presentations of IPF (Javaheri et al., 1980). The development of IPF has been associated

with known genetic conditions, such as Niemann-Pick disease (Nicholson et al., 2006) and Hermansky-Pudlak syndrome (Carter, 2012). Finally, mouse studies have shed light onto the genetic predispositions for IPF in inbred strains. For example, upon exposure to bleomycin, a known fibrogenic stimulus in the lungs of mice, C57BL/6 strains are more likely to develop lung fibrosis than are BALB/c strains (Ortiz et al., 1998). These lines of evidence have led to the belief that there are strong genetic components leading to the development of IPF. Additionally, individuals exposed to similar concentrations of asbestos have been shown to experience different outcomes in the development of lung diseases, indicating that genetic underlying factors may exacerbate IPF development (Polakoff et al., 1979).

Although genetic research into the underlying etiologies of IPF has a 30 year history, only a few genes have been shown to be directly associated with IPF development. Recently, a genome-wide association study (GWAS) identified seven new loci associated with IPF, with results suggesting that genes involved in deoxyribonucleic acid (DNA) repair, immunity and cell-cell adhesion may play a role in the development of the disease (Fingerlin et al., 2013). Recent estimations predict approximately 80% of genetic factors contributing to IPF development remain unknown (ATS/ERS, 2013); therefore the work from both GWAS and molecular studies will help in determining additional genetic factors involved in IPF development. The known genes associated with the progression of IPF are involved with surfactant protein production, telomere maintenance and production of the mucosal lining in the lung (Steele et al., 2013).

1.3.1 Surfactant Proteins

Pulmonary surfactant is composed of proteins and phospholipids secreted by alveolar type II epithelial cells. They are essential for normal lung function, by lowering surface tension at the epithelium/ air interface. There are many surfactant proteins, but surfactant protein A and surfactant protein D are particularly important in reducing fluid accumulation in the lung, eliciting immune responses and regulating inflammatory responses (Wright, 2004).

Surfactant proteins were first associated with IPF in 2001 in an infant with idiopathic nonspecific interstitial pneumonia (Nogee et al., 2001). Nogee and colleagues identified a heterozygous, intronic nucleotide substitution in surfactant protein C (*SFTPC),* a gene responsible for producing surfactant protein C. Since this discovery, many families with UIP spanning multiple generations have been reported with *SFTPC* mutations (Thomas et al., 2002).

Since discovery of the association between *SFTPC* mutations and IPF, studies have revealed mutations in additional surfactant protein genes, such as surfactant protein A1 (*SFTPA1)* and surfactant protein A2 (*SFPTA2)* in many FPF patients (Steele et al., 2013). Using a genome-wide linkage scan, Wang and colleagues identified a linkage signal on chromosome 10 in a large family with FPF and subsequently identified a glycine to valine amino acid substitution in codon 231 of *SFTPA2* (Wang et al., 2009). In a 2003 study, association between two synonymous coding single nucleotide polymorphisms (SNPs), as well as one nonsynomous SNP, was linked with IPF in 84

sporadic cases (Selman et al., 2003). Heterozygous mutations in surfactant protein coding genes *SFTPC* and *SFTPA2* account for a small proportion of IPF cases, while mutations in genes encoding telomerase enzymes contribute a larger genetic predisposition (ATS/ERS, 2013)

1.3.2 Telomerase Enzymes

Telomeres are short tandem repeats of DNA that occur at the ends of chromosome by means of the telomerase complex, which slowly decrease with each cell division. Once a critical repeat size is reached, a normal cell senescence program is initiated. Malfunctions in the telomerase complex were first reported in individuals with dyskeratosis congenita (DKC), a genetic condition characterized by abnormal skin pigmentation, nail dystrophy, oral leukoplakia, and pulmonary fibrosis (Vulliamy et al., 2006). DKC displays high genetic heterogeneity, and has been shown to follow multiple modes of inheritance as a result of mutations in different genes. In 1998, an X-linked inheritance of DKC was shown to be a result of mutations in dyskeratosis congenital-1 (*DKC1)*, a gene encoding the dyskerin protein which is responsible for telomere stabilization and maintenance. (Heiss et al., 1998). Recently, a novel mutation in *DKC1* was found in a kindred displaying familial interstitial pneumonia (FIP), as well as clinical signs of DKC, including hyperpigmentation and a dyskeratotic rash of the hands (Kropski et al., 2014). Additionally, autosomal dominant forms of DKC have been shown to be caused by mutations in *TERT* and *TERC* (Armanios et al., 2005), as well as autosomal recessive forms resulting from mutations in NOP10 ribonucleoprotein (*NOP10),* a protein involved in ribosome biogenesis and telomerase maintenance (Rashid et al., 2006; Walne

et al., 2007).). IPF is seen in 20% of patients diagnosed with DKC, with respiratory failure being the second most common cause of death in these individuals (Dokai, 2000).

Mutations in the telomerase genes *TERT* and *TERC* are associated with IPF in both sporadic and familial cases (Vulliamy et al., 2001) and account for approximately 10% of cases (Armanios et al., 2005). *In vitro* analysis demonstrated decreased telomerase activity, as well as telomere shortening in IPF patients (Armanios et al., 2007), with suggestions that mutations in telomerase enzymes may result in premature shortening of telomeres in the alveolar epithelium. This hypothesis, combined with a decrease in epithelium regeneration and excessive apoptotic response of the alveolar epithelial cells, is believed to contribute to the development of IPF (Selman et al., 2014). Furthermore, many of the IPF patients who test negative for mutations in *TERC* and *TERT* nevertheless demonstrate shortened telomeres (Steele et al., 2013), suggesting that there may be other genes involved in telomere maintenance.

1.3.3 MUC5B

Recently, a promoter polymorphism in *MUC5B* has been associated with IPF segregation in many cohorts throughout the world (Seibold et al., 2011). *MUC5B* encodes a mucin, whose expression is increased in lung tissue and alveolar lesions of many IPF patients (Seibold et al., 2011). Using a genome wide linkage analysis, Seibold and colleagues reported an association between this promoter variant in *MUC5B*, *rs35705950,* and disease phenotype in 83 families. Seibold and colleagues hypothesise that this promoter SNP may lead to the overexpression of *MUC5B.* An increase in mucin

production may result in blockages of the bronchioles and reduced clearance of inhaled substrates, leading to inflammation and changes in the extracellular matrix (ECM) of the alveoli. Additionally, previous work in the Woods lab has identified an association with *rs35705950* in two FPF families, families R1136 and R0942 (Pirzada, 2012). Using a case-control study design, the presence of *rs35705950* was evaluated in 110 affected IPF individuals and 277 controls. Odds ratios for IPF affected individuals who were homozygous or heterozygous for the minor allele of this SNP were 12.2 (95% confidence interval, 3.3 to 44.7, $P < .001$) and 5.4 (95% confidence interval, 3.3 to 9.6, $P < .001$), respectively (Pirzada, 2012). Using SImplified rapid Segregation Analysis (SISA) (Møller et al., 2011), the likelihood that co-segregation happened by chance was 1.56%, suggesting the *rs35705950* promoter variant may have a gene-dosage effect in the development of IPF in these two families.

Recently, Zhang and colleagues suggested the *rs35705950* variant may not be the only genetic contributor to IPF in patients initially described by Seibold and colleagues. They concluded that functional annotation on the effects this variant has on protein production and thus disease development should be conducted before the phenotypic impact of this polymorphism can be properly interpreted (Zhang et al., 2011). Polygenic inheritance must also be considered when evaluating linkage between SNPs and disease phenotype in genetically heterogenic, reduced penetrant diseases, such as IPF. As the phenotypic consequences of this variant are still undetermined, the possible role of other variants should be further investigated in these two families.

1.3.4 Human Leukocyte Antigen

Human Leukocyte Antigen (HLA) includes a locus of genes responsible in maintaining the immune system. There are a variety of *HLA* genes, which include the major antigens such as HLA-A, HLA-B and HLA-C. These proteins specifically act as cell surface antigens and contribute to the major histocompatibility complex (MHC), which plays an important role in host defense and cell immunity. *HLA-A* allelic polymorphisms have been previously associated with IPF (Zhang et al., 2012), but have also been associated with other disorders such as lung cancer (Araz et al., 2014), breast cancer (Leong et al., 2011), and autoimmune diseases (Gough et al., 2007).

1.3.5 Polygenic Inheritance in R0851

The literature suggests that IPF is a multifactorial complex disease, with genetic and clinical heterogeneity (Steele et al., 2013). It is estimated that approximately 80% of the genetic contribution to IPF remains unknown, suggesting that many familial cases have an underlying polygenic form, including family R0851. As shown in Figure 7, a novel *TERT* variant, c.1892G>A - predicted to be pathogenic - was found to segregate completely in a nuclear family of R0851 (Fernandez et al., 2012). Upon testing additional affected relatives using Sanger sequencing, the variant was not present in distantly related family members. Additionally, the affected sons in the nuclear family possessing the *TERT* variant were diagnosed in their late 20's, a much younger age than is typically seen in FPF patients. This observation suggests there may be an additional variant in this particular nuclear family that may be shared by other affected relatives. However, families with *TERT* mutations are shown to display genetic anticipation (Armanios et al.,

2005), in which the age of onset decreases with subsequent generations. This may also explain the earlier age of onset in second generation members of this family. Family R0851 demonstrates the complexity that is seen in autosomal dominant diseases with variable expressivity and reduced penetrance, and whose disease progression is also affected by environmental factors, polygenic inheritance and perhaps genetic anticipation. The concept that multiple, medium penetrant genetic factors may affect disease development is difficult to control for, yet should be considered when analyzing data.

1.4 Previous Work Completed by Others

1.4.1 Patient Recruitment and Assessment

The recruitment of patients and construction of family pedigrees was conducted by Dr. Bridget Fernandez (clinical geneticist) and Ms. Barbara Noble (research nurse). Upon recruitment, individuals signed consent forms granting researchers access to use their personal medical history, family history and DNA samples. The diagnosis of IPF was made using the guidelines suggested by ATS/ERS (2002) (Table 1). The majority of IPF patients were diagnosed based on PFTs interpreted by respirologist Dr. George Fox, as well as independent assessment of HRTCs by Dr. Rick Batia and Dr. Eric Sala using the Royal Brompton Hospital scoring system (Wells et al., 2003). PFTs included diffusing capacity of carbon monoxide, forced vital capacity and forced expire volume in one second. In assessing HRCT, scans were scored at five different levels in the lung. Scores were determined for mean coarseness of fibrosis, presence of lower lobe dominance and mean proportion of abnormal lung.

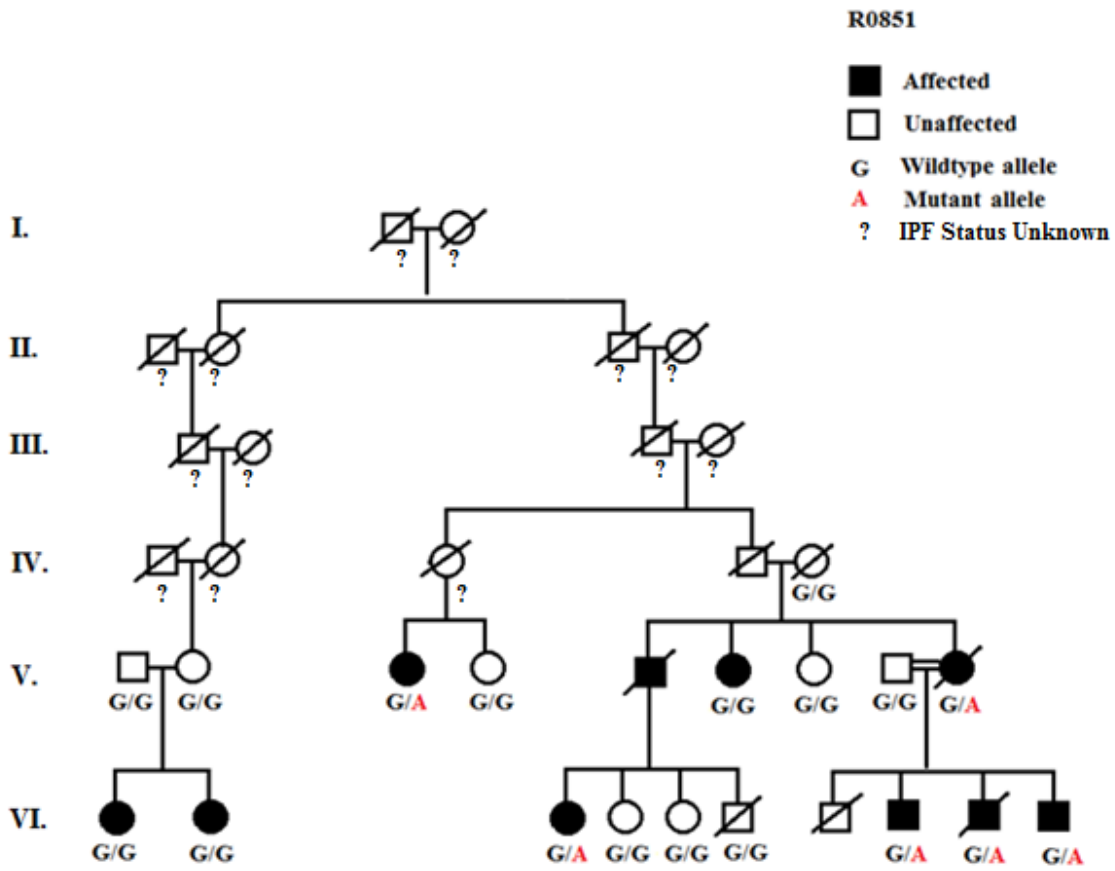


Figure 7: Segregation of *TERT* variant in R0851

A novel *TERT* variant, c.1892G>A, was found to segregate in one branch of family R0851. This predicted pathogenic variant was not found in distantly related family members also affected with IPF. It is believed there may be multiple genetic contributions to IPF in this particular family.

As shown in Appendix A, the HRCT scoring system translates to different classifications of IPF, including "typical IPF", "probable IPF", "possible IPF" or "unaffected". Taking into consideration all clinical data [PFTs, HRTCs, tissue biopsy - if available - family history and pedigree analysis] each patient was classified as "definitely affected', "probably affected", "possibly affected" or 'unaffected". Patients clinically classified as "definitely affected" were shown to have UIP based on a lung biopsy or autopsy report, while those clinically classified as "probably affected" met ATS/ERS criteria for non-invasive diagnosis, as outlined in Table 1. However, for the purpose of the study, patients clinically classified as "definitely" or "probably affected" according to the clinical assessment given by Dr. Fernandez, Dr. Fox, Dr. Batia and Dr. Sala, were categorized as "affected".

1.4.2 Genome- Wide Linkage Analysis Using Microsatellite Markers and Fine Mapping

Previously, a 10 centimorgan (cM) genome-wide scan was conducted by Ms. Laura Edwards (M.Sc. candidate) using 382 autosome microsatellite markers in two families: R0942 and R0851 (Edwards, 2006). A two-point linkage analysis was conducted, where a Logarithm of Odds (LOD) score greater than 3(θ=0) was considered significant for linkage, and a score less than -2 (θ=0) was deemed significant for exclusion of linkage. Results from both families showed no significant linkage association; however, each family had loci in which linkage was suggested, with scores greater than 1.00.

For family R0942, the maximum LOD score was 2.19 for marker D16S423, located at 16p13.3, while family R0851 showed 4 markers with a LOD score ranging from 1.00 to 3.00, suggesting possible linkage on at 16p13. Fine mapping of this region using an additional 10 microsatellite markers was conducted by Ms. Laura Edwards and Mr. Fady Kamel (Memorial University) and was the basis of Mr. Kamel's thesis project (Kamel, 2010). This region was further analysed using a genome-wide SNP scan on five affected families: R0851, R0892, R0896, R0942, R1136 and R1487. Of the five families, R0942 was shown to have a haplotype segregating with the disease, a 17.38 megabase (Mb) region, containing 386 genes on chromosome 16p13.3 (Kamel, 2010).

SNP marker genome-wide scan analysis revealed suggested linkage on chromosome 1 for R0851, with LOD scores greater than 1.00. Multipoint linkage analysis also identified one region of suggested linkage, marker rs930027 on chromosome 18. However, there were no genes in this region suggestive of involvement in IPF development, such as wound healing and fibrosis in the lung, and this region was excluded from further investigation (Kamel, 2010). Results from two-point parametric linkage analysis demonstrated only two markers with heterogeneity LOD scores (HLOD) over 3.00, indicating statistical significance for the markers *rs942631* (HLOD =3.22) and marker *rs3130922* (HLOD =3.15), located at chromosomal loci 6p24.3 and 6p21.3, respectively (Kamel, 2010). This region was further analysed using a candidate gene selection method.

1.4.3 Candidate Genes Sequenced

Results from the microsatellite genome-wide scan, microsatellite marker fine mapping and SNP marker genome-wide scan identified three loci of interest (chromosome 1p, 6p24.3 and 6p21.3) for families R0851, R0892, R0896, R0942 and R1136, with statistical association for IPF development. From these loci, 13 genes were selected as candidate genes, based on gene function, and Sanger sequencing to determine potential IPF causal variants (Kamel, 2010). All exons of each gene were sequenced but no causative variants were found (Kamel, 2010). An additional nine genes were sequenced in R0942, based on potential candidate gene loci by Mr. Ashar Pirzada (Pirzada, 2012). These genes were selected based on previously identified loci with suggested linkage, 6p24.3-6p23 and 16p13.3, and were selected from a list of genes: 386 genes from chromosome 16p13.3 and 134 genes from chromosome 6p24.3-6p23. Of the nine genes sequenced, 28 variants were found, all of which were excluded upon Sanger sequencing due to lack of segregation and presence in control samples (Pirzada, 2012). Of the total 28 families enrolled in the FPF NL study, three families tested positive for three different *TERT* mutations (R0892, R0851 and R1254). However, as previously mentioned, the *TERT* mutation found in R0851 did not fully explain the genetic contribution to IPF development in this family, therefore, additional genetic analysis was required.

1.4.4 Telomere Length Assay

Since many individuals who test negative for *TERT* and *TERC* mutations still have shortened telomeres (Wuyts et al., 2013), telomere length assays were conducted on

FPF patients from multiple families (Fernandez et al., 2012). Venous blood samples were sent to Repeat Diagnostics in British Columbia, Canada. Lymphocytes and granulocytes were isolated from venous blood samples by *in situ* hybridization and flow cytometry. Telomere lengths for two types of white blood cells, granulocytes and lymphocytes, were analysed and compared to average values calculated for Caucasian control populations. As shown in Figure 8, telomere lengths were markedly reduced in both granulocytes and lymphocytes in R1136 (Figure 8A). In R0942, telomere lengths were within the normal range for age matched controls (Figure 8B). These results reflect the heterogenous pathophysiology of IPF development. A diagnosis of IPF is often accompanied by shortened telomeres (Steele et al., 2013). However, some families have normal telomeres, as demonstrated in R0942. The above subgrouping of IPF-families demonstrates telomere length heterogeneity, similar to other that of other studies (Diaz de Leon et al, 2010), suggesting that there are multiple pathways involved in IPF development, some of which are unrelated to telomere maintenance.

As previously discussed, it is known that mutations in both surfactant and telomerase proteins can result in the development of IPF. However, there has yet to be convincing evidence to support a link between the surfactant and telomerase gene products. Current findings suggest there may be several developmental pathways involved in the IPF pathway, including genes involved in host defense, aging and cell-cell adhesion (Mathai et al., 2014). Furthermore, recent research investigating the contribution of a specific *TERT* variant, *rs2736100*, and a common *MUC5B* promoter SNP, *rs3570590*, has determined that although both SNPs produce statistically significant

associations for the development of ILD, they do so independently of one another, supporting the genetic heterogeneity of ILDs (Rongrong et al., 2014). These findings provide evidence that there are multiple pathways involved in the development of IPF. Previous work in the Woods lab utilized a genomic linkage and candidate gene approach to attempt to identify such genetic variants. For this project, a new approach, whole exome sequencing (WES), was used to determine novel or rare genetic variants in the exomes of IPF patients.

1.5 Illumina HiSeq Whole Exome Sequencing

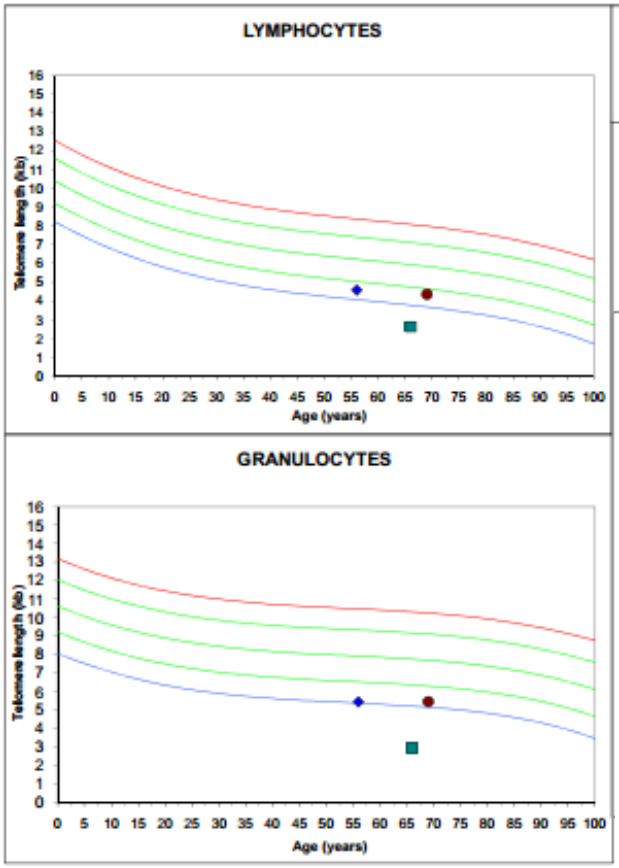
WES involves the targeting and capture of exons, the protein coding sequences of the genome. The human genome consists of approximately 180,000 exons, yet represents only 1% of the entire genome (Choi et al., 2009). About 85% of known mutations with large effect on human disease are found in the exome (Choi et al., 2009). It is anticipated that additional highly penetrant rare variants can be found by WES of familial cases and implemented in a clinical diagnosis (Raddatz-Sikkema et al., 2013). Given an estimated 1.2 million coding variants per human exome (Fu et al., 2013), WES utilizes high throughput DNA sequencing to annotate massive sequences of DNA in parallel. Since 2007, several next generation sequencing (NGS) technologies have been developed differing in the method of exome capture, read length, depth of coverage and run time. The filtering and analysis of WES data has recently been applied to rare Mendelian disorders. For example, WES has been successful in identifying causative variants in autosomal recessive disease such as Miller Syndrome (Ng et al., 2010) as well as for clinical diagnosis of genetic diseases, such as Bartter Syndrome (Choi et al., 2009).

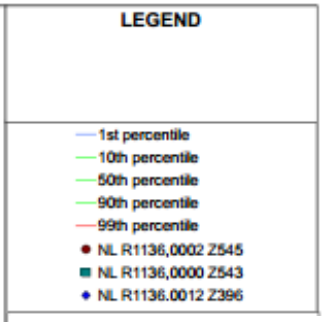
A) Telomere Length Assays in Family R1136

Sample	Age	Sex	Lymphocytes MTL *	Lymphocytes MTLN ***	Lymphocytes INT ****	Granulocytes MTL *	Granulocytes MTLN ***	Granulocytes INT ****
			(kb)	(kb)		(kb)	(kb)	
NL R1136,0002 Z545	69	F	**4.4**	5.8	**L**	**5.5**	7.7	**L**
NL R1136,0000 Z543	66	M	2.6	5.9	VL	3.0	7.8	VL
NL R1136.0012 Z396	56	M	4.6	6.2	L	5.5	8.0	L

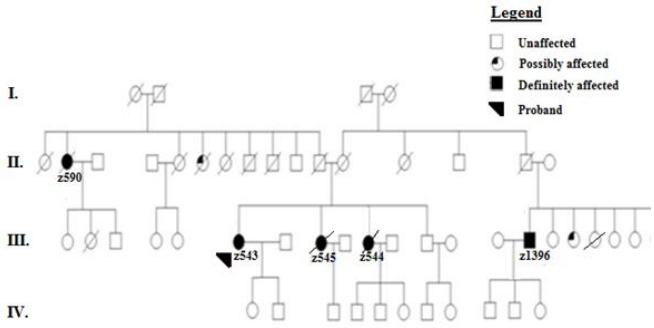
* MTL = Median Telomere Length
*** MTLN = Normal Median Telomere Length at age (50th percentile)
**** INT = Telomere length interpretation

INT:
VH = Very High
H = High
N = Normal
L = Low
VL = Very Low





Pedigree of R1136

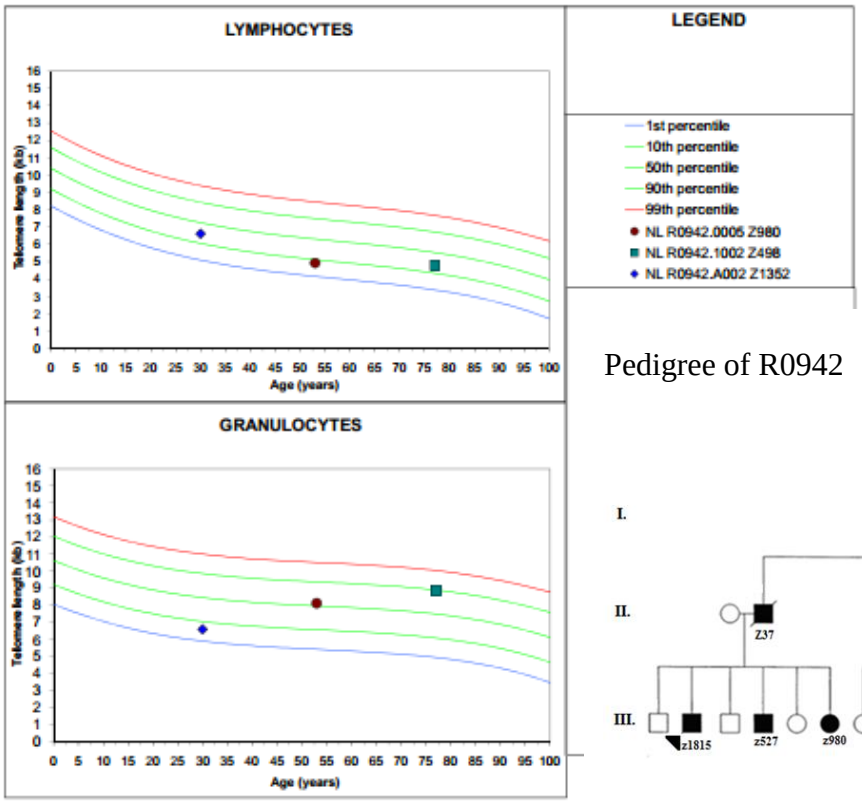


B) Telomere Length Assays in Family R0942

Sample	Age	Sex	Lymphocytes MTL *	Lymphocytes MTLN ***	Lymphocytes INT ****	Granulocytes MTL *	Granulocytes MTLN ***	Granulocytes INT ****
			(kb)	(kb)		(kb)	(kb)	
NL R0942.0005 Z980	53	F	**4.9**	6.3	**L**	**8.2**	8.0	**N**
NL R0942.1002 Z498	77	M	4.8	5.5	N	8.9	7.5	N
NL R0942.A002 Z1352	30	F	6.6	7.2	N	6.6	8.5	L

* MTL = Median Telomere Length
*** MTLN = Normal Median Telomere Length at age (50th percentile)
**** INT = Telomere length interpretation

INT:
VH = Very High
H = High
N = Normal
L = Low
VL = Very Low



Pedigree of R0942

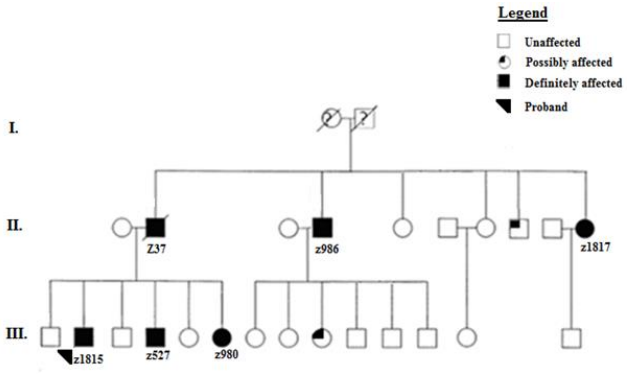


Figure 8: Telomere Length Assays for R1136 (A) and R0942 (B)

Analysis of telomere length assays for both R1136 (A) and R0942 (B) showed that telomere lengths were below normal for members of R1136 and within normal range for R0942.

As described in the following sections, the filtering of WES data utilizes a variety of bioinformatics programs to annotate variant calls. This project utilizes the programs Sorting Intolerant from Tolerant (SIFT), PolyPhen2, Genomic Evolutionary Rating Profile (GERP) and Grantham to predict whether specific nucleotide base changes and amino acid substitutions are detrimental in protein structure and function. SIFT is an algorithm used to predict whether a specific amino acid change is likely to affect protein function (Ng et al., 2003). The program does not rely on protein structure, but on the conservation of amino acids. SIFT scores range from 0 to 1, with a score of less than 0.05 thought to be detrimental in affecting protein function. Like SIFT, PolyPhen2 also evaluates amino acid changes, but PolyPhen2 evaluates the effect of amino acid changes on protein structure (Adzhubei et al., 2013). The algorithm uses both paralogues and orthologues to assess structure and functional changes in relation to disease. The scoring system for Polyphen2 ranges from 0 to 2. A score of 0.000 is considered a "benign" change, while a score of 0.999 represents a "probably damaging" change. Grantham scores categorizes codon replacements into various classifications based on chemical dissimilarity. Classes range from 0-50 for "conservative" changes, 51-100 for "moderately conservative" changes, 101-150 for "moderately radical" changes and greater than 151 for "radical" changes (Li et al., 1984). Finally, GERP scores are used to determine evolutionary constraint of a specific locus based on sequence conservation amongst varying taxa (Cooper et al, 2005). Scores range from -12.3 to 6.17, with positive scores indicating increased evolutionary constraint and stronger nucleotide conservation.

These bioinformatics programs, as well as other variant annotation tools, were used in the analysis of WES data.

1.6 Hypothesis and Objectives

1.6.1 Hypothesis

There is strong evidence to suggest that IPF has a large genetic component. Approximately 80% of causative genes for IPF remain unknown, and it is believed that multiple novel genetic variants in previously unidentified FPF susceptibility genes will be found in FPF families around the world. Therefore, I hypothesise that – using WES - the causative genetic variants segregating through the Newfoundland FPF cohort will be found in genes not previously associated with IPF.

1.6.2 Objectives and Rationale

The objective of this study is to identify novel FPF susceptibility genes in a Newfoundland cohort, specifically in families R0942 and R1136.

1. To identify and categorize genetic variants found in the exomes of 24 FPF patients from 14 families. This will be accomplished by WES.
2. To analyze lists of moderate- and high-impact exome sequencing variant lists, compiled by Genome Québec in order to identify candidate genes.
3. To determine segregation of candidate variants in FPF families
4. To assess the frequency of candidate variants in NL control families.

The identification of causative variants in susceptibility genes for IPF has multiple benefits. Primarily, identifying IPF causal genes will help better understand the pathophysiological and molecular mechanisms underlying the etiology and development of this fatal disease. With a greater understanding of the disease mechanisms, we may be able to develop specific biomarkers appearing at earlier disease stages. Clinicians may be able to screen additional FPF families for mutations in these genes, for example through the inclusion of these causal genes in a gene panel. Once a family's mutation is known, asymptomatic first-degree relatives have the option of determining whether or not they are mutation positive. This may allow also allow molecular classification strategies for the diagnostic hierarchy of ILD. While there are currently no pharmacotherapeutic interventions that stop the natural disease history in pre-symptomatic patients or in people with very early disease, this may change when the underlying pathophysiology is better understood. For example, the identification of additional IPF-associated genes lend to the development of gene therapy. Once all IPF genes are known, hopefully therapies will be developed that delay or prevent disease onset.

2.0 Materials and Methods

2.1 Patient and Family Recruitment

The worked described in this thesis continues the previous work conducted by other members of the Woods lab using a novel strategy as outlined in Figure 9. As of January 2014, 146 research participants have been enrolled in the current IPF study, consisting of 79 FPF patients and 68 sporadic IPF patients, with no known family history. Ascertainment of research participants was conducted between January 2006 and July 2011 by a team of local neurosurgeons and research nurse Ms. Barbara Noble, under the supervision of Dr. Bridget Fernandez (Fernandez et al., 2012). These 79 FPF patients represent 28 families, from various regions of Newfoundland, as shown in Figure 10. For the work completed in this thesis, DNA samples from 24 individuals from 14 Newfoundland FPF families were selected for WES, which was performed at McGill University and Genome Québec, as shown in Figure 10 and Table 3. For WES, patients were selected from families with the most number of affected individuals. The individuals selected for WES within each family consisted of the proband, an affected first degree relative and an affected distantly related relative. The rationale for patient selection is that distant relatives are more likely to share genetic variants by chance, thus eliminating a large number of non-causative, shared variants that are seen amongst closely related relatives. At the current time, unaffected individuals were not selected for WES analysis, as clinically unaffected individuals may develop IPF at a later age due to variable expressivity of the condition, which could confound results. Analysis of the high impact variants was performed with a specific filtering strategy as discussed below.

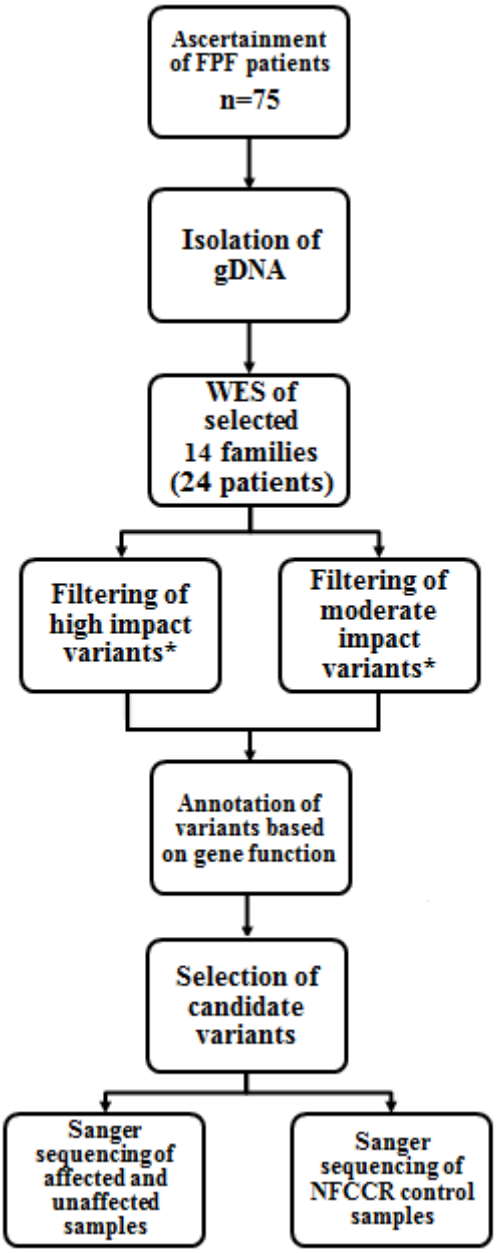


Figure 9: Overall Thesis Study Design

Flowchart outlining the overall thesis study design as described in detail in Section 2.

*Filtering of high and moderate impact lists is described in detail in Section 2.3 and Figure 14.

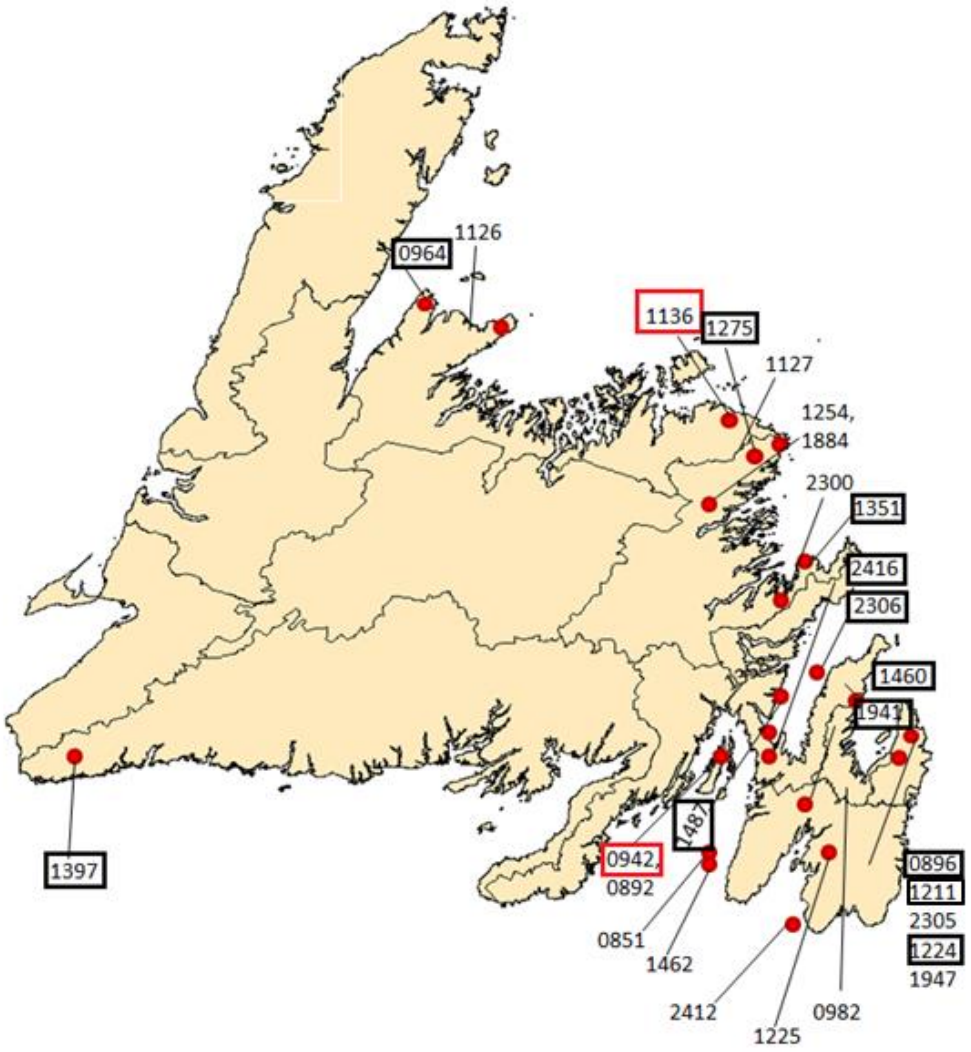


Figure 10: Geographic Origins for Newfoundland Familial Pulmonary Fibrosis Families Enrolled in Study. Reproduced with permission from Fernandez et al., 2012. Copyright Respiratory Research.

Family numbers for all IPF families enrolled in the study are depicted above and indicate the location of their community of origin, including R1136 and R0942 (red boxes).

Table 3: List of 24 Individuals Sequenced using Whole Exome Sequencing by McGill University and Genome Québec

Sample Number	Family	Community of Origin
z980	R0942	Chance Cove, Trinity Bay
z37	R0942	
z986	R0942	
z543	R1136	Musgrave Harbour, Northeast coast
z544	R1136	
z1396	R1136	
z900	R1487	South Dildo, Trinity Bay
z902	R1487	
z1373	R1487	
z529	R0896	St. John's
z730	R0896	
z618	R0896	
z659	R1351	Duntara, Bonavista, Bay
z660	R1351	
z624	R1224	St. John's
z690	R1224	
z1392	R1460	Harbour Grace
z525	R0964	Coachmans Cove, Baie Verte Peninsula
z647	R1275	Templeman, Bonavista Bay
z1647	R2416	
z1470	R1941	Indian Point
z1547	R2306	
z733	R1397	
z616	R1211	St. John's

2.2 Whole Exome Sequencing Methodology

2.2.1 Sample Selection

Between July 2006 and 2011, venous blood samples were collected from affected and unaffected research participants and DNA was extracted using a Promega extraction protocol (Appendix B). In April 2012, DNA samples from 24 individuals (Table 3) were sent to Genome Québec, for WES using an Illumina sequencing platform. These 24 individuals come from 14 different families, from various communities in Newfoundland (Figure 10). For the purpose of this thesis, analysis of the WES was focused primarily on families R1136 (Figure 11) and R0942 (Figure 12).

Families R1136 and R0942 have been previously studied extensively and were selected, from all families sequenced using WES, for further analysis. Firstly, both families have multiple affected members with adequate DNA available. Both R1136 and R0942 have linkage data that has identified loci of suggested linkage, as well as telomere length assays and extensive clinical information that may be used to help narrow down lists of candidate variants. As shown in Figure 8B, family R0942 displays normal telomere lengths, while family R1136 have markedly decreased telomeres, suggesting that two different biological pathways may be involved in the development of IPF. By studying these two families, both of whom have extensive pre-existing data and may have two distinct biological causes for their disease, our understanding of the underlying biological pathophysiology of IPF may be enhanced. Other reasons for selecting these two families included recruiting large numbers of affected family members in each

family (family R1136: five affected members, family R0942: six affected members) and previously generated genotypic data.

2.2.2 Exome Capture

The following sections 2.2.2 – 2.2.5 are written based on information provided by Genome Québec regarding their specific sequencing and filtering methodology (Schwartzentruber, 2012).

The first step when targeting exonic regions of DNA for massive parallel sequencing involves the shearing of genomic DNA (gDNA) into double stranded fragments, approximately 200 base pairs in length. This is completed through ultrasonication. The sheared gDNA fragments are ligated to adaptors to create blunt end fragments and the inserts are amplified using PCR to create a genomic library. In order to capture exonic regions of the genome, DNA "baits" complementary to target sequences are hybridized to each genomic library fragment. These baits contain biotinylated uridine which has a strong binding affinity to the protein streptavidin. Additionally, streptavidin-coated magnetic beads are used to pull biotinylated DNA baits which are hybridized to the gDNA library, therefore isolating the target regions of interest. The fragments are then eluted, isolated and amplified using PCR, and further sequenced using the Illumina HiSeq platform.

R1136

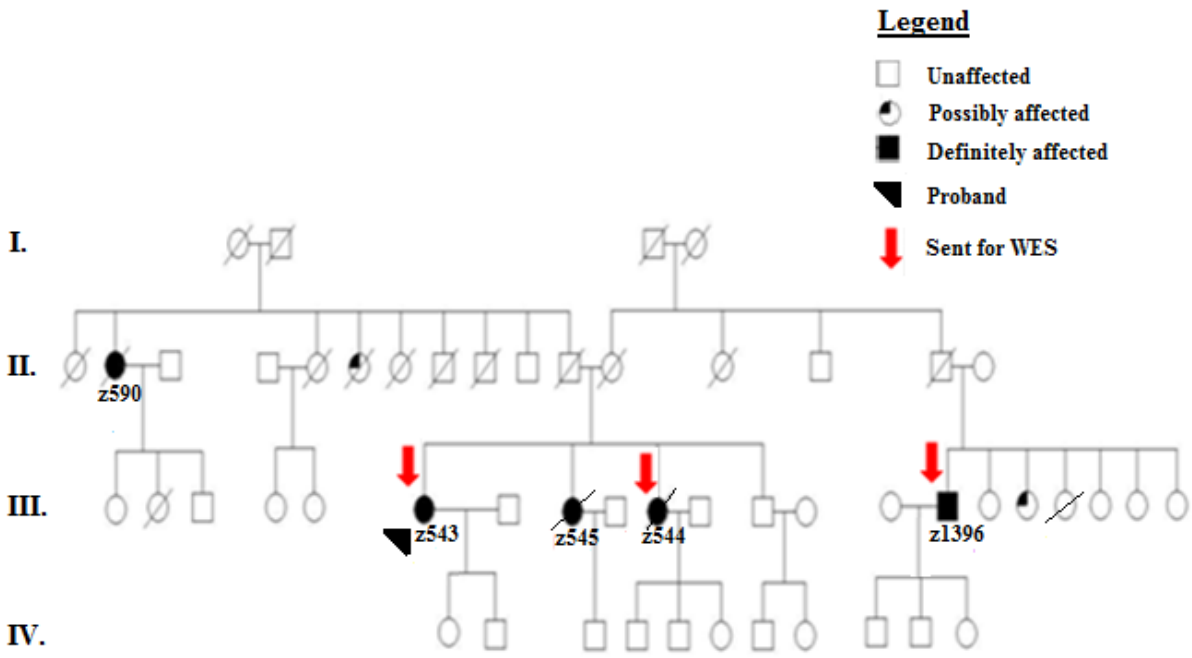


Figure 11: Pedigree for R1136

Family R1136 was selected for WES based on large family size, availability of DNA and pedigree structure. The red arrows indicate individuals who had WES through Genome Québec.

R0942

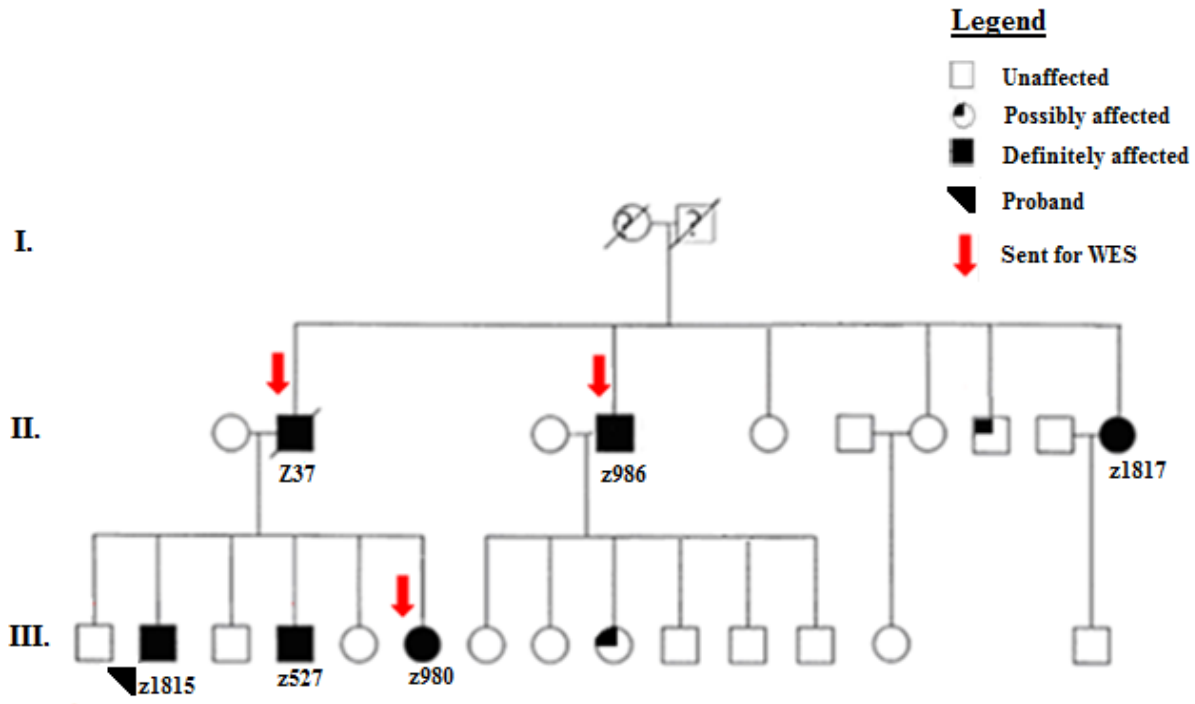


Figure 12: Pedigree for R0942

Family R0942 was selected for WES based on large family size, availability of DNA and pedigree structure. The red arrows indicate individuals who had WES through Genome Québec.

2.2.3 Primary Analysis: Illumina HiSeq Sequencing

The initial base calling in NGS is referred to as primary analysis. Illumina HiSeq uses a flow-cell to immobilize adaptor-ligated fragments and amplify each fragment. This process involves solid-phase bridge amplification, which creates approximately 1000 copies of each fragment in close proximity. Each flow-cell contains eight lanes, and after amplification, each lane contains anywhere from 60-80 billion bases be sequenced. This method, also known as Solexa sequencing, utilizes the incorporation of a single, reversible terminator fluorescently labeled base. The phosphate group that is released after incorporation of the labeled bases into the DNA strand releases a fluorescent dye specific for correctly incorporated base [either adenine (A), guanine (G), cytosine (C) or thymine (T)] based on the unique signal intensity emitted by each nucleic acid. The result is the sequencing of approximately 1000 copies of each captured exonic fragment. These individual sequenced fragments are called "reads" and are approximately 90 base pairs in length for Illumina HiSeq sequencing.

2.2.4 Secondary Analysis: Quality Control, Alignment and Coverage

Once initial base calling is performed by the Illumina HiSeq Sequencing platform, secondary analysis, involving quality control (QC), and reference sequence alignment is completed using the software programs Samtools (Li et al, 2009), Burrows-Wheeler Aligner (BWA) (Burrows et al, 1994), Genome Analysis Toolkit (GATK) (McKenna et al, 2010) and Dindel (Albers et al, 2011). The first set of QC parameters involves end trimming and is based on quality scores. Upon sequencing, each base is given a quality score which is built on the probability that the incorporated base is incorrect. This score is

calculated on a logarithmic scale and ranges from 2-41. For example, a score of 20 represents a 1% chance that the incorporated base is wrong. Typically scores greater than 30 are required to pass QC and these scores have been shown to decrease exponentially once read length surpasses approximately 90 bases for Illumina HiSeq sequencing. As a result, end-trimming is a conducted to increase the quality of the reads produced. This step also removes adaptor sequences which were ligated during exome capture. Additionally, reads are aligned to a human reference sequence, specifically the current GRChr38 human reference genome, a step that also produces quality scores, similar to base calling, and represents the chance that an error in read alignment has occurred.

Secondary analysis protocols implemented by Genome Québec also include marking duplicate reads. As previously mentioned, Illumina HiSeq sequencing based on solid-phase bridge amplification produces multiple copies of the same fragment. Single reads are marked as the actual read while the remainder are marked as duplicates. This is done using the NGS software Picard (Sarovich et al, 2014), which uses a set of tools to reorganize Samtool data files. QC parameters during reference sequence alignment also account for insertion and deletion (INDEL) realignment. Small INDELs may result in frameshifts which have an obvious effect on sequence alignment. INDEL realignment will fix read alignment to the reference sequence in genomic regions where INDELs have occurred. The entire exome capture, and primary and secondary analysis protocols are demonstrated in Figure 13.

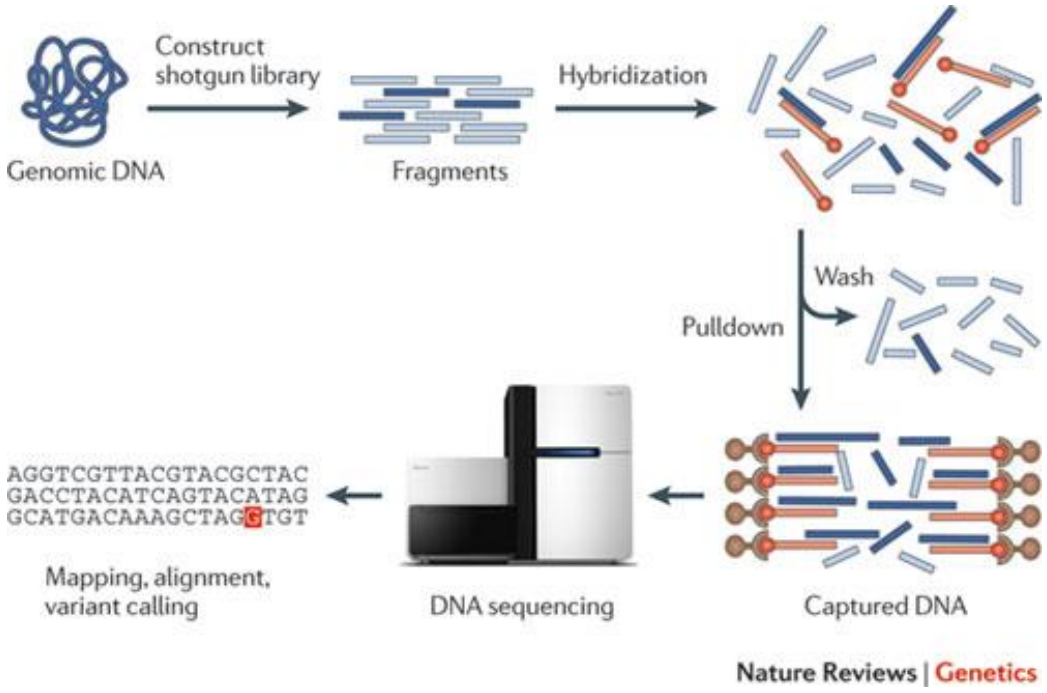


Figure 13: Workflow for Whole Exome Sequencing using Next Generation Sequencing. Reprinted from Bamshad et al. (2011) with copyright permission (Appendix E)

High throughput sequencing of targeted genomic regions, such as exomes, involves the initial construct of a DNA library using sheared gDNA. Targeted sequences are subsequently captured through the hybridization of adaptors specific to target gDNA loci. A strong binding interaction between streptavidin magnetic beads and biotinylated adaptors of the targeted DNA results in the capture of targeted DNA and elution of unwanted DNA fragments. These targeted DNA fragments are amplified by PCR and sequenced using a NGS platform, such as Illumina. The multiple reads produces by this method are then aligned to a reference sequence, subjected to quality control, nucleic acid base calling and variant calling.

Once reads are accurately aligned to the GRCh38 human reference sequence, the coverage for each base is determined. Coverage is a numerical factor for how many reads are aligned to a specific locus in the genome. It is calculated based on a formula that takes into account the read length of the original genome or exome (O), the number of reads (N) and the average read length (L) under the function: Coverage = N X L / G. As sequencing of varying loci differ depending on varying sequence motifs and guanine/ cytosine (GC) content, the average coverage for a gene varies from exon to exon. Additionally, each nucleodtide is given a Phred score, which measures the quality of the nnucledotide base sequence for each position. They are defined as a logarithmic property which is related to the base-calling error probabilities: $Q = -10\log_{10}P$, where Q is the Phred quality score and P is the probability that the base call is incorrect. For example, Phred score of 30 represents a 1 in 1000 likelihood that the base call at a specific position is incorrect. Phred scores of 30 is the cut-off used by Genome Québec and in this project.

2.2.5 Variant Call Format

Once primary and secondary analysis has been completed, sequences not matching the reference genome are highlighted and called as variants. Genome Québec uses the program Samtools to generate a variant call format (VCF). This program has the ability to call SNPs and INDELs, as well as remove potentially false positive variants. Genome Québec also uses the bioinformatic program ANNOVAR (Wang et al, 2010) to analyse variant calls. ANNOVAR will first determine the variant type, ie: if the variant is predicted to be a missense, nonsense, intronic, exonic, frameshift, splice site, etc (Wang et al, 2010). ANNOVAR also uses various databases, such as the single nucleotide

polymorphism database (dbSNP) (Sherry et al, 2001), the National Heart, Lung and Blood Institute (NHLBI) Exome Variant Server (NHLBI Variant Server, 2014), and 1000 Genomes (The 1000 Genomes Project Consortium, 2012) to determine the prevalence of the variants in control populations. Lastly, ANNOVAR incorporates other bioinformatic programs, such as SIFT (Ng et al., 2003), Polyphen2 (Adzhubei et al., 2013) and GERP scores (Cooper et al, 2005), which aid in predicting the structural, and thus functional effects that amino acid changes have on the protein function.

In a typical exome, between 15,000 and 20,000 variants will be called in a single individual (Nq et al., 2009; Stitziel et al., 2011). Of these variants, 10,000 are predicted to be non-synonomous, including missense, frameshift insertions, frameshift deletions, stop-gain and stop-loss variants, and thus will change the amino acid in the coding sequence. Non-synonomous variants have the potential to create a truncated protein if the variant in question results in a frameshift, occurs in a splice site or results in the introduction of a premature stop codon. I conducted this tertiary analysis using both the high- and moderate- impact lists generated by Genome Québec.

2.3 High Impact Variants

2.3.1 Filtering of High Impact Variants

Using the secondary analysis techniques described above, a moderate and high impact variant list was compiled and supplied in a Microsoft Excel spreadsheet by Genome Québec. The high impact list contained variants that were predicted to have the most dramatic effect on protein structure and included splice site and nonsense variants,

while the moderate impact list contained missense variants. I first analysed the high impact list following specific filtering criteria, as shown in Figure 14. The initial spreadsheet contained WES data on all 24 FPF samples sequenced. The information provided in this file include the reference and alternative alleles, chromosomal and genomic position of variants, the frequency of variants in 1000 Genomes, the read depth and coverage, functional information for each gene, SIFT, Polyphen2 and GERP scores, as well as a link to dbSNP for reported variants. The above variant information was initially used to filter all variants.

The WES data was provided in an Excel file, and contained the data for all individuals, but was not organized based on individuals belonging to the same family. Therefore, another filtering step involved grouping the data from affected individuals into families. I next removed all variants found in more than 20 of the 24 affected individuals (85%). This was done to remove common variants, as the likelihood that all families have the same mutation is low since IPF is a genetically heterogeneous disease and the families studied are not known to be related. I then removed any variants found in the 1000 Genomes database with a minor allele frequency (MAF) greater than 0.05, thereby eliminating common variants in the population. Finally, I analyzed variants segregating within each family. Specifically I searched for variants that were shared by all affected family members of a family. Although complete concordance amongst those sent for WES was ideal, variants found in multiple, but not all affected family members, were also investigated.

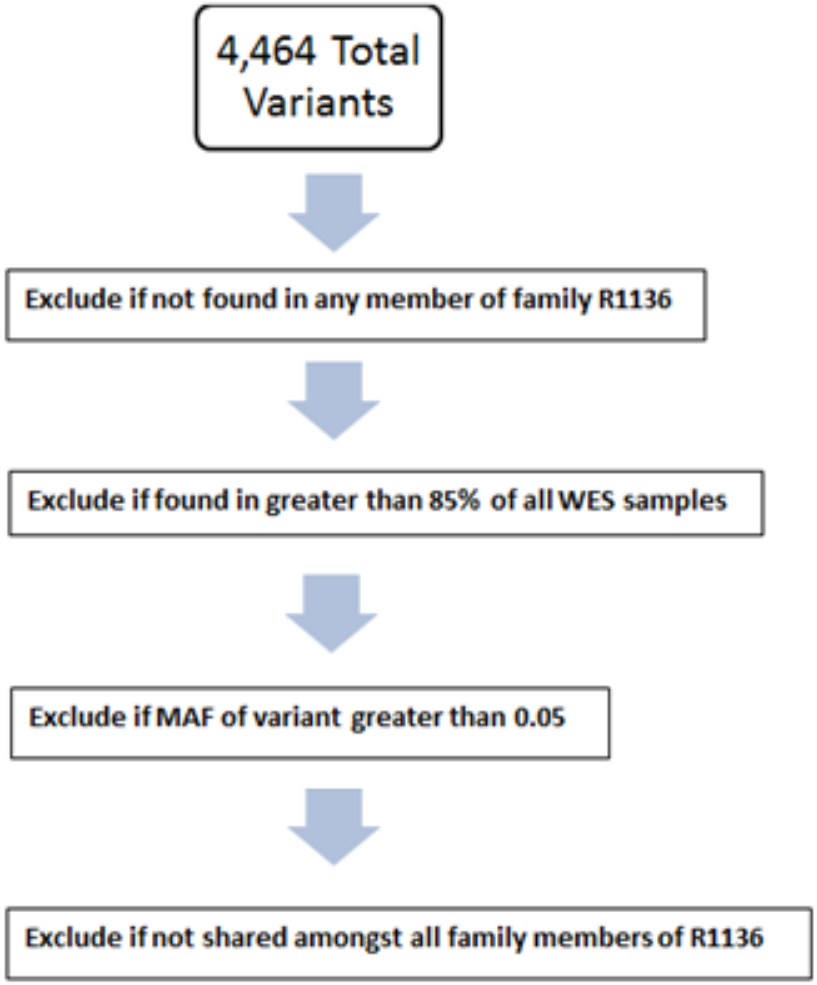


Figure 14: Criteria for Filtering of High Impact Variant List from Whole Exome Sequencing Data

A high impact variant list, containing nonsense and splice site mutations, was compiled by Genome Québec. Initial filtering of the 4,464 high impact variants consisted of grouping any variant found in any affected family member for specific families. Next, variants found in 85% or more of the samples sequenced were eliminated. Additionally, any variant found to have a MAF greater than 0.05 in 1000 Genomes was eliminated. Finally, variants found segregating among multiple family members for each particular family were included.

2.3.2 Filtering Based on Gene Function

After filtering familial sets of data using the methods outlined in Figure 14, I filtered variants based on gene function. Specifically, variants found in genes potentially involved in IPF developmental pathways, such as those with roles in aberrant wound healing, fibrotic growth pathways, mucosal secretion and those highly expressed in the lung were further investigated. To determine gene function and predicted biological pathways, the bioinformatics program the "Database for Annotation, Visualization and Integrated Discovery" (DAVID) was used. DAVID was developed by the Laboratory of Immunopathogenesis and Bioinformatics (Fredrick, Maryland, USA) and consists of a set of bioinformatic tools used to interpret large lists of genes derived from genomic studies such as microarrays, proteomic studies and exome analysis (Huang et al., 2009). Using DAVID, specific attention was paid to variants found in genes that interact with, or that in some other way are associated with the known IPF susceptibility genes: *TERT*, *TERC*, *SFTPA2*, *SFTPC* and *MUC5B*. Periodic review of current literature was fundamental in assessing any variants in newly associated IPF genes. For example, a recently published GWAS paper identified seven new IPF loci and confirmed existing associations for *TERT* (Fingerlin et al., 2013). Variants in these new loci were assessed in all families using the high and moderate impact lists as well as thorough analysis of the raw data, filtered using NextGENe software. Additionally, variants found in genes which are predicted to be benign based on their location in highly polymorphic were removed (Ju et al., 2011). Variants in these genes are predicted to be benign based on the their common occurrence in the general population. As IPF is thought to be caused by rare heterozygous mutations,

it is unlikely that the causative variants would be found in the exons of genes that are frequently mutated and found in healthy individuals.

2.4 Filtering Using NextGENe Software

2.4.1 Introduction to NextGENe

In addition to identifying potential causal IPF variants for both the high and moderate impact lists, raw data was filtered using the software program NextGENe version 3.4.2 produced by Softgenetics. Filtering of the raw data allowed for confirmation of the variants called in the high and moderate impact lists. Additionally, by using NextGENe, coverage and read depth for specific loci were assessed to determine the quality of specific reads. NextGENe is a software tool that allows for secondary and tertiary filtering of NGS data. It is compatible with a variety of sequencing platforms, including Solid, Roche and Illumina. NextGENe performs sequence alignment to a reference genome, specifically the 19th version of the Human Reference Genome, GRCh37. The software will also perform QC measures, such as trimming ends and filtering based on low coverage reads. After reads are aligned and QC is performed, base calls may be analysed using a variety of filtering techniques. Filtering can be done based on coverage, read depth and read length during the initial alignment phase.

2.4.2 NextGENe Filtering Steps: Secondary Analysis

In January of 2013, raw data was made available on the external server Nanuq, provided to us by Genome Québec. Raw data files for all 24 samples were available in both the forward and reverse run. Both forward and reverse files were downloaded for all

families and were approximately 3 million kilobytes (Kb) each in size. These files were downloaded in the FastQ format. The first step in uploading raw data into NextGENe requires the conversion of FASTQ files to FASTA format. Both the forward and reverse converted FASTA files were loaded into the NextGENe input, along with a file containing version 19 of the Human Reference Genome, GRCh37, to align reads to the human reference sequence. The next step enabled modification of filtering criteria, through adjustment of read length, coverage and read depth. Default parameters were utilized and included capturing coverage of at least 10X for each read, greater than 85% base matching for each individual read and a read length of at least 50 base pairs. Lastly, an output file was selected, where the raw data analysis could be found upon completion. The entire alignment and QC took approximately three hours per two samples analysed.

2.4.3 NextGENe Filtering Steps: Tertiary Analysis

Upon alignment, additional filtering was assessed using the mutation report tool. This tool allows for filtering of coding DNA (cDNA), messenger RNA (mRNA), missense, silent, splice site and nonsense variants, as well as variants which have or have not been reported in dbSNP. The version of NextGENe used in this study also allows for incorporation of external databases, such as COSMIC: Catalogue of Somatic Mutations in Cancer, as well as the Database for Non-synonymous SNPs' Functional Predictions (dbNSFP) version 2.1. I decided to use dbNSFP for the analysis of the IPF raw data as it incorporates the bioinformatic programs SIFT, LRT, MutationTaster, FATHMM, Mutation Assessor and Polyphen2 which are helpful in evaluating the effects of variants on DNA sequence, protein structure and function. In conjunction with the above filtering

tools, a variant comparison tool (VCT) was also used to compare raw data from multiple samples within the same family. The VCT permits filtering based on shared mutations, different mutations, mode of inheritance and multiple mutations in the same gene. This tool was used frequently to compare raw data from samples within a family and verify variant calls from Genome Québec.

2.5 Sanger Sequencing of Candidate Genes

Upon identification of potential causal variants using the above filtering criteria, verification of variants in samples, as well as investigation of variants in control samples was conducted using Sanger sequencing. This protocol involves the amplification of a locus of interest using PCR, purification of PCR product using an Exonuclease (EXO) and Shrimp Alkaline Phosphatase (SAP) solution, and sequencing of the purified DNA strand using an ABISeq protocol in a temperature controlled thermocycler machine. All primer sequences and thermocycler protocols for each variant or exon sequenced can be found in Appendix C and D, respectively.

2.5.1 Polymerase Chain Reaction Protocol

PCR was used in the amplification of DNA fragments. Briefly, gDNA was used, either from stock, at a concentration of 100 ng/µl, or diluted using ultra-pure deionized water to a final concentration of 50 ng/µl. All materials for each PCR reaction were supplied by Invitrogen Canada Inc, 5250 Mainway, Burlington, ON, L7L 5Z1). For each reaction, a cocktail of 10.225 µl of deionized water, 1.5 µl of 10X PCR Reaction Buffer, 0.375 µl of 100 ng/µl nucleotide triphosphates (dNTPs), 0.5 µl of 10 µM forward primer,

0.5 µl of 10 µM reverse primer, 0.75 µl of 50 Mm of MgCl, 0.15 µl Platinum Taq DNA Polymerase and 1.00 µl of gDNA was prepared and added to a 96 well PCR plate along with negative controls. Each plate was capped, vortexed, centrifuged and placed in an Eppendorf or Biometra Thermocycler. Program protocols were determined for each primer set, as outlined in Appendices C and D. Upon completion of the PCR protocol, product amplification was verified using gel electrophoresis. In a flask, 50 µl of 1X TBE Buffer was added to 1.00 g of Ultrapure agarose, heated for 75 seconds and allowed to cool briefly. Additionally, 3.75 µl of SYBR Safe DNA Stain Gel (Life Technologies, Ontario, Canada) and the mixture was allowed to cool to touch before setting in the gel chamber with a well comb. 3.5 µl of 3X loading dye was added to each well, along with 3.0 µl of PCR product. 1 µl of DNA ladder was also added to 3.5 µl of loading dye for comparison of band size. The gel was placed in a gel chamber and ran for 30 minutes at 130 V. An image of the gel was taken using an AlphaImager. PCR was considered successful if there was no contamination in the negative PCR control and if predicted band size was produced.

2.5.2 Exonuclease /Shrimp Alkaline Phosphatase

Upon successful PCR amplification, purification of the PCR product was performed to remove unwanted salt reagents as well as phosphates. This was completed using a cocktail mixture of 0.5 µl of EXO (10 U/µl), 0.5 µl SAP (1 U/µl), and 7.5 µl of water. In a new 96 well PCR plate, 8.5 µl of ExoSap cocktail mixture was added, followed by 8.0 µl of PCR product. The plate was capped, vortexed and centrifuged

briefly and placed in an Eppendorf thermocycler under the "ExoSap" program (Appendix D).

2.5.3 ABISeq Protocol

For successful Sanger sequencing to occur, an ABI sequencing mixture was added to the EXOSAP/PCR mixture. Briefly, two cocktail mixtures were prepared: both consisting of 0.5 µl of BigDye Terminator v3.1 Cycle Sequencing Mix, 2.0 µl BigDye Terminator v3.1 5X Sequencing Buffer (Applied Biosystems, Ontario, Canada) and 13.83 µl of dH_2O. To one mixture, 0.67 µl of 250 ng/ µl of forward primer was added, while the second mixture contained 0.67 µl ng/ µl of reverse primer. These mixtures were vortexed and 19 µl of each was added to a 96 well sequencing plate. Additionally, 4 µl of "EXOSAP" product was added to each well. The plate was capped, vortexed, briefly centrifuged and placed in an Eppendorf thermocycler under the ABIseq protocol (Appendix D).

Ethanol precipitation was used to precipitate the DNA from its aqueous solution. Upon completion of the ABI sequencing cycle, 65 µl of 95% ethanol and 5 µl of 0.125 µM of EDTA was added to each well. The plate was capped and set aside in the dark for between an hour and overnight. Upon precipitation, the plate was centrifuged for 45 minutes at 3000 g. After spinning, the ethanol was decanted from the wells. The plate was inverted over paper towel and centrifuged at 200 rpm to remove all ethanol. Additionally, 150 µl of 70% ethanol was added to each well, the plate was capped and centrifuged for 15 minutes at 3000 g. Ethanol was decanted, centrifuged at 200 rpm and

allowed to dry in the dark for 30 minutes. Upon complete evaporation of the ethanol, 10 µl of High Di Formamide was added to each well. A septa was added to each plate, the plate was vortexed, centrifuged and placed in the thermocycler under the denaturing program (Appendix D). The plate was placed in an ABI 3130 Prism Sequencer and raw sequencing data of the desired PCR product was produced. The data generated for all affected, unaffected and control samples sequenced was analysed using Sequencher 5.0 program.

2.5.4 Control Samples

Sanger sequencing of control samples was performed to determine the frequency of potentially causal variants in the Newfoundland population. These control samples were obtained from the Newfoundland Colorectal Cancer Research (NFCCR) study. Controls were ethnically matched, with no personal history of cancer and were obtained by random digit dialing for the purpose of the NFCCR (Green et al., 2007). By using Newfoundland control samples, the prevalence of variants in the Newfoundland population was assessed to gain a better understanding of the rarity of specific variants captured using WES.

3.0 Results

3.1 Variants Calls from High Impact List Generated from 24 Affected Patients

3.1.1 Sequencing of Previously Associated Pulmonary Fibrosis Genes

Results from the WES analysis were first obtained in December 2012. Initial analysis of the WES included the investigation of the high impact variant list provided by Genome Québec, which contained 4,464 nonsense and splice site variants found in any of the 24 affected FPF patients.

I first searched for variants found in genes previously associated with IPF. An *HLA-A* variant, c.delG, was identified in 18 affected patients. Because of the previous association with IPF, Sanger sequencing of the *HLA-A* variant was conducted in affected samples. From the high-impact list, a total of 18 of the 24 patients were predicted to have the above *HLA-A* variant. Sanger sequencing of the 18 affected samples as well as 41 controls did not identify the variant in any samples sequenced. Since the variant was not previously reported, it is likely that the variant call was a false-positive and was subsequently eliminated from the list of potential causative variants.

Additionally, two intronic variants were found in all 24 affected samples in a known IPF gene, *SFPTA2*. One of these variants was a predicted splice acceptor site, g.79559511C>T. This variant was previously reported by dbSNP and 1000 Genomes, and was shown to have a MAF of 0.34. Because of the high frequency of the variant in control populations, it was excluded from the analysis. The second intronic variant, g.79559667C>T was predicted to be a splice donor site and was previously reported with

a MAF of 0.06. Because the MAF was close to elimination threshold of 0.05 and the fact that it was found in a gene previously shown to cause IPF when mutated, further investigation was warranted.

The second intronic SFTPA2 variant was confirmed by Sanger sequencing in 7/24 patients who had WES: z543 and z544 from R1136, z1647 from R2416, z1547 from R2306, z659 and z660 from R1351 and z616 from R1211 (see Table 3). Upon these results, Sanger sequencing of NFCCR controls was conducted to determine the prevalence of this variant in the Newfoundland population. A total of 31 controls were sequenced and the variant was found in 25 of the 31 controls. Because it was found in 25 control samples with no known history of IPF, I also excluded this variant from analysis. As there were no other variants in genes previously associated with IPF, I began filtering the high impact variant lists to identify potential causative variants in novel IPF genes.

3.1.2 Filtering of High Impact Variant Lists

A list of high impact variants found in all 24 affected IPF patients was provided by Genome Québec. This high impact list also contained bioinformatic results and MAF data based on in-house analysis. This information was utilized in the initial analysis of the high impact variants. Following the filtering criteria outlined in Figure 14, six families containing more than one affected family member who had WES were selected for filtering: R1136 (Figure 11) and R0942 (Figure 12), as well as R1351, R1224, R1487 and R0896 (Appendix G). Initial filtering, which included grouping individuals into families, generated multiple high impact lists and resulted in 1,103 variants for R1487,

1,133 variants for R0942, 1,120 variants for R0896, 883 variants for R1224, and 1,149 variants for R1351 (Table 4).

Upon grouping individuals into families, variants were eliminated based on those found in 85% or greater of individuals. This was done as it is unlikely that all families have the same genetic cause of IPF, given the clinical and genetic heterogeneity that has been observed. As Table 4 demonstrates, the number of variants left at this stage of filtering included 864 variants in R1487, 894 variants in R0942, 881 variants in R1136, 1,011 variants in R0896, 646 variants in R1224 and 1,149 variants in R1351. The number of variants in each family was reduced again based on a MAF in 1000 Genomes less than 0.05. Subsequently the lists were reduced to 826 variants for R1487, 855 variants for R092, 841 variants for R1136, 961 variants for R0896, 613 variants for R1224 and 879 variants for R1351. Finally, I eliminated variants that were not in every affected member of each family, thereby identifying shared variants in each family. This reduced the number of variants in each family to 178 variants for R1487, 168 variants for R0942, 132 variants for R1136, 89 variants for R0896, 244 variants for R1224 and 263 variants for R1351. Although there were variants overlapping in multiple families, some variants were unique to each family. High impact variant lists for families R0896, R1351, R1487 and R1224 can be found in Appendix H. Although all families were analysed using the above filtering methodology, families R1136 and R0942 were shown to have more compelling variants of interest. Because of this, as well as the fact that the majority of work on IPF in the Woods lab has been conducted in these two families, both R1136 and R0942 are the focus of the thesis hereafter.

3.2 Filtering of High Impact Variants in R0942

3.2.1 Initial Filtering of R0942

Upon filtering the high impact variant list using the filtering methodology described in section 2.3 (see Figure 14), initial analysis highlighted 168 variants that passed filtering criteria in family R0942. These variants were found in all three affected family members sequenced using WES (z37, z986, and z980, see Figure 12), were found in fewer than 85% of total samples sequenced and had a frequency less than 5% in 1000 Genomes. The bioinformatics program DAVID was used to annotate the 168 genes which helped to identify Gene Ontology (GO) terms consistent with involvement in IPF pathways, for example, expression in the lung, cell-cell adhesion, fibrotic growth, immune response, and ECM involvement (Table 5).

3.2.2 Elimination of Variants and Candidate Gene Selection

As shown in Table 5, a total of 14 variants were highlighted as having potential roles in IPF development. Of these 14, ten were eliminated based on a MAF greater than 0.05 in either dbSNP or NHLBI Exome Variant Server. Additionally a variant found in *MUC3A* was not further pursed based on the highly polymorphic nature of this gene. The remaining variants were selected for Sanger sequencing, ie: variants found in inteleukin 32 (*IL32),* fibroblast growth factor 4 (*FGFR4)* and cytokine induced apoptosis inhibitor 1 (*CIAPIN1)*. Results for both *IL32* and *FGFR4* are described in the following sections. Due to technical difficulties, confirmation of the variant found in *CIAPIN1* was not completed, but will be conducted in future work.

Table 4: Filtering of High Impact Variant List from Whole Exome Sequencing Data in Six Newfoundland Families with Idiopathic Pulmonary Fibrosis

Family	**R1487**	**R0942**	**R1136**	**R0896**	**R1224**	**R1351**
Total variants found in all affected members of each family	1103	1133	1120	1250	883	1149
Total variants found in 85% or less of individuals (20 out of 24)	864	894	881	1011	646	911
Total variants with a MAF <5% in 1000 Genomes	826	855	841	961	613	879
Total variants shared in all members of the same family	178	168	132	89	244	263

Table 5: Fourteen Genes of Interest in R0942 using DAVID to Annotate High Impact Variant List

Gene	Chromosome Location	Genomic Position	Gene Function	Variant Information
ADAMST1	21q21.3	28216059	Disintegrin and metalloproteinase with thrombospondin motif; Involved in inflammatory process	A>G Intronic variant; *rs400852* MAF:0.21
IL32	**16p13.3**	**3119297**	**Member of cytokine family, expression increased after activation of T cells; Induces production of TNF-α (TNF-α implicated in PF)**	**InsG Frameshift *rs144971189* MAF: unknown (11/24 WES samples)**
EGFR	7p11.2	55214348	Epidermal growth factor receptor; mutations and overexpression implicated in lung cancer	C>T Splice site variant *rs2072454* MAF 0.45
CIAPIN	**16q21**	**57463181**	**Cytokine induced inhibitor of apoptosis; dependant on growth factor stimulation**	**G>A Nonsense variant *rs187678010* MAF<0.01 Variant only in R0942 (3/24 WES samples)**
COLQ	3p25.1	15515695	Acetylcholoesterase collagenic tail peptide; collagen like molecule associated with acetylcholinesterase in skeletal muscles	C>T Splice site variant *rs2305016* MAF:0.46
DNAH11	7p15.3	21582963	Member of dynein heavy chain family; involved in respiratory cilia; mutations implicated in Kartagner syndrome	G>T Nonsense variant *rs2285943* MAF:0.38
FGFR4	**5q35.1**	**176517136**	**Fibroblast growth factor receptor4; Role**	**InsGTGT Splice site variant**

			in angiogenesis, wound healing; overexpression may lead to cell transformation and cancer; mutations associated with human developmental disorders	***rs144400190*** **MAF: N/A (14/24 WES samples)**
GSDMB	17q21.2	38064469	Gasdermin-like protein; implicated in regulation of apoptosis, linked to cancer	T>C Splice site variant *rs11078928* MAF: 0.33
GZMB	14q11.2	25103414	Encodes enzyme responsible for cell lysis during cell-mediated immune response; apoptosis execution	G>A Nonsense variant *rs22738414* MAF:0.28
HYDIN	16q22.2	71100838	Gene encodes protein involved in cilia motility	T>C Intronic variant *rs74361942* MAF:0.28
ITGB2	21.q22.3	46328099	Integrin protein; involved in cell adhesion	C>T Splice site variant *rs760462* MAF: 0.14
MREIIA	11q21	94225807	Involved in telomere length maintenance, homologous recombination	C>T Splice site variant *rs496797* MAF:0.47
MUC3A	7q22.1	100552738	Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung	C>T Nonsense variant *rs79874934* MAF: N/A
PRAM1	19p13.2	8567475	Protein involved in T cell receptor mediated signalling; neutrophils	T>C Start loss missense *rs968502* MAF: 0.30

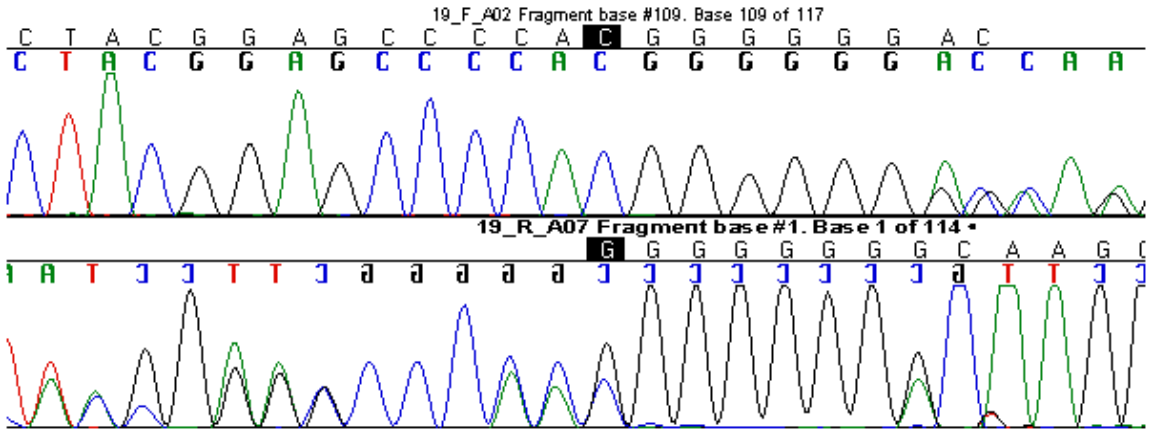
3.2.3 IL32

Through filtering of the high impact variant list, a previously reported insertion variant, c.508_509insG, was found in *IL32* in all three affected individuals who had WES in R0942. Sanger sequencing was performed on all samples predicted to have the c.508_509insG variant by WES, which included three members of R0942 (z986, z9890, z37, see Figure 12), as well as z690 (R1224), z1547 (R2306), z616 (R1211), z1470 (R1941), z647 (R1275), z1392 (R1460), z529 and z618 (R0896, see Table 3).

Sequencing chromatogram results, as shown in Figure 15, confirmed the variant in all 11 affected samples that were tested. Additionally, there was a second variant uncovered, c.508_509insG, further downstream that segregated within all samples tested. Both of these insertions occurred in a region rich in G repeats.

To further investigate both *IL32* variants, I tested a set of 28 NFCCR control samples. Sequencing results demonstrated the c.508_509insG variant in eight control samples. For the second variant, c.516_517insG, Sanger sequencing of the 28 control samples positively confirmed the presence of the variant in either the forward or reverse direction for all samples. Because these results were seen at a high rate in Newfoundland controls, these two variants were eliminated.

A)



B)

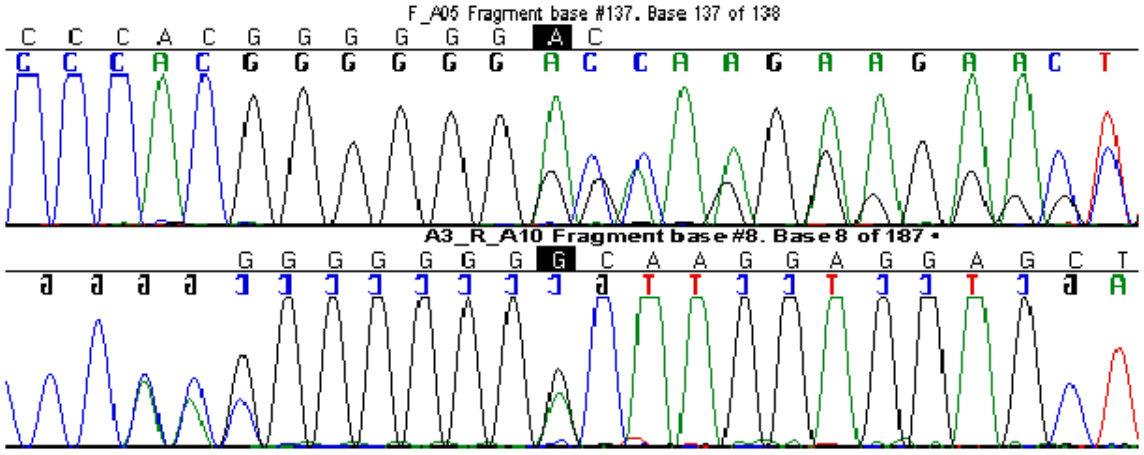


Figure 15: Sanger Sequencing of *IL32* variant in an Affected Idiopathic Pulmonary Fibrosis Patient (A) and a Newfoundland Control (B)

Sanger sequencing of both the IPF cases (A) and controls (B) revealed insertion of guanines at two positions. This resulted in a frameshift and subsequent termination of base calls.

3.2.4 FGFR4

Sanger sequencing was performed on all 6 affected members of R0942 (Figure 12) to further investigate the insertion variant, c.92-255_92-254insGTGT, found in *FGFR4*, a receptor for fibroblast growth factor with a functional role in angiogenesis and wound healing. Sanger sequencing of all affected (6) and unaffected (11) members of R0942 confirmed the variant in the three affected members found using WES; however, it was also present in all the unaffected members who were sequenced. As well, this variant was not found to segregate with disease phenotype in other families. For example, WES data shows the variant to be present in one member of R1487 (z902), but was not found in the other two family members sent for WES (z900 and z1373). Because the *FGFR4* variant did not segregate with disease phenotype in either R0942 (upon Sanger sequencing) or in R1487 upon (interpretation of WES data), it was eliminated from the list of potential causative variants.

3.3 Filtering of Moderate Impact List in R0942

3.3.1 Initial Filtering of Moderate Impact List

To further investigate the WES data, a moderate impact list was provided by Genome Québec. Provided in an Excel file, this moderate impact list consists of 56,692 missense variants for the 24 DNA samples and are predicted to have a less damaging effect compared to the high impact variants. Similar to the high impact variant list, initial analysis was based on bioinformatics, MAF and functional gene annotation information provided within the moderate impact list. As this project aims to find novel IPF associated genes, I initially attempted to filter based on the same methodology used for

the high impact variant list (filtering based on presence in 1000 Genomes, prevalence in less than 85% of samples, variants shared by all family members); however, upon initiation of this method, it was apparent that there were too many variants to investigate individually.

Alternatively, I decided to focus on variants in genes previously known to be associated with the IPF developmental pathway. Specifically I focused on genes previously associated with IPF, such as those involved in surfactant protein production and the telomerase complex. Table 6 shows a list of variants selected based on their functional importance and potential relation to the IPF pathway. All variants highlighted had relatively low or unreported MAF. To investigate these variants further, I looked for variants that were found in multiple members of the same family. I excluded common variants found in multiple families but not segregating with the disease in each family. Of interest were variants found in *TERF1*, telomerase protein component 1(*TEP1)* and telomere maintenance interacting protein 1 (*TTI1).* These three genes encode proteins involved in the telomerase complex and help to maintain telomere length. The *TTI1* variant, c.484G>T; p.A162T, was not further pursued, as it was only found in one patient, z902 of family R1487. Although by WES the variant did not segregate with affected phenotypes, it was not ruled out as it is possible this variant is a modifier and may have an effect on disease progression. A second variant in *TTI1*, c.3083G>T, p.R1028K was also not further pursued as it did not segregate in all families, only R1136, and was predicted to be benign by both PolyPhen2 and SIFT (Table 6). Additionally, variants found in both *TERF1* and *TEP1* were highlighted and further investigated, as discussed in

the following sections. A c.311G>T missense variant was found in *TERF1* (Table 6) in R0942, z37 and z980 (Table 3, Figure 12). Although this variant was initially shown not to segregate with the disease in R0942, it was further investigated due to its critical role in telomerase function, predicted pathogenicity and the potential that multiple genetic variants may have a role in IPF development in this family.

Table 6: Thirty-Four Moderate Impact Variants from Whole Exome Sequence Data Filtered Based on Previously Associated Idiopathic Pulmonary Fibrosis Genes or Predicted Pathways

Gene	Allele	MAF	Amino Acid	Poly-phen 2	SIFT	GERP	Number of Sample Positives (X/24)	Gene Description
ABCA3	C>T	NR	A852T	D,D	0	5.54	1	Surfactant production
ABCA3	C>T	NR	V659M	B,B, B	0.03	6.17	1	Surfactant production
ABCA3	G>T	NR	Q6K	B,B, D	0.64	4.81	1	Surfactant production
SFTPA2	C>G	0.02	V50L	B	0.63	2.91	3	Surfactant protein
SFTPB	G>T	NR	A330D	B,B	0.64	4.86	1	Surfactant protein
SFTPC	C>A	0.21	T138N	P,B, P	0.09	5.49	8	Surfactant protein
TELO2	G>T	NR	A240S	B	0.74	4.87	1	Telomere maintenance
TELO2	G>T	NR	G372V	B	0.28	5.01	1	Telomere maintenance
TELO2	G>T	NR	E529Q	D	0.03	5.3	1	Telomere maintenance
TELO2	C>T	NR	R530C	D	0.01	5.3	1	Telomere maintenance
TELO2	G>C	NR	E532D	D	0.07	5.3	1	Telomere maintenance
TELO2	G>T	NR	A548S	D	0.4	5.3	1	Telomere maintenance
TELO2	C>A	NR	L234M	B	0.25	4.62	1	Telomere maintenance
TELO2	C>G	NR	L302V	P	0.38	5.15	1	Telomere maintenance
TELO2	C>A	NR	P305T	D	0.33	5.15	1	Telomere maintenance
TELO2	C>T	N/A	R170C	B	0.2	5.33	1	Telomere maintenance
TELO2	G>A	<0.01	S317N	B	0.66	4.86	1	Telomere maintenance

TEP1	C>A	NR	W2001L	D,D,D	0.13	5.62	1	Telomerase protein
TEP1	G>T	NR	P1121T	B,B,B	0.21	5.84	1	Telomerase protein
TEP1	T>G	NR	H462P	D,D	0.01	5.43	1	Telomerase protein
TEP1	G>A	<0.01	R1278W	P,P,P	0	5.44	2	Telomerase protein
TEP1	C>T	<0.01	R500Q	B,B	0.48	5.7	2	Telomerase protein
TEP1	T>C	0.05	K368R	D	0.13	5.69	2	Telomerase protein
TEP1	C>T	0.03	R1047Q	B,B,B	1	5.62	1	Telomerase protein
TEP1	C>T	0.03	R110Q	P,D,D	0.26	5.76	2	Telomerase protein
TEP1	T>C	0.03	H2454R	D,D	0.6	5.81	3	Telomerase protein
TERF1	G>T	N/A	S104I	B,D	0.73	4.85	2	Telomerase protein
TERT	C>A	NR	C89F	P,P,P	0.06	3.59	1	Telomerase protein
TERT	C>T	<0.01	A779T	.	0.61	4.26	2	Telomerase protein
TERT	C>T	0.01	A279T	B,B,B	0.29	3.43	1	Telomerase protein
TTI1	C>T	<0.01	A162T	P	0.39	5.35	1	Telomerase protein
TTI1	C>T	0.02	R1028K	B	0.29	5.3	4	Telomerase protein
TTI2	G>A	<0.01	R324C	D,D	0	5.5	1	Telomerase protein
ZNF268	G>A	0.01	R163H	.	.	3.79	3	Telomerase associated zinc finger

*Refers to prediction of amnio acid change on protein structure

D = Probably Damaging; P = Possibly Damaging; B = Benign

3.3.2 TERF1

TERF1 encodes Telomeric Repeating Binding Factor, a protein involved in the negative regulation of the telomerase complex. Because of its involvement with the telomerase complex and telomere maintenance, it was identified as a potential IPF causative gene, as the genes *TERT* and *TERC*, whose functions are also essential for telomere maintenance, have been previously implicated in IPF. Although this missense variant (c.311G>T, p.S104I) was not found in all three family members from R0942 sequenced using WES, it was further pursued as the function of this gene may be a highly important contributor to telomere maintenance within a polygenic model. Results from Sanger sequencing confirmed the presence of the *TERF1* variant in both z37 and z980, as well as the absence of the variant in z986. Additionally, the variant was shown to segregate in the nuclear family of R0942 (Figure 16), and was confirmed in the proband (z1815), the affected brother of the proband (z527), the affected sister of the proband (z980) and the unaffected sister of the proband (z502). As shown in Figure 16, the variant was not found in two affected family members, z986 and z1817, or any of their unaffected children. This variant is reported in both dbSNP and NHLBI Exome Variant Server with a MAF <0.01. It is predicted to be pathogenic by both PolyPhen2, with a score of 1.00, and Grantham, with a score of 142, suggesting the amino acid change from a serine to isoleucine most likely has a damaging effect on protein structure and function. The wildtype allele is highly conserved, with a GERP score of 3.9, indicating evolutionary conservation.

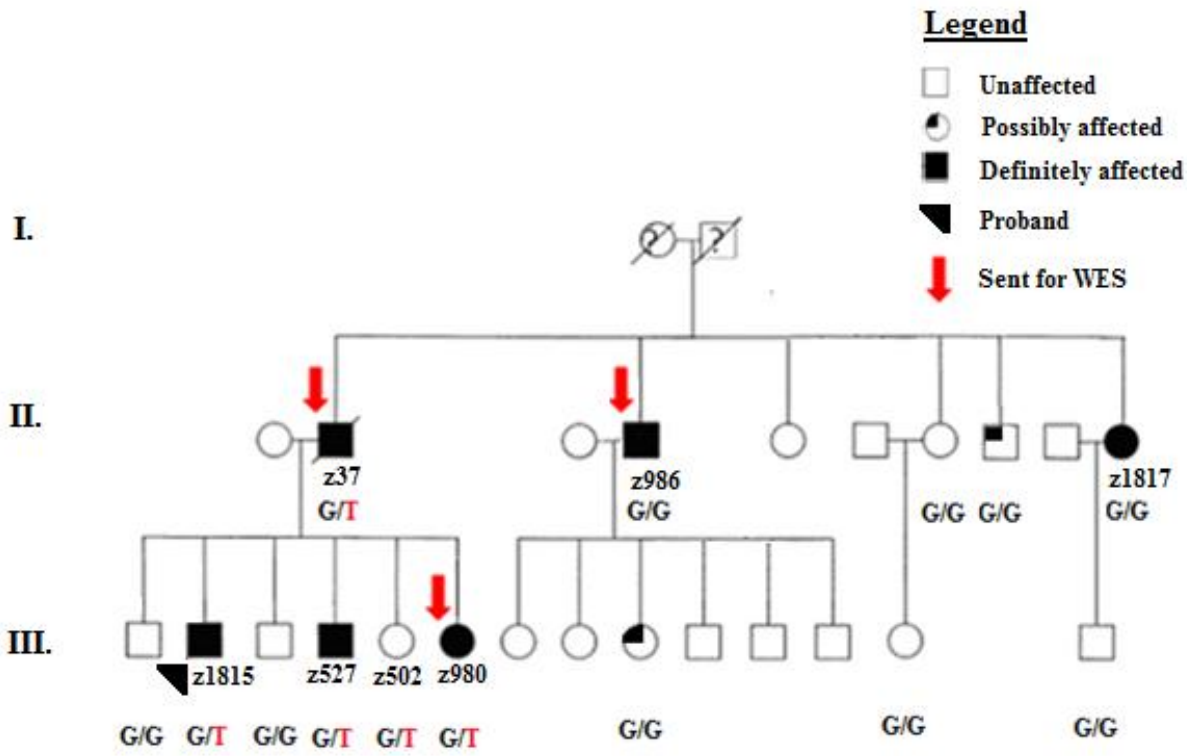


Figure 16: Segregation of *TERF1* Missense Variant in R0942

Sanger sequencing of all members of R0942 revealed a rare missense variant, c.311G>T, segregating in the nuclear family of R0942. This rare missense variant (MAF< 0.01), results in a serine to isoleucine amino acid change (p.S104I) and is predicted to be pathogenic by PolyPhen2 (score 1.0), with a GERP score of 3.9, suggesting strong conservation at this locus for the wildtype allele.

As this variant was shown to segregate in the nuclear family, Sanger sequencing of NFCCR control samples was conducted to assess the frequency of the variant in the Newfoundland population. A total of 40 controls were sequenced, with the variant not found in either the forward or reverse direction for any of the controls sequenced.

3.4 Filtering of High Impact Variants in R1136

3.4.1 Initial Filtering of R1136

For the filtering of the high impact variant lists involving variants called in members of R1136 (Figure 11), filtering strategies were similar to that of R0942 (Figure 12, Figure 14). Upon filtering the high impact variant list for R1136, a total of 132 high impact variants of interest were identified. The next step in filtering of the high impact variants was annotation of these variants based on gene function. To evaluate the gene function, tissue expression and biological pathway of the genes in which these variants were found, the bioinformatic program DAVID was used. By searching for specific GO terms associated with IPF development (expression in the lung, cell-cell adhesion, fibrotic growth, immune response, extracellular matrix), a list of variants and candidate genes was compiled for R1136, as shown in Table 7, consisting of nine variants of interest. These variants were further investigated using dbSNP, NHLBI Exome Variant Server and Ensembl to determine the MAF and chromosomal location.

Table 7: Nine Genes of Interest in R1136 using DAVID to Annotate High Impact Variant Lists

Gene	Chromosome Location	Genomic Position	Gene Information	Variant Information
EGFR	7p11.2	55214348	Epidermal growth factor receptor; Ubiquitously expressed; associated with non-small cell lung carcinoma; may play a role in the malformation of pulmonary airways	C>T Synonymous variant *rs2072454* MAF: 0.45
MUC3A	7q22.1	100552738	Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung, involved in epithelial structure maintenance	C>T Nonsense variant *rs79874934* MAF: 0.01
HLA-DRB5	6p21.32	32487426	Part of MHC; involved in immune response; Associated with sclerosis and sarcoidosis	G>C/T/A Nonsense variant *rs1071751* MAF: 0.49
DUOX1	15q15.3	45457127	Dual oxidase 1; Associated with MS and sarcoidosis; mRNA greatly expressed in fibrofatty lesions; Critical role in mucin expression in airway epithelial; Regulated expression of DUOX1 during alveolar maturation; Role in microbial defense	C>T Splice site variant *rs1769193* MAF: 0.31
ITGB2	21q22.3	46328099	Integrin beta 2 (complement component 3 receptor): Associated with leukocyte adhesion deficiency; recurrent bacterial infections, deficient in adhesion	T>C Splice site variant *rs760462* MAF: 0.14

CD109	**6q13**	**74476710**	**Cell surface antigen: regulation of TGF-β signalling: High expression in lung, trachea, aorta, placenta, uterus**	**C>T Nonsense variant Not reported (3/24 WES samples)**
PDGFRB	5q32	149501688	Platelet derived growth factor receptor; Cell surface tyrosine kinase receptor; important in cell proliferation, cellular differentiation, growth; High levels found in lung fibrosis patients	Del CTCA Splice site variant *rs56202461* MAF: 0.08
DSP	**6p24**	**7542148**	**Desmoplakin; important in anchoring adjacent cells; associated with ARVCD and autoimmune diseases; cardiac fibrosis. Keratosis; recently identified as an IPF susceptibility loci**	**Ins A Frameshift *rs66469215* Initally no MAF data; later MAF: 0.19 (7/24 WES samples)**
GSDMB	17q21.2	38064469	Gasdermin B; Localization in apical region of gastric chief cells and colonic surface mucous cells; Polymorphisms may contribute to childhood asthma	T>C Splice site variant *rs11078928* MAF: 0.33

3.4.2 Elimination of Variants and Candidate Gene Selection

As shown in Table 7, a list of nine variants was compiled from the high impact variant list based on segregation within affected family members of R1136, predicted deleterious effect on protein structure and low prevalence in the 1000 Genomes database. These variants were also selected based on gene function, as these genes were thought to potentially play a role in IPF development. Further investigation into these nine variants involved the databases dbSNP and NHLBI Exome Variant Server to determine MAF. Any variant with a MAF greater than 0.05 was eliminated, thus reducing the number of potential variants from nine to three variants, initially including variants in desmoplakin (*DSP*), *CD109* and mucin 3 A (*MUC3)A.* Because mucin genes are highly polymorphic, the *MUC3A* variant was not further pursued. The remaining variants in *DSP* and *CD109* were chosen for further investigation. At the time of filtering, there was no publically available information regarding the MAF of the variant in the gene *DSP*, which has recently been shown to be associated with IPF (Fingerlin et al., 2013). As there was initially little known about the frequency of this specific variant in *DSP*, and *DSP* has recently been associated with IPF, both *CD109* and *DSP* were selected as candidate genes.

3.4.3 DSP and Surrounding Genes

As filtering of both the high impact variant lists was being conducted, a GWAS study published in April 2013 identified seven new loci associated with IPF (Fingerlin et al., 2013). The GWAS study included 1,616 individuals with IIPs and 4,683 controls, with replication analyses in 876 cases and 1,890 controls. The main findings of the study

confirmed previously existing associations with the genes *TERT*, *TERC* and *MUC5B*, and also identified seven new loci including associations with *FAM13A* at 4q22, *OBFC1* at 10q24, *ATP11A* at 13q34, *DPP9* at 19p13, *DSP* at 6p24 and chromosomal regions 7q22 and 15q14-15. Initially, analysis of the high impact variant list was performed for variants found in and around these five genes and two additional loci. Through analysis of the high impact variant list, the variant c.-1_1insA was identified in the gene *DSP* in seven samples of the 24 sent for WES. The *DSP* gene encodes desmoplakin, a major protein component of desmosomes which are involved in cell-cell adhesion. Previous studies have shown *DSP* to be associated with ARVCD (Gerull et al., 2004), a fatal cardiac condition. Several NL families have been identified with ARVCD due to a founder mutation in the *TMEM43* gene (Merner et al., 2008). Additionally, DNA demethylation of *DSP* has been shown to decrease expression of *DSP* in eight of eleven lung cancer cell lines (Yang et al., 2012). Because of the increased expression of *DSP* in lung tissue, the suggested association of *DSP* with IPF based on GWAS data, and the fact that *DSP* is situated in a chromosomal location (6p24) previously identified as having suggested linkage with IPF (in family R0851) based on a genome wide scan using SNP markers in the Woods lab (Kamel, 2010), the variant found in *DSP* was further investigated.

The variant of interest, c.-1insA, was found using WES at the start site of *DSP* in seven individuals: all family members of R1136 (z543, z544 and z1396, Figure 11), two of three members from R1487 (z900 and z902), two of three members from R0896 (z730 and z618), and one of one member of R1941 (z1470) (see Table 3 and Appendix G. As

the mutation was found in the three family members of R1136, all affected family members of R1136 were Sanger sequenced (z543, z544, z1396, z545 and z590), as well as affected family members of R1487 (z900, z902, z1373). At the time of sequencing, the variant was not reported in either dbSNP or NHLBI Exome Variant Server databases, and therefore remained a strong candidate. Sanger sequencing confirmed the variant in the three family members of R1136 who were sequenced using WES; however, it was not found in the additional two affected members. Unexpectedly, four months after sequencing was completed for the *DSP* variant in both families, information pertaining to the variant's frequency was made public on NHLBI Exome Variant Server, with a MAF of 0.19, and was predicted to be "probably non-pathogenic". Because the variant did not segregate in R1136 and it has been found to have a high MAF, it was eliminated from the lists of potential candidates.

As the high impact variant list did not demonstrate any additional variants in genes newly associated with IPF from the recently published GWAS study, I investigated variants in the IPF associated genes, as well as variants in genes in close proximity to *DSP* using NextGENe. As *DSP* is located in a loci of suggested linkage in family R0851 (Figure 7; Kamel, 2010), there is a possibility that there may be a causal IPF variant in linkage disequilibrium of *DSP*. Using the raw exome data filtered through NextGENe, I looked for variants in genes previously associated with IPF (*SFTPA1, SFTPA2, SFTPC, TERT, TERC*), genes newly associated with IPF *(DSP, FAM13A, OBFC1, ATP11A, DPP9, MUC2 and TOLLIP*) and genes located at 6p24 in close proximity to *DSP* (*SSR1, CAGE1, RIOK1, SNRNP48* and *BMP6*). Filtering of the raw data was broad and included

looking for missense, splice site and nonsense variants found in any sample, as well as both reported and unreported variants. Upon filtering, there were no variants found to segregate in either R1136 or R0942 that had a MAF <0.05; therefore, none of the variants found in the above listed genes were further investigated (Appendix F). Although there were no variants of interest found, this region should not be eliminated entirely. Linkage analysis has suggested this locus, 6p24, is a region of suggested linkage with IPF (Kamel, 2010). However, the genetic contribution of IPF in R0851 is complex and only partially explained by a *TERT* variant segregating in the nuclear family. Based on these results and the current GWAS study identifying this locus, there may be pathogenic variants in downstream genes or regulatory regions that may be in linkage disequilibrium and should be further investigated.

3.5 *CD109* Variant in R1136

Upon filtering of the high impact variant list in R1136, a nonsense variant in the gene *CD109* was identified in all members of R1136 sequenced by WES (z543, z544 and z1396) (Figure 17). This nonsense variant, c.1474C>T; p.R492X, was found in exon 13 of *CD109*, a gene consisting of 33 exons and shown to be a negative regulator of TGF-β. Furthermore, this variant was not previously reported in 1000 Genomes, dbSNP or NHLBI Exome Variant Server, and was shown to have a highly conserved GERP score of 4.95 (NHLBI Exome Variant Server, 2014). Because of the high impact nature of this nonsense variant, segregation of the variant in affected family members sent for WES and its functional role in TGF-β regulation, *CD109* was selected as a potential candidate gene and the c.1474C>T variant was further investigated.

3.5.1 Sanger Sequencing of CD109 variant in R1136

Sanger sequencing of all five affected and 19 unaffected family members of R1136 was conducted to determine segregation of the variant. As shown in Figures 17 and 18, the variant was confirmed in the three DNA samples from R1136 sent for WES. Additionally, the variant was found in 5/19 other unaffected family members. However, it was not found in the two additional affected members, z545 or z590, as shown in Figure 17.Of the clinically unaffected family members, the variant was found in the sister of z1396 who was diagnosed with chronic obstructive pulmonary disease (COPD) (See Appendix H). Granted the sister of z1396 does not fit the clinical diagnostic criteria for IPF, as outlined in Table 1, we know that pulmonary conditions, specifically IPF, can manifest itself in different symptoms, due to variable expressivity of the condition. Although the variant did not segregate completely with disease phenotype, it was not eliminated for multiple reasons. First, the variant itself was not previously reported in either the 1000 Genomes database, dbSNP or NHLBI Exome Variant Server. As there are only a select few genes associated with IPF, it is believed that the causative variants exist in previously unassociated genes and the variants themselves are either extremely rare or not previously reported. Secondly, the variant is a predicted pathogenic, nonsense variant that falls in the middle of an exon in the gene *CD109,* which can be linked to IPF through its regulation of TGF-β. Lastly, although the variant is not shown to segregate completely with the disease phenotype, it is predicted that the development of IPF is both clinically and genetically heterogeneous, therefore there may be multiple variants that may explain disease development in this family. It is quite possible there may be unaffected family

members who have yet to develop IPF due to the late onset of the disease. Furthermore, IPF is a reduced penetrance disease; therefore, certain family members may not develop IPF even if they have causative IPF variants. For these reasons, the *CD109* variant was further investigated through the Sanger sequencing of NFCCR controls.

3.5.2 Sanger Sequencing of Newfoundland Control Samples

Sanger sequencing of 203 NFCCR control samples was conducted to investigate the prevalence of the c.1474C> T nonsense variant in *CD109* in the Newfoundland population. The variant was not found in any of the 203 control samples sequenced, suggesting this unreported variant is also rare in the Newfoundland population. Because this variant has not been previously reported and did not appear in any control samples, further exploration of this variant was warranted.

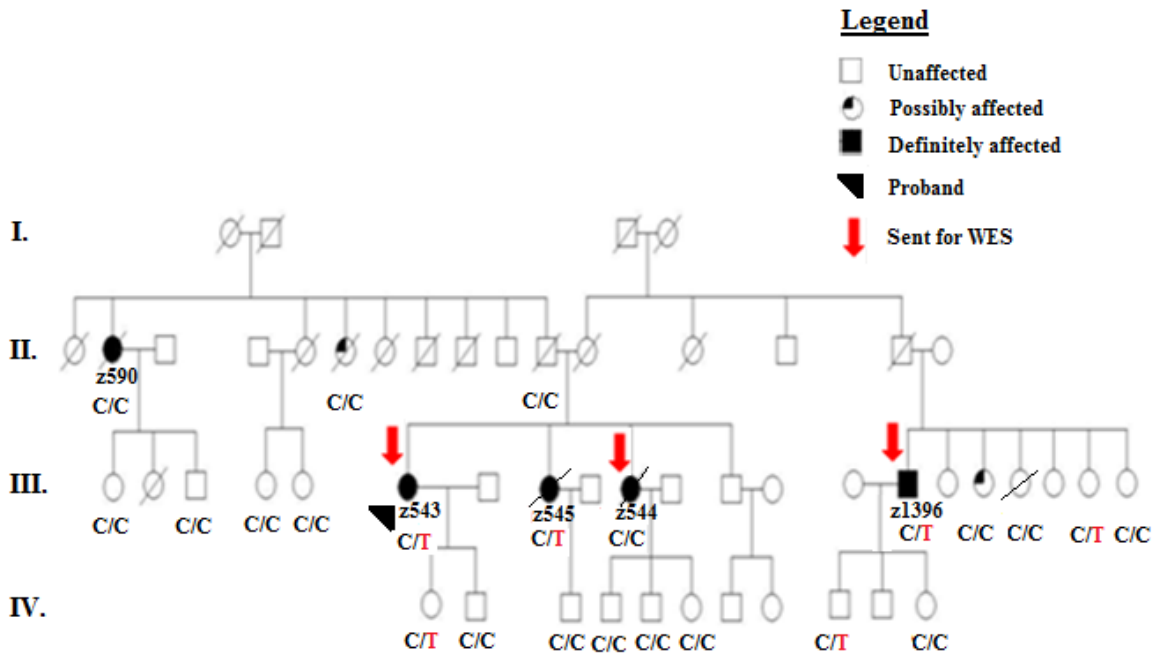
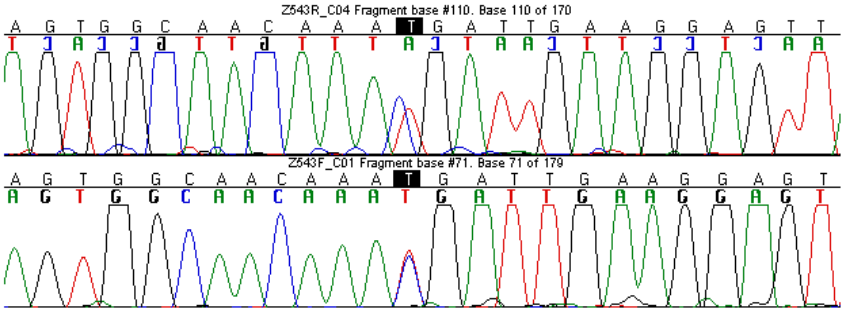


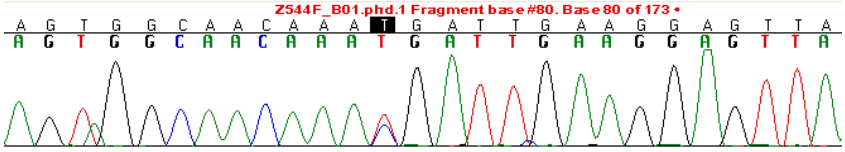
Figure 17: Sanger Sequencing Results for *CD109* Nonsense Variant in R1136

Sanger sequencing results of all members of R1136, for whom DNA samples were available revealed a heterozygous nonsense variant, c.1474C>T; p.R492X, in three of the five affected family members diagnosed with IPF. The variant was also found in five unaffected family members.

A)



B)



C)

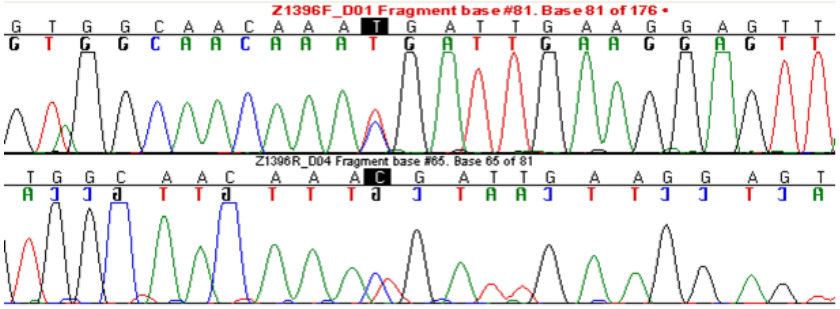


Figure 18: Sanger Sequencing of Affected Individuals in R1136

Sanger sequencing of three affected family members from R1136 who have WES by Genome Québec confirming the c.1474C>T nonsense variant in *CD109*, a negative regulator of TGF-β.

3.5.3 Additional CD109 Variants in Familial Pulmonary Fibrosis Samples

Investigation into additional variants in *CD109* was conducted in two ways. First, a list of variants found in *CD109* was compiled and analysed from the moderate impact variant list, as shown in Table 8. A total of eight moderate impact missense mutations were found in the 24 exomes analysed by WES. Of these eight, four were initially eliminated based on a MAF greater than 0.05 (c.2108A>C, c.2390A>G, c.2533G>A and c.3722C>T). Secondly, Sanger sequencing of selected exons was performed for 54 FPF samples in which there was adequate DNA aliquots available. Exons that were sequenced were chosen based on whether a complete or partial exon fell within a functional protein domain of *CD109*, as shown in Figure 19. Domains are conserved structural sequences of proteins that can act independently of the protein and provide a specific function to the protein; therefore, mutations that occur in genes encoding specific protein domains may have a larger impact than genetic variants translated to other protein locations.

Table 8: Eight Moderate Impact Variants Uncovered in *CD109* Gene in DNA Samples from 24 Familial Pulmonary Fibrosis Patients Analysed by Whole Exome Sequencing

Variant	Exon	Amino Acid Change	Number of Sample Positives (X/24)	MAF	GERP	Poly-Phen2	SIFT	Segregation in Family
c.314G>A	4	R105H	1	0.004	-2.26	0.011 (B)	0.13	No
c.2108A>C	19	Y703S	19	0.44	5.16	0.00 (B)	0.62	Yes
c.2365A>G	21	K789E	1	0.004	-5.58	0.002 (B)	0.89	No
c.2390A>G	21	N797S	11	0.39	1.63	0.107 (B)	0.28	Yes
c.2533G>A	21	V854I	11	0.39	-9.1	0.015 (B)	0.6	Yes
c.3025G>A	25	V1009M	2	0.01	4.62	1.0 (D)	0.06	No
c.3671C>T	29	T1224M	18	0.40	-1.95	0.81 (P)	0.13	Yes
c.3877A>G	31	R1310G	2	0.004	2.15	0.037 (B)	0.35	Yes

The above variants were filtered by Genome Québec, as described in sections 2.2.3 – 2.2.5. Variants were analysed based on MAF and GERP scores, with scores greater than 3.00 indicative of conservation. PolyPhen2 scores were also analysed, with scores less than 0.5 considered benign (B) and scores greater than 0.8 deemed probably damaging (P) or damaging (D), as well as SIFT scores, with scores over 0.05 considered benign (B). Segregation of variants within multiple family members was also analysed. If variants were shown to be most likely benign, non-conserved or have a MAF greater than 0.05, they were eliminated from the study.

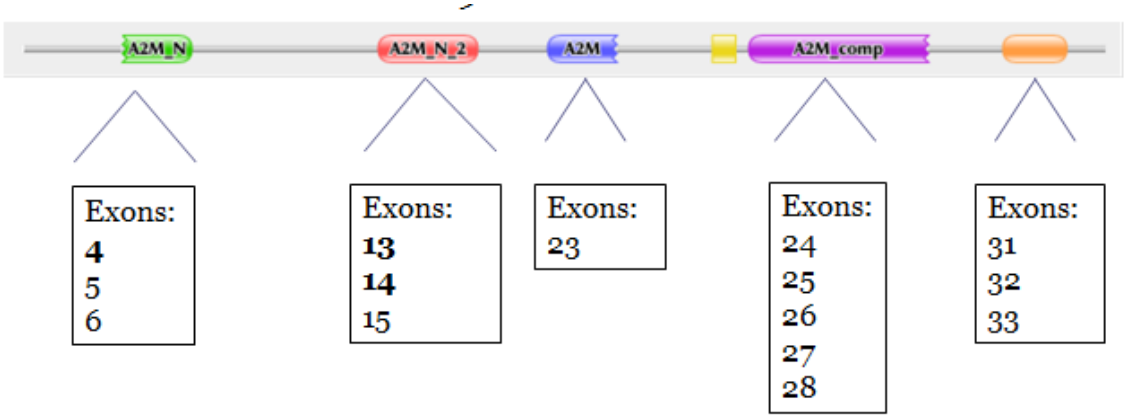


Figure 19: Exons Sequenced by Sanger Sequencing in Functional Domains of *CD109*. Reproduced with permission from Pfam Protein Families Database (2014).

A schematic representation of the functional domains contained in the CD109 protein. The CD109 protein contains six functional domains, with each domain comprised of various exons, as shown above.

3.5.4 Sanger Sequencing of CD109 Gene

Sanger sequencing of all exons found within functional domains in both the forward and reverse directions confirmed the presence of variants previously reported using WES (Table 8), and revealed additional missense variants, represented in Table 9. Further investigation of additional variants using Ensembl, dbSNP and NHLBI Exome Variant Server found these variants to be synonymous mutations, with a MAF greater than 0.05, as shown in Table 9. These variants were further eliminated as being potentially pathogenic based on a high MAF and predicted low impact nature of the synonymous variants.

Of interest were variants c.314G>A, which was found in two sisters (z544 and z545) of R1136 (Figure 20), c.1474C>T, found in three members (z543, z544 and z1396; Figure 17) of R1136 and c.3877A>G which was found in both an affected mother (z660) and daughter (z659) of R1351 (Appendix G). These three variants all had relatively low MAF or were unreported, and were predicted to be pathogenic by either SIFT or Polyphen2 (Table 9).

Table 9: Eight Missense Variants Uncovered in *CD109* Gene in DNA Samples from 54 Familial Pulmonary Fibrosis Patients Analysed by Sanger Sequencing

Variant	Exon	Amino Acid Change	Number of Sample Positives (X/24)	GERP	PolyPhen2	SIFT	MAF
c.314G>A	4	R105H	2	-2.26	0.011 (B)	0.13	0.004
c.645C>T	6	Y215Y	24	1.01	N/A	N/A	0.33
c.1474C>T	13	R492X	3	3.14	N/A	0.70	N/A
c.2108A>C	19	Y703S	19	5.16	0.00 (B)	0.62	0.44
c.2365A>G	21	K789E	1	-5.58	0.002 (B)	0.89	0.004
c.2390A>G	21	N797S	21	1.63	0.107 (B)	0.28	0.39
c.2533G>A	21	V854I	21	-9.1	0.015 (B)	0.6	0.39
c.3025G>A	25	V1009M	4	4.62	1.0 (D)	0.06	0.01
c.3671C>T	29	T1224M	18	-1.95	0.81 (P)	0.13	0.40
c.3877A>G	31	R1310G	2	2.15	0.037 (B)	0.35	0.004
c.4122T>G	33	A374A	20	-0.80	N/A	N/A	0.88

Sanger sequencing of 54 affected FPF samples confirmed the presence of multiple variants in *CD109* reported in the moderate impact list supplied by Genome Québec

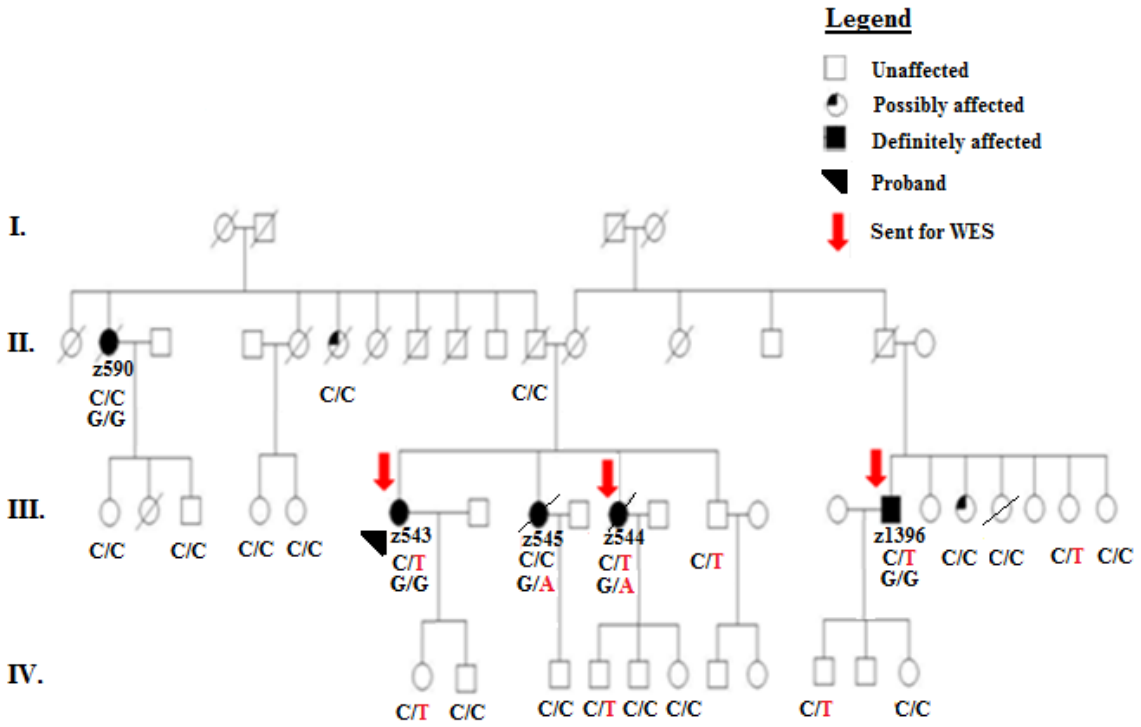


Figure 20: Segregation of *CD109* Variants in R1136

Two variants found in *CD109* were carried by family members of R1136. A c.1474C>T nonsense mutation (top) was found in three affected family members (z543, z544 and z1396), as well as a c.314G>A missense variant (bottom) found in two affected family members (z544 and z545).

3.6 Filtering of Moderate Impact List in R1136

To further investigate all variants in R1136, filtering and analysis of a moderate impact list was conducted. Primarily, variants found in genes previously associated with IPF or suggested to be associated with IPF were analyzed, as shown in Table 6. Of the variants listed, a missense variant in *TEP1* was identified and further pursued in members of R1136. *TEP1* encodes the protein Telomerase Associated Protein-1, and functions as a component of the telomerase ribonucleoprotein complex. The variant found in R1136 was a c.4156C>T, resulting in the amino acid change of an arginine to a tryptophan (p.R1386W), with a GERP conservation score of 4.54 (NHLBI Variant Server, 2014). It is predicted by PolyPhen2 to be "probably damaging", based on a score of 0.999, and is shown to also be a "moderately radical amino acid change", based on a Grantham score of 101. The variant has a MAF less than 0.01 and was further pursued based on the predicted damaging nature and low frequency of the variant. All five affected and 19 unaffected family members of R1136 were tested for the missense variant. As shown in Figure 21, which summarizes the genotype from family R1136, Sanger sequencing confirmed the variant in the two samples sent for WES, z543 and z1396. It was not found in z544, the third sample sent for WES, or z590. It was, however, found in z545, another affected family member and sibling of z543 and z544. Additionally, the variant was found in five unaffected family members. Because of the incomplete segregation of this variant within the family, a small set of NFCCR control samples were sequenced. Upon sequencing of 30 DNA samples, one control tested positive for the variant. Because this variant was found in a control sample, as well as the fact that the variant did not segregate

perfectly with the disease phenotype, this variant was eliminated from the possible list of causative IPF variants.

From the analysis of Table 6, an additional variant in *TEP1* was identified in another family, R1351. This variant, c.1103A>G; p.K368R, was found in both family members of R1351 and was predicted to be damaging by PolyPhen2. However, it was shown to have a MAF of 0.05 and was therefore not a priority for current investigation. This variant should be pursued in future IPF studies conducted in the lab to determine whether there is complete segregation of the variant within the family, as well as its presence in control populations.

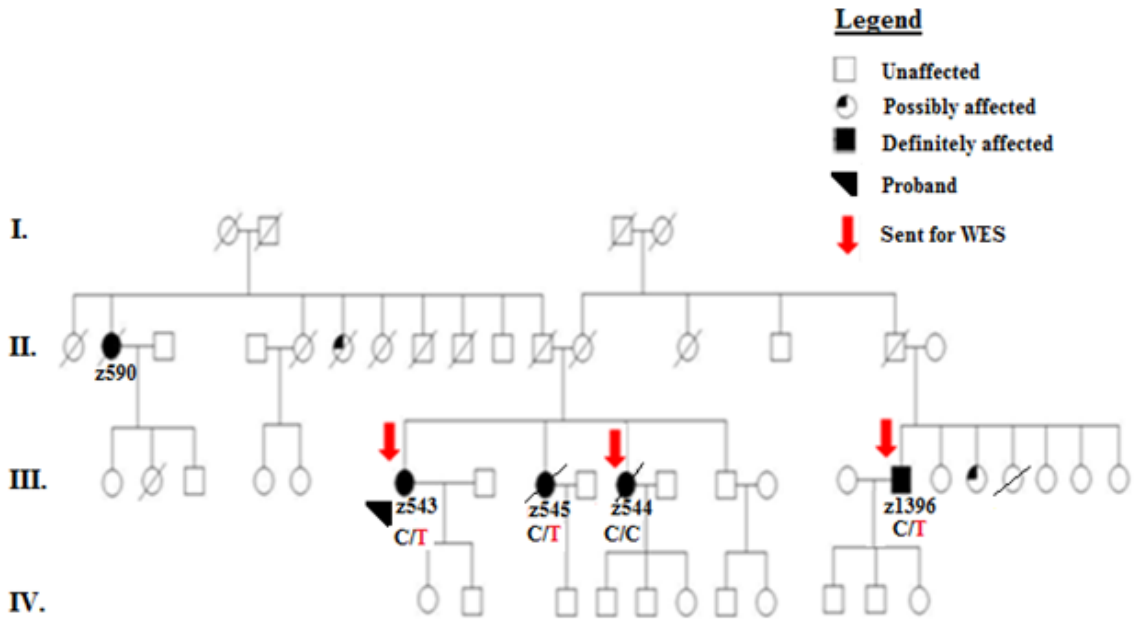


Figure 21: Segregation of *TEP1* Variant in R1136

WES identified a missense variant, c.4156C>T, in two family members of R1136. Sanger sequencing of all affected and unaffected family members confirmed the variant in both individuals sent for WES (z543 and z1396), as well as an additional affected family member (z545) and three unaffected family members. Sanger sequencing of 30 Newfoundland population controls confirmed the variant in one control.

4.0 Discussion

As more research unveils the complexities of ILD development, it is becoming apparent that IPF is heterogeneous both clinically and developmentally. It is evident from the numerous Newfoundland and global families that IPF has a strong genetic component, as many families demonstrate an apparent autosomal dominant mode of inheritance with multiple first degree relatives affected (van Moorsel et al., 2010). However, the etiology behind IPF progression in many of these families remains unknown, with only a small handful of genetic variants confirmed to be causal in the development of the disease globally. Although many IPF families appear to follow a Mendelian autosomal mode of inheritance, it is suggested by the recent ATS/ERS update concerning IPF classifications (ATS/ERS, 2013), multiple studies (Xaubet et al, 2003; Talbert et al, 2005), as well as through the analysis of WES data in the current study, that the development of IPF may indeed be polygenic.

The aim of this research project and thesis was to determine novel genetic variants in two FPF families, R0942 and R1136, sent for WES. Functional gene and variant annotation were two components that were taken into consideration when filtering variants. For example, a c.508_509insG variant was discovered to be found in multiple affected individuals in a gene *IL32*. This gene encodes a pro-inflammatory cytokine, which is important in the activation of pro-inflammatory cytokine genes *IL-1β* and *TNF-α*. Additionally, *IL32* is located in region of suggested linkage, 16p13.3, which was identified by Mr. Fady Kamel (Kamel, 2010), thus making this variant a potential

candidate. However, further Sanger sequencing of control population samples revealed this variant to be present in 8 of 28 controls and therefore, the variant was eliminated. Extensive initial analysis of the WES data did not reveal any obvious causative variants, as many of the potential pathogenic variants did not segregate with all affected family members. As many families are predicted to have polygenic inheritance for IPF development, this study took a broader approach and looked for variants segregating incompletely in affected family members. This was demonstrated in multiple variants of interest in telomerase associated genes, including variants in *TERF1* and *CD109*.

4.1 Implications of Variants in Telomerase Genes

Mutations in *TERC* and *TERT* account for the highest percentage of the known genetic contribution to IPF (Armanios et al., 2007), accounting for a combined total of approximately 10%. Of the 25 NL families reported by Fernandez et al. in 2012, three are segregating *TERT* mutations. Interestingly, almost 80% of patients suffering from IPF are also shown to have shortened telomeres (Steele et al., 2013), which is demonstrated in multiple IPF families in Newfoundland, including R1136 (Figure 8). Therefore, it is plausible that there may exist previously unreported mutations within telomerase associated genes that may alter the structure and function of the telomerase complex, exacerbating the development of IPF.

4.1.1 Telomerase Complex

As previously mentioned, the telomerase complex is responsible for adding short nucleotide repeats, called telomeres, at the ends of replicating chromosomes. They

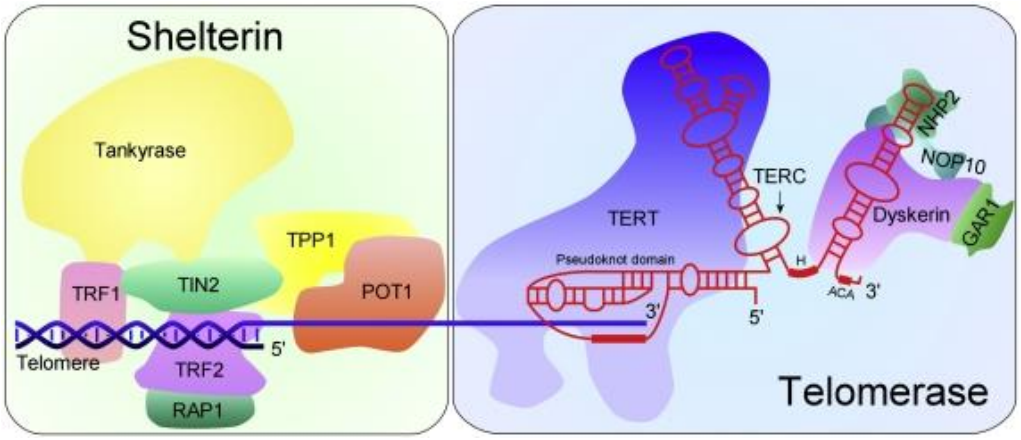
function in protecting the loss of genetic information during each round of replication. As organisms age, telomeres become shorter until cell replication ceases. Telomere shortening has therefore been associated in aging, as well as specific diseases such as cancer (Djojosubroto et al., 2003), Cri-du-chat syndrome (Zhang et al., 2003) and IPF (Liu et al., 2013).

As shown in Figure 22, there are many protein components involved in the telomerase complex, including TERF1 (TRF1) and TEP1. TERF1 acts specifically as a component in the shelterin nucleoprotein complex, which is associated with the telomerase complex throughout every cell cycle. TERF1 acts as an inhibitor of telomerase, limiting the elongation of telomeres. Conversely, the protein TEP1 acts to increase processivity, aiding in the addition of telomeres to the 3' end of replicating DNA. Altogether, the telomerase complex is comprised of many proteins that all function to maintain cellular replication. Mutations in genes encoding any protein component may therefore have a detrimental effect on telomerase efficiency. Figure 22 demonstrates the complexity involved in the telomerase complex.

The c.311G>T variant in *TERF1* was shown to segregate in four individuals within the nuclear family of R0942 (Figure 16). Although *TERF1* is predicted to inhibit telomerase activity, the functional effect of this variant is unknown and may have a hypermorphic effect on gene function, also known as a gain of function mutation. If specific *TERF1* variants result in hypermorphic effects on protein function, one would predict to see increased inhibition of the telomerase complex and therefore shortened

telomerase. Likewise, the genetic effect of the *TEP1* variant found in R1136 may result in diminished protein function, resulting in reduced telomerase activity and shortened telomeres. There is a possibility that these genetic factors may result in changes to the ECM of the lung which may lead to the progression of IPF. However, until functional work is conducted to determine the phenotypic effect of variants found within telomerase associated genes, such as *TERF1* and *TEP1*, one can only speculate on the consequences of these variants and they cannot be ruled out regardless of incomplete segregation. Although the c.311G>T variant in *TERF1* variant was predicted to be damaging and can be linked to the IPF pathways, the individuals in R0942 were not shown to have decreased telomere lengths (Figure 8B). Further investigation into the biological function of *TERF1* and its role in telomere maintenance is needed to better understand the potential relationship between *TERF1* variants and IPF.

A)



B)

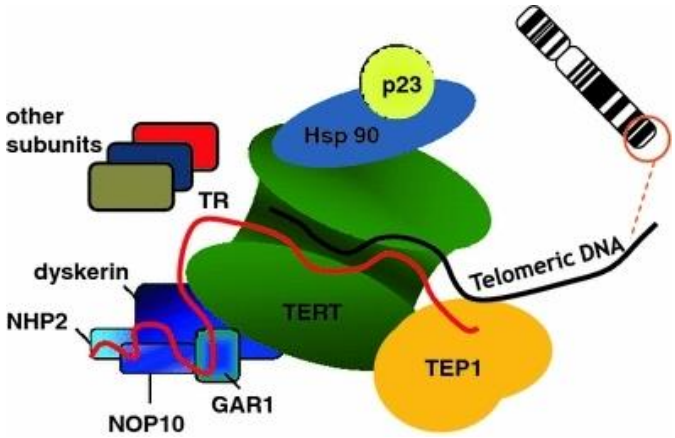


Figure 22: Proteins Involved in the Telomerase Complex Including (A) TERF1 (TRF1) and (B) TEP1. Reproduced with permission by (A) Kirwan et al., 2009. Copyright Biochimica and Biophysica Acta (B) Wojtyla et al., 2011. Copyright Biology Reports.

The telomerase complex involves many proteins, including TERF1 (A) and TEP1 (B) that come together to form a complex which regulates the addition of telomeres on the ends of chromosomes. Mutations in telomerase genes account for various cancers, aging syndromes and lung disorders including IPF and DKC.

As previously mentioned, IPF is thought to be a genetically and developmentally heterogeneous disease. A subset of families are believed to carry multiple mutations, as is already seen in R0851, in which a *TERT* variant is found segregating in a nuclear family but not in the remaining affected family members (Figure 7). This could be similar for R0942, where the c.311G>T variant in *TERF1* may exasperate IPF symptoms due to a malfunctioning telomerase complex. Additionally, both R0942 and R1136 were reported as having a common promoter polymorphism in *MUC5B*, rs35795950. Although this variant is shown to be common within the global population, it has been previously associated with IPF in multiple cohorts (Mathai et al., 2014). Perhaps it is the combination of the *MUC5B* promoter polymorphism as well as the variants found in *TERF1* and *CD109* that are causal in the development of IPF in both R0942 (Figure 23) and R1136 (Figure 24), respectively. Further testing of other cohorts for these variants, as well as functional work in either a mouse model or human cell line may shed light onto the phenotypic consequences of these variants.

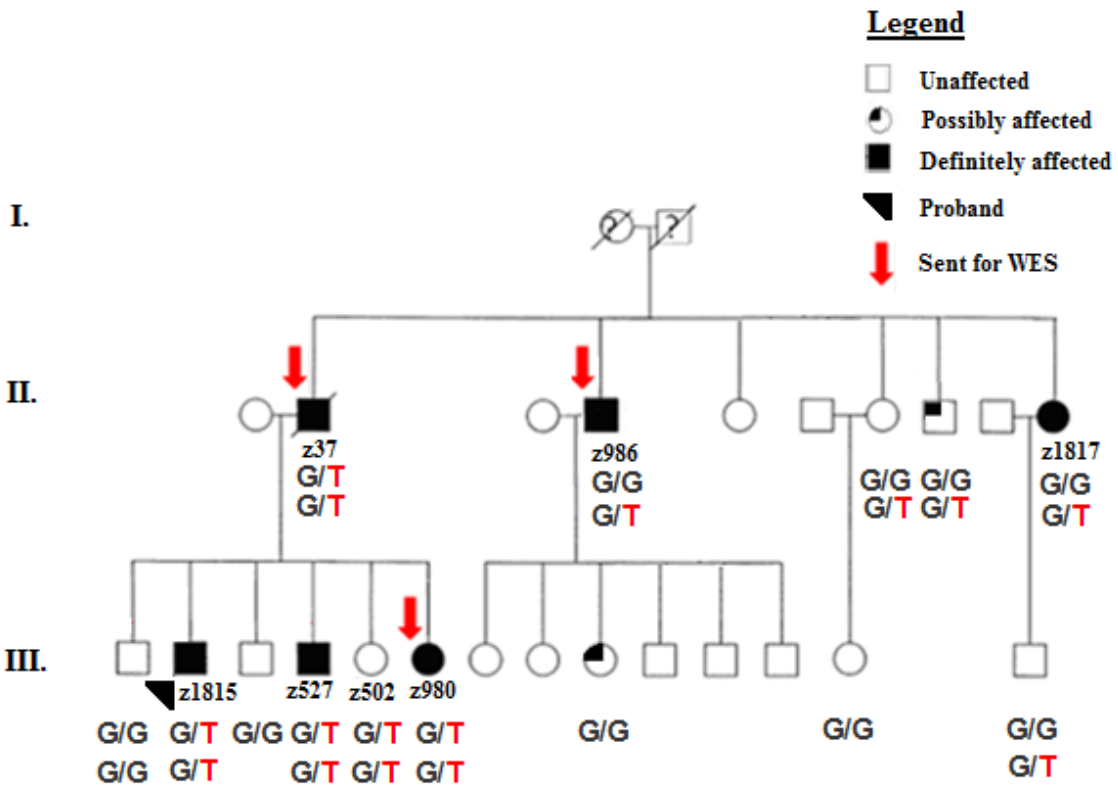


Figure 23: Segregation of *MUC5B* rs35795950 promoter variant and a c.311G>T *TERF1* variant in Family R0942

A rare c.311G>T missense variant (top row) in *TERF1* was uncovered using WES and shown to segregate in four affected and one unaffected member of a nuclear family of R0942. A previously reported promoter variant, *rs35795950* (bottom row), in *MUC5B* was shown to segregate in all six affected and four unaffected members of family R0942. It is believed both variants may play a role in IPF development in this family.

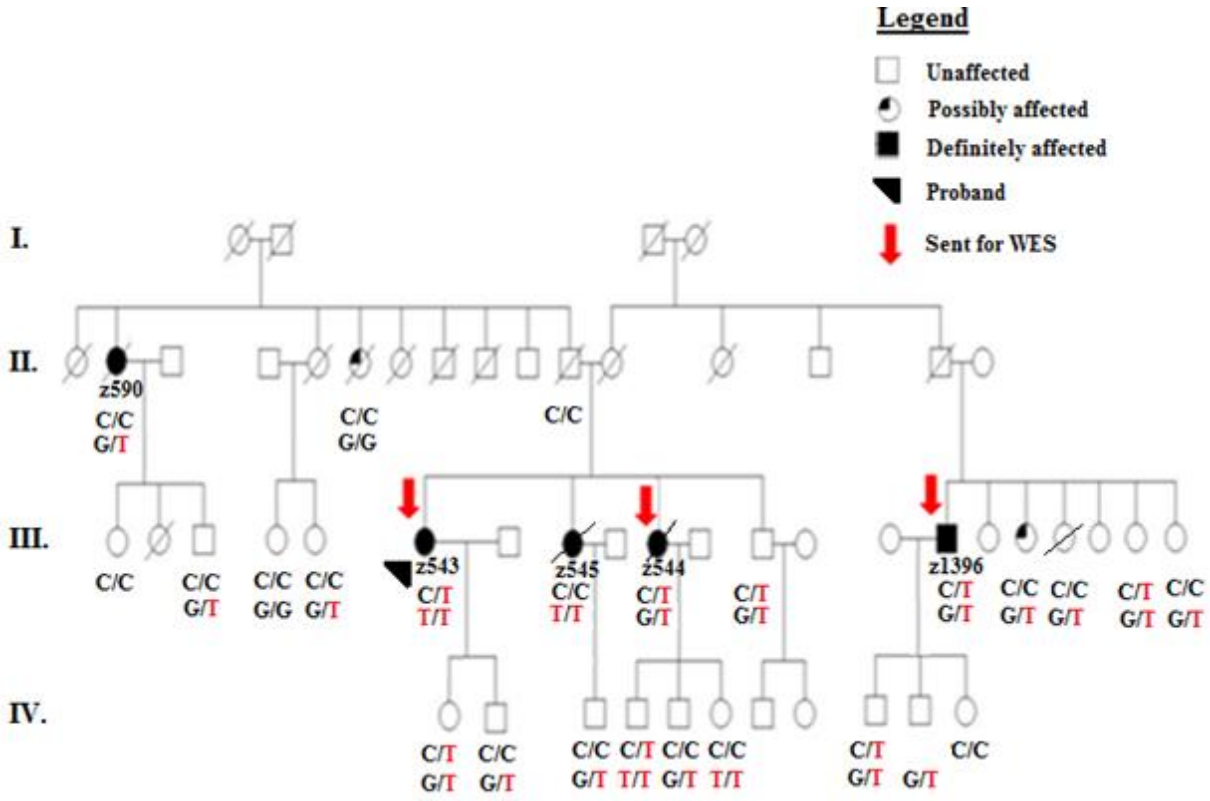


Figure 24: Segregation of *MUC5B rs35795950* promoter variant and a c.1474C>T *CD109* variant in Family R1136

A rare c.1474C>T nonsense variant (top row) in *CD109* was uncovered using WES and shown to segregate in three affected and five unaffected members of family R1136. A previously reported promoter variant, *rs35795950* (G>T; bottom row), in *MUC5B* was shown to segregate in all five affected and 15 unaffected members of R1136 in both the homozygous and heterozygous genotype. It is believed both variants may play a role in IPF development in this family.

4.2 *Genetic Variants found in CD109*

Along with the c.311G>T variant found in *TERF1* in R0942 and R1136, respectively, a previously unreported nonsense variant, c.1474C>T in the gene *CD109* was uncovered in family R1136. This variant was shown not to segregate entirely with all affected family members; however, from analysis of the family pedigree presented in Figure 17, it is suggestive that there may be at least two causes of IPF in this particular family. The proband (z543) was diagnosed with IPF at age 60, while her two sisters (z544 and z545) were diagnosed only a few years apart from each other at the ages of 61 and 64, respectively. Although the proband is still alive, both sisters died from IPF, with z544 and z545 dying five and nine years post-diagnosis, respectively. Additionally, a first cousin of this sibship on the maternal family line was diagnosed at age 53 and is currently in stable condition, as well as an aunt on the paternal side who was diagnosed at the age of 92 and died two years later. Diagnosis of IPF on both the maternal and paternal side of the family suggests there may be more than one genetic contributor in this family. Although all three sisters were diagnosed around the same age, z545 (sister of the proband) was also diagnosed with lung cancer and emphysema and was a heavy smoker for 52 years. She was exposed to printers ink while working in a print shop for many years of her life, thereby exposing her to multiple environmental assaults that may have added to her underlying lung conditions. As shown in Table 9, z545 was shown to have an additional variant in *CD109*, a c.314G>A in exon 4. Although this variant, which is shared by z544 (sister of the proband), is not shown to be evolutionarily conserved according to GERP, it has a predicted low MAF of 0.004. Although all siblings presented

around the same age with symptoms of IPF, there is the possibility that there may be a partial environmental cause of IPF for z545. There are therefore, multiple contributing factors adding to the lung conditions seen in z545, which is a reason that the *CD109* nonsense variant was not excluded in this family even with the non-segregation seen in z545. As this family is not known to have any consanguineous relationships, it is suggestive that there may be at least two causes of IPF running through this family. This is also suggested given the segregation of FPF from both the maternal and paternal side of family R1136 (see z590 and z1396 from Figure 17). This is an important concept to keep in mind when analyzing WES data for reduced penetrant, autosomal dominant conditions, as variants not completely segregating with all affected family members should not necessarily be eliminated. Incomplete segregation of a pathogenic variant has already been seen in family R0851, as described in Figure 7. There is a possibility that the variants in *TERF1* and *CD109* may also be pathogenic, yet incompletely segregate with IPF status.

4.2.1 Relationship between CD109 and TGF-β

CD109, encoded by the *CD109* gene, is a glycosylphosphatidylinositol (GPI)-linked glycoprotein. This 170 kilodalton protein functions as a cell surface antigen, with specific binding sites for TGF-β, basic fibroblast growth factor, IL-1β, and platelet derived growth factor (Bizet et al., 2011). As previously mentioned, TGF-β has long been implicated in the development of IPF (Khalil et al., 1991), although the exact mechanisms are still unknown. Overactive signalling of TGF-β is thought to increase the accumulation of the ECM, a symptom seen in many IPF patients who are generally

reported as having exaggerated ECM deposition. As previously described, TGF-β binds to its receptors, TGFβRI and II, resulting in the activation of tyrosine kinase activity and phosphorylation of SMAD signalling molecules. The phosphorylation and activation of SMAD proteins permits the translocation of these molecules into the nucleus where they may act as transcription factor binding proteins for various target genes. However, different signalling mechanisms exist in which TGF-β binds to alternative receptors which affect gene transcription, as is seen in the presence of *CD109*. As depicted in Figure 25, the CD109 protein is shown to bind to TGFβRI and II. Additionally, CD109 associates and binds with calveloin-1, a major protein component of the caveolae. Together, CD109 bound to calavelin-1 and the TGF-β receptor complex results in the internalization of CD109 and TGF-β receptors into a caveolae. This culminates in the localization of TGF-β receptors into the caveolar compartment where degradation of TGF-β receptors occurs. Therefore, in the presence of the CD109 protein, activation of TGF-β signalling is decreased, which may have profound effects on downstream target genes.

4.2.2 TGF-β and Pulmonary Fibrosis

The role of TGF-β involves the activation of various signalling cascades which ultimately affect gene expression and cell growth. TGF-β mediated growth acceleration or inhibition is dependent upon the specific cell type. Fibroblast differentiation into myofibroblasts occurs when TGF-β signalling is increased; therefore, TGF-β is thought to play an important role in the life cycle of fibroblast formation. Increased fibroblast formation is often seen in IPF (Steele et al., 2013) and is a staple to the development of

the disease. Not only is TGF-β thought to increase fibroblast formation and exaggerate ECM deposition, recently it has been shown that TGF-β signalling can affect the expression of a known IPF susceptibility gene, *TERT*. Researchers Li et al. in 2006 first demonstrated an association between TGF-β and *TERT* via SMAD-mediated regulation (Li et al., 2006). Using a human breast cancer cell line, it was shown that TGF-β signalling induces SMAD3 translocation into the nucleus where it binds to the *TERT* gene promoter and suppresses gene transcription (Figure 26). Additionally, knockout experiments resulting in the silencing of *SMAD3* gene expression were shown to eliminate *TERT* suppression, supporting the association between TGF-β and *TERT* (Lacerte et al., 2008), findings similar to that of Li and colleagues in 2006. It is possible that dysregulation of TGF-β signalling may alter *TERT* expression, leading to shortened telomeres and the development of IPF and other fibrotic diseases.

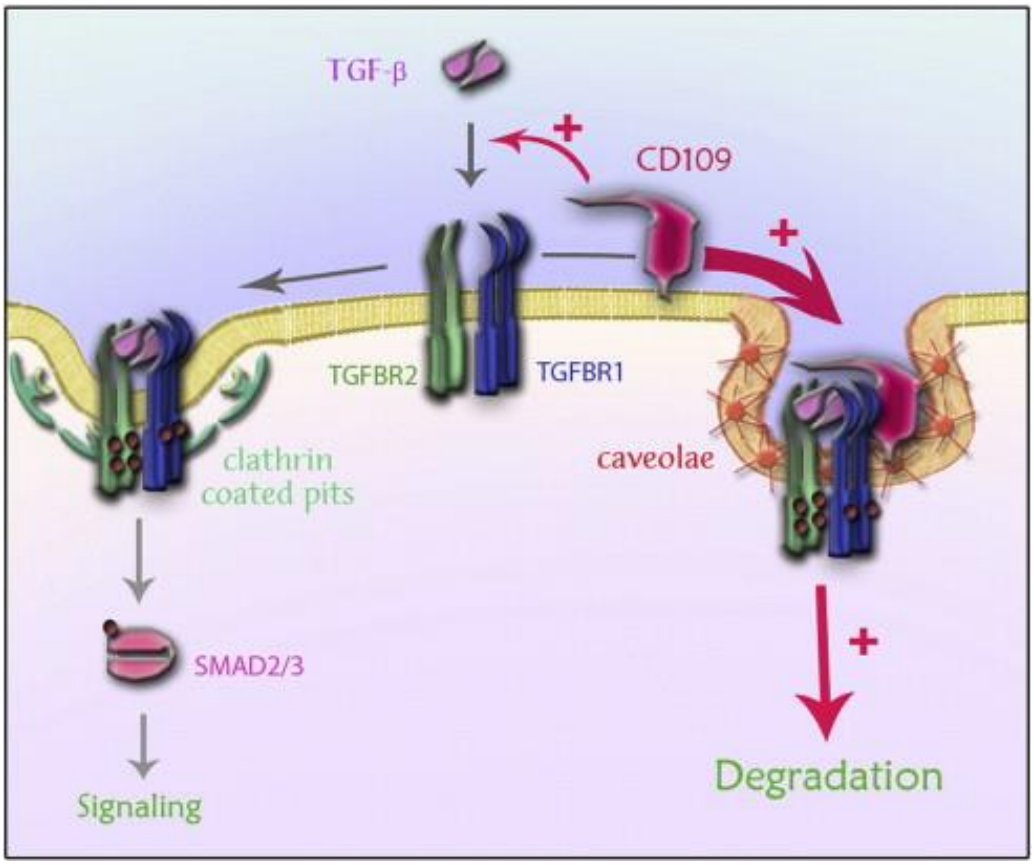


Figure 25: Schematic Model of the Potential Mechanism by Which CD109 May Regulate TGF-β Receptor Internalization and Degradation. Reproduced with permission by Bizet et al., 2011. Copyright Biochimica et Biophysica Acta (BBA) - Molecular Cell Research

TGF-β is shown to bind to its receptors, TGFβRI and II, forming a complex resulting in the endocytosis of its receptors and activation of the tyrosine kinase subunit. This activation allows for the phosphorylation and subsequent activation of SMAD2/3 signalling molecules, which can further regulate gene transcription upon translocation into the nucleus. Alternatively, TGF-β may bind to CD109 antigens at the surface. The proposed mechanism is thought to involve the endocytosis of TGF-β bound to both the TGFβR11/ TGFβRII complex and CD109 antigen. This results in the formation of a caveolae pit, resulting in the degradation of TGF-β and its receptors, therefore reducing TGF-β mediated signalling.

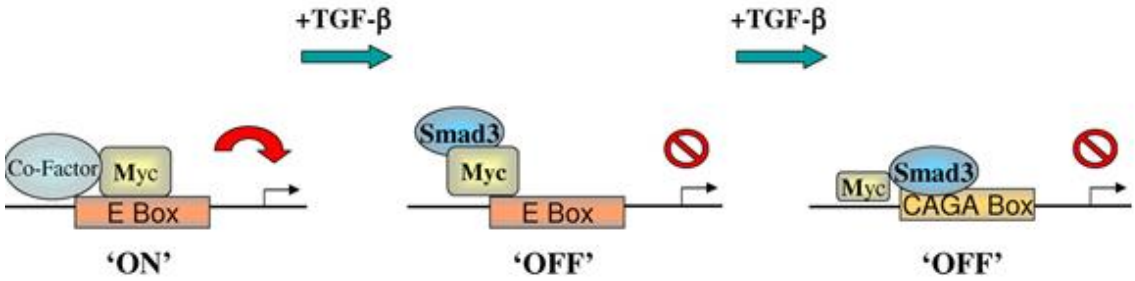


Figure 26: Model of the Mechanisms by Which TGF-β Induces *TERT* Gene Suppression. Reproduced with permission by Li et al., 2006. Copyright The Journal of Biological Chemistry.

A model mechanism by which TGF-β supresses *TERT* gene expression. In the absence of SMAD molecules, Myc can bind to the enhancer box (E box) of the *TERT* gene, resulting in increased gene activation. However, in the presence of SMAD3 molecules via TGF-β signalling, SMAD3 binds and sequesters Myc, preventing binding to the E Box of *TERT* and therefore reduced transcriptional activity. Additionally, the SMAD3 and Myc complex may bind to a downstream CAGA Box, resulting in suppression of *TERT* transcriptional activity.

4.2.3 Association of CD109 in Fibrotic Conditions and Human Disease

Although *CD109* is a relatively novel annotated gene, with much work left in understanding its exact functional consequences, research has demonstrated the importance of *CD109* in many human conditions. Changes in expression patterns of *CD109* are seen in many cancers, including uterine cancer (Semczuk et al., 2013) and lung squamous cell carcinomas (Sato et al., 2007). Interestingly, changes in *CD109* expression have been implicated in many fibrotic conditions. Mouse models have allowed researchers to study such conditions, including IPF and fibrotic skin diseases. For example, transgenic mice overexpressing *CD109* as well as wildtype littermates were injected with bleomycin to induced scleroderma, a fibrotic skin condition resulting in hardening of the epidermis. Transgenic mice were shown to have significantly less skin fibrosis, as measured by dermal thickness, collagen crosslinking and fibronectin content compared to wildtype littermates (Vorstenbosch et al., 2013). Similar results were found in a recent study, where both wildtype and transgenic mice overexpressing *CD109* were subjected to dorsal excisions, creating a novel murine hypoxic wound model (Winocour et al., 2014). Many fibrotic conditions, including scleroderma and IPF, result in hypoxic wound conditions; therefore, this model serves to evaluate wound healing in an environment suitable to the disease phenotype and similar to that of IPF. In another study, isolated hypoxic wound healing tissue was shown to have less fibronectin and collagen type-1 expression in *CD109* transgenic mice, as well as less dermal thickening seven days post wounding compared to wildtype and non-hypoxic control mice (Winocour et al., 2014). These results imply that *CD109* may have a protective effect against fibrotic

wound healing in the presence of hypoxic conditions. As IPF progresses, hypoxia is often seen due to a thickening of the alveolar interstitium which results in a decreased diffusion of oxygen rich arterial blood (Robinson et al., 2012); therefore, this model may serve to represent late stage IPF progression.

CD109 expression patterns have also been shown to be altered in scleroderma fibroblasts isolated from human cell lines. Scleroderma has similar phenotypic findings to IPF, as both diseases are a result of over production of pro-fibrotic proteins leading to the hardening of epidermal cells. Interestingly, pulmonary distress is often seen in specific scleroderma cases, such as diffuse scleroderma. In these cases, one or more of the internal organs is affected, including the lungs (Solomon et al., 2013), with as many as 90% of patients having interstitial lung abnormalities on HRCT scans (Schurawitzki et al., 1990). Patients with systemic scleroderma may even be misdiagnosed as having ILD due to a lack of clinical symptoms in the epidermis of the skin, a condition known as scleroderma sine scleroderma (Toya et al., 2009). Trans-infection of *CD109* specific siRNA in both normal and fibroblast cells isolated from scleroderma patients has been shown to decrease expression of TGF-β signalling, resulting in a marked decrease in fibronectin and collagen accumulation (Man et al., 2012). *CD109* dysregulation has also been implicated in psoriasis, a chronic skin disease characterized by red, itchy patches of skin. Researchers Litvinov and colleagues demonstrated a decrease in CD109 protein expression in the keratinocyte of a patient with psoriasis compared to controls (Litvinov et al., 2011). Although there has yet to be mutations reported in *CD109* that can be

attributed to any heritable human disease, it is suggestive from the literature that *CD109* expression patterns, when altered, may play an important role in fibrotic conditions.

4.2.4 Implications of CD109 Mutations in Relation to Pulmonary Fibrosis

The role of the CD109 protein as a cell surface antigen involves the interaction and regulation of profibrotic proteins and growth factors, such as TGF-β. As TGF-β has been previously implicated in IPF, and has recently been linked to altering *TERT* gene expression, variants in *CD109* warranted further investigation. Two variants in *CD109* were found in multiple family members of R1136: a c.1474C>T nonsense variant was found in two first and one second degree relatives, as well as an additional rare missense variant, c.314G>A, in two first degree relative. As nonsense variants are mainly predicted to be protein truncating mutations, one would expect a decrease in *CD109* gene expression with regards to the c.1474C>T nonsense variant. As *CD109* is a negative regulator of TGF-β, one would expect that decreased protein expression of *CD109* may increase TGF-β expression. This has various implications as described in Figure 27. Increased TGF-β expression may result in increased ECM deposition which may result in the accumulation of profibrotic proteins such as collagen and fibronectin. This exaggerated ECM may simultaneously recruit additional profibrotic proteins, resulting in a positive feedback cascade in which permanent changes to the alveolar interstitium occur (Figure 4). Additionally, as TGF-β is predicted to be a negative regulator of *TERT* expression, an increased expression of TGF-β may result in a decreased expression of *TERT*. These combined factors may increase the risk of developing IPF and/ or exacerbate existing conditions.

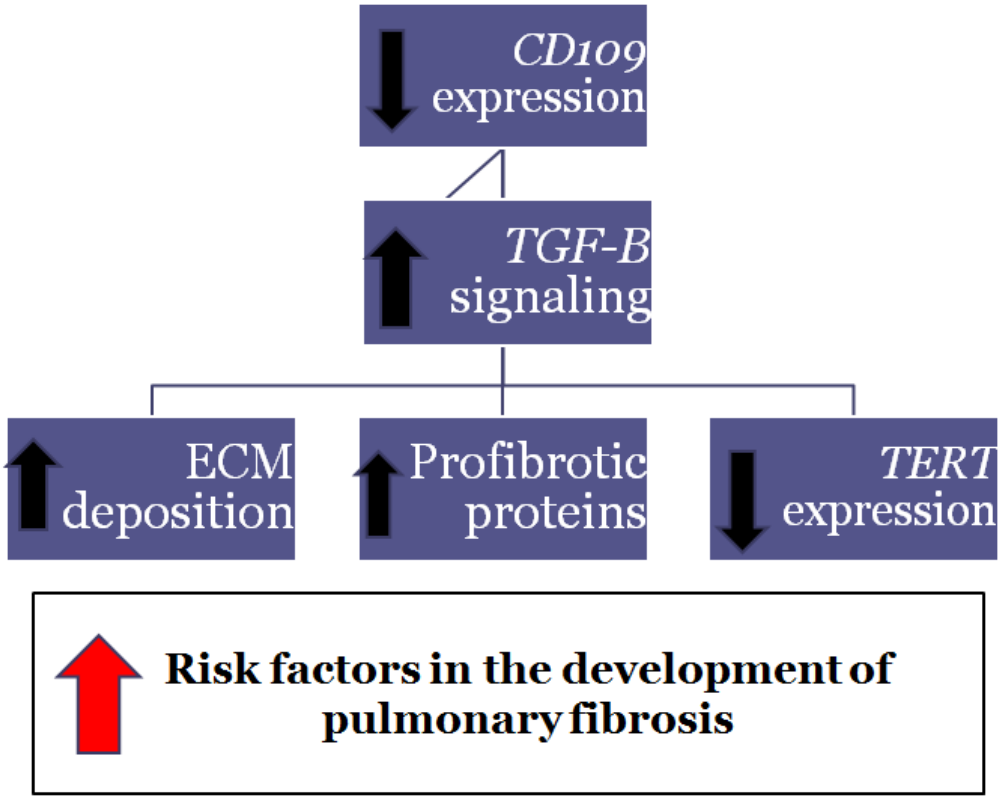


Figure 27: Potential Mechanism by Which *CD109* Mutations May Contribute to the Development of Pulmonary Fibrosis

A novel c.1474C>T variant was found in the gene *CD109*. This nonsense variant may potentially result in decreased *CD109* expression or truncation of the CD109 protein. As *CD109* is a negative regulator of TGF-β, this may result in an increased expression and activity of TGF-β. This in turn may have implications in the development of IPF, as increased TGF-β may result in an increase in ECM deposition and accumulation of profibrotic proteins. Additionally, TGF-β has been shown to be a negative regulator of *TERT*, a known IPF susceptibility gene. Increased expression of TGF-β via a decreased expression in *CD109* may result in decreased expression of *TERT*, resulting in shortened telomeres and the development of IPF.

4.2.5 Importance of Newfoundland Controls

The use of population specific controls is essential for determining the pathogenicity of novel and seemingly rare variants. A variant that may be unreported or rare in public databases, such dbSNP or the 1000 Genomes Database, may in fact have a higher prevalence in specific populations due to population stratification (Matheison et al., 2012), underlying the importance of collecting and integrating control demographics in genetic studies. Certain variants of interest, such as the g.79559667C>T variant in *SFTPA2,* c.508_509insG variant in *IL32* and c.4156C>T variant in *TEP1* were all reported as having low or unreported MAF in dbSNP and the 1000 Genomes Database. However, upon sequencing of NL controls, these variants were all shown to have higher prevalence in the NL population. The use of population specific controls is therefore a useful tool in assessing the prevalence and pathogenicity of variants in specific populations. What may be a rare variant in the general population, or specific databases, may in fact be more common and benign in the NL population due to the NL population architecture. In the future, the development of an in-house control database containing WES data for healthy, NL controls will be beneficial in assessing variants uncovered by WES.

4.3 Limitations of Study

4.3.1 Limitations in Whole Exome Sequencing

The study of genetic conditions has allowed for the discovery of many novel causative variants in Mendelian and familial disease. Although technology has progressed rapidly and has been a boom in most fields within genetics, there remain

numerous limitations in uncovering genetic sources of human disease. The use of WES has uncovered causative variants in many diseases (Bamshad et al., 2013) and has allowed for the amplification and fast sequencing of thousands of DNA sequences in short amounts of time. However, there are drawbacks to this technology. A drawback to WES includes the fact that sequencing of the exome only includes 1-2% of the entire human genome. Although it is estimated that 85% of human diseases can be attributed to the exome (Bolstein et al., 2003), more and more research is shedding light onto the importance of non-coding DNA, as revealed in the ENCODE project (Farnham, 2012). Non-coding DNA regions are shown to play more of an important role than once previously thought, with regulatory and intronic regions demonstrated to be important in altering gene expression. For example, a recent study published in Nature demonstrated the importance on non-coding regulatory regions to human disease (Weedon et al., 2014).

Using whole genome sequencing (WGS), Weedon and colleagues were able to identify pathogenic variants in 10 probands of families diagnosed with pancreatic agenesis, an autosomal recessive condition in which partial or complete development of the pancreas is absent. In total, four variants were identified using WGS: a family containing a homozygous g.23508437A>G missense variant, a homozygous g.23508363A>G mutation, a homozygous g.23508305A>G mutation and compound heterozygous g.23508365A>G and g.23508446A>C mutations. Additionally, one of the families was shown to have a 7.6 kb pair deletion spanning the same loci as the above variants (chromosome 10: 23,502,416–23,510,031). The shared haplotype was shown to span a 7.6KB region, approximately 25 KB downstream from the *PTF1A* gene. This

PTF1A gene encodes pancreas transcription factor 1 subunit alpha and is shown to have functional importance in early pancreatic development (Sellick et al., 2004). These variants were shown to segregate with affected patients within their respective families and the pathogenicity of these variants was further confirmed through functional annotation. In conclusion, the variant was predicted to fall within a regulatory region of the *PTF1A* gene and shown to decrease gene expression, leading to the development of pancreatic agenesis in multiple families. This study demonstrates the importance of regulatory regions in gene transcription and how variants in these regions may have a detrimental effect on gene transcription and as well as human disease. Therefore, there is the potential that using WES compared to WGS may prevent the discovery of causative variants in non-coding regions.

Reduced coverage in specific areas may result in variants being lost during QC measures. Additionally, exome capture may not fully identify and sequence all exons for specific samples due to technical difficulties that can occur during WES. This may have detrimental impacts when analyzing both singleton and familial exome data, as segregation of variants with disease may not be accurate. Similar to this, WES involves many QC measures, including trimming the ends of low quality reads and eliminating low coverage reads, which have the potential to lose important variants if low quality reads are filtered out.

4.3.2 Drawbacks to Study Design

The current study design for this project included the filtering of variants that segregated mainly within all affected family members and were shown to be rare through

a MAF less than 0.05. Furthermore, variants were selected based on gene function, with protein functions relating to the IPF developmental pathway investigated more stringently than other unannotated or unrelated gene candidates. This design has two drawbacks; firstly, WES of 24 exomes revealed many variants in unannotated genes. There is the possibility that these genes may have functional importance in the development of IPF, yet little is known regarding its function. Secondly, filtering of variants was largely based on gene function, with many variants excluded as the gene function did not fit the predicted developmental pathway of IPF. Because the etiological mechanisms underlying the development of IPF is still largely unknown, research into the specific pathways underlying IPF progression will aid in annotating gene function and variants found within genes not previously thought to be associated with IPF development.

Additionally, the use of WES has proven to be challenging in many novel gene discovery studies, especially those concerning autosomal dominant conditions. In a 2014 report published by Finding of Rare Disease Genes in Canada (FORGE) in the American Journal of Human Genetics, the use of WES was examined for the use of novel gene discovery. It was concluded that the use of WES for moderately sized autosomal dominant families is problematic. Many families studied were left with too many heterozygous variants to accurately analyze, whereas others were shown to have few candidate genes (Beaulieu et al., 2014). This observation is also seen in the current study, where many families, including R0942 and R1487 showed a small number of candidate genes within the high impact variant lists; however, the moderate impact variant lists required extensive filtering beyond the scope of what was manageable. The analysis of

NGS sequencing, specifically using WES and WGS, has proven to be challenging; from the analysis of large sets of sequencing data to the storage of the data on adequately sized hard drives. As previously stated, the fact that IPF is a highly clinically and genetically heterogeneous disease, potentially developing from multiple genetic variants not segregating completely within the family proves difficult when studying WES data.

Although the diagnostic criteria for IPF are clearly stated, the ATS/ERS update admits that there is variability with the manifestation of ILD symptoms due to the heterogenic nature of the disease (ATS/ERS, 2013). Because there are many factors involved in the development of IPF including environmental factors, such as cigarette smoking, chemical inhalation, environmental irritants, as well as genetic and potentially autoimmune responses, determining one cause for IPF in each patient may not be realistic. Detection of genetic contributors to autosomal dominant disease proves difficult when these environmental factors come into play and must be taken into account when analyzing WES data.

4.4 Future Work

The continual analysis of the WES data on 24 affected IPF patients will be a focal point in future work conducted with the present project. The current thesis has described variants found in only two families; however, data exists on multiple members from other families. Continuous filtering of the high and moderate impact lists, as well as the raw WES data using NextGENe or equivalent programs, may reveal other interesting variants that may warrant further investigation.

With regards to the specific variants described in this current thesis, further genetic work may be undertaken in the form of collaborations with researchers studying IPF in other cohorts. The identification of novel and putatively pathogenic variants in *CD109* and *TERF1* in additional IPF cohorts would provide more evidence that these genes may be new IPF susceptibility genes. Alternatively, functional work using either a knockout mouse model or human cell lines may give insight into the phenotypic effects of specific variants found in *CD109* and *TERF1*.

4.5 Conclusion

The use of WES and WGS in the identification of novel disease causing variants is one of the latest technologies that have allowed researchers to broaden their knowledge of specific disease etiologies and progressions. Although there exists difficulties with the interpretation of this technology, specifically with autosomal dominant diseases, researchers are gaining a better understanding of how to interpret WES data which will prove useful in future work.

The current study involved the use of WES of DNA samples belonging to 24 affected IPF patients, from 14 different families. The analysis and filtering of this data uncovered multiple potential causative variants in two families, R0942 and R1136, including *TERF1* and *CD109*, respectively. Sanger sequencing of all affected and unaffected family members in this family demonstrated incomplete segregation of both variants; however, neither variant was found in control samples that were tested. The implications of variants in both these genes may result in alterations of the telomerase

complex, which is thought to be integral in the development of IPF. Future work will require the functional annotation of these variants. The discovery of novel variants will not only help researchers better understand the molecular etiologies uncovering the development of IPF, but may one day lead to a more efficient treatment plan and prolongation of life.

References

Adzhubei, I., Jordan, D., & Sunyaev, S. (2013). Predicting Functional Effect of Human Missense Mutations using PolyPhen-2. *Current Protocols in Human Genetics*, 76:7.20.1-7.20.41.

Albers, C., Lunter, G., MacArthur, D., McVean, G., Ouwehand, W., & Durbin, R. (2011). Dindel: accurate indel calls from short-read data. *Genome Research*, 21(6):961-973.

Alder, J., Chen, J., Lancaster, L., Danoff, S., Su, S., Cogan, J., et al. (2008). Short Telomeres are a Risk Factor for Idiopathic Pulmonary Fibrosis. *Proceedings of Natational Academy of Science*, 105(35): 13051-13056.

Araz, O., Yilmazel Ucar, E., Meral, M., Yalcin, A., Acemoglu, H., Dogan, H., et al. (2014). Frequency of Class I and II HLA Alleles in Patients with Lung Cancer According to Chemotherapy Response and Five-Year Survival. *Clinical Respiratory Journal*, Epub.

Armanios, M., Chen, J., Chang, Y., Brodsky, R., Hawkins, A., Griffen, C., et al. (2005). Haploinsufficiency of Telomerase Reverse Transcriptase Leads to Anticipation in Autosomal Dominant Dyskeratosis Congenita. *Proceedings of the National Academy of Sciences*, 102: 15960-15964.

Armanios, M., Chen, J., Cogan, J., Alder, J., Ingersoll, R., Markin, C., et al. (2007). Telomerase Mutations in Families with Idiopathic Pulmonary Fibrosis. *New England Journal of Medicine*, 356: 1317-1326.

ATS/ERS. (2002). American Thoracic Society/ European Respiratory Society International Multidisiplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. *American Journal of Respiratory and Critical Care Medicine*, 277-302.

ATS/ERS. (2013). An Offical American Thoracic Society, European Respiratory Society Statement: Update of the International Multidisciplinary Classification of the Idiopathic Interstitial Pneumonias. *American Journal of Respiratory Critical Care Medicine*, 733-748.

Bamshad, M., Ng, S., Bigham, A., Tabor, H., Emond, M., Nickerson, D., et al. (2013). Exome Sequencing as a Tool for Mendelian Disease Gene Discovery. *Nature Reviews*, 12: 745-755.

Baumgartner, K., Samet, J., Coultas, D., Stidley, C., Hunt, W., T, C., et al. (2000). Occupational and Environmental Risk Factors for Idiopathic Pulmonary Fibrosis: A Multicenter Case-Control Study. *American Journal of Epidemiology*, 152: 307-315.

Baumgartner, K., Samet, J., Stidley, A., Colby, T., & Waldron, J. (1997). Cigarette Smoking: A Risk Factor for Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 155: 242-248.

Beaulieu, C., Majewski, J., Schwartzentruber, J., Samuels, M., Fernandez, B., Bernier, F., et al. (2014). FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project. *The American Journal of Human Genetics*, 94: 809-817.

Bissell, D. (2001). Chronic Liver Injury, TGF-Beta and Cancer. *Experimental and Molecular Medicine*, 33:179-190.

Bizet, A., Liu, K., Tran-Khanh, N., Saksena, A., Vorstenbosch, J., Finnson, K., et al. (2011). The TGF-Beta Co-Receptor, CD109, Promotes Internalization and Degradation of TGF-Beta Receptors. *Biochimica et Biophysica Acta*, 742-753.

Botstein, D., & Risch, N. (2003). Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease 33. *Nature Genetics*, 33: 228–237.

Burrow, M., & Wheeler, D. (1994). A block-sorting loss-less data compression algorithm. *Students Representative Counsel Research*, 124.

Carter, B. (2012). Hermansky-Pudlak Syndrome Complicated by Pulmonary Fibrosis. *Proceedings (Baylor University Medical Center)*, 25(1):76-77.

Choi, M., Scholl, U., Ji, W., Liu, T., Tikhonova, I., Zumbo, P., et al. (2009). Genetic Diagnosis by Whole Exome Capture and Massively Parallel DNA Sequencing. *Proceedings of National Academy of Science USA*, 106(45): 19096-101.

Cooper, G., Stone, E., Asimenos, G., Program, N. C., Green, E., Batzoglou, S., et al. (2005). Distrubtion and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901-913.

Cronkhite, J., Xing, C., Raghi, G., Chin, K., Torres, F., Rosenblatt, R., et al. (2008). Telomere Shortening in Familial and Sporadic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 178(7): 729-737.

Davies, H., Richeldi, L., & Walters, E. (2003). Immunomodulatory Agents for Idiopathic Pulmonary Fibrosis. *Cochrane Database System Review*, CD003134.

Diaz de Leon, A., Cronkhite, J., Katzenstein, A., Godwin, J., Raghu, G., Glazer, C., et al. (2010). Telomere Lengths, Pulmonary Fibrosis and Telomerase (TERT) Mutations. *PLOS One*, 5(5):e10680.

Djojosubroto, M., Choi, T., Lee, H., & Rudolph, K. (2003). Telomeres and Telomerase in Aging, Regeneration and Cancer. *Molecules and Cells*, 15(2): 164-175.

Dokai, I. (2000). Dyskeratosis Congenita in all its Forms. *British Journal of Hematology*, 102: 15960- 15964.

Edwards, L. (2006). The Search for Genetic Loci Linked to Pulmonary Fibrosis in Newfoundland. *Honours Thesis, Memorial University*.

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). (2014, February 12). Seattle, WA, USA.

Farnham, P. (2012). Thematic Minireview Series on Results from the ENCODE Project: Integrative Global Analyses of Regulatory Regions in the Human Genome. *The Journal of Biological Chemistry*, 287: 30885-30887.

Fernandez, B., Fox, G., Bhatia, R., Sala, E., Noble, B., Denic, N., et al. (2012). A Newfoundland Cohort of Familial and Sporadic Idiopathic Pulmonary Fibrosis; Clinical and Genetic Features. *Respiratory Research*, 13(64): 1-10.

Fingerlin, T., Murphy, E., Zhang, W., Peljto, A., Brown, K., Steele, M., et al. (2013). Genome-Wide Association Study Identifies Multiple Susceptibility Loci for Pulmonary Fibrosis. *Nature Genetics*, 45: 613-620.

Fu, W., O'Connor, T., Jun, G., Kang, H., Abescasis, G., Leal, S., et al. (2013). Analysis of 6,515 Exomes Reveals the Recent Origin of Most Human Protein-Coding Variants. *Nature*, 493: 216-220.

Garber, K. (2013). At the Frontiers of Lung Fibrosis Therapy. *Nature Biotechnology*, 31(9): 781-783.

Gauldie, J. (2002). Pro: Inflammatory Mechanisms are a Minor Component of the Pathogenesis of Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory Critical Care Medicine*, 165: 1205-1206.

Gerull, B., Heuser, A., Wichter, T., Paul, M., Basson, C., McDermott, D., et al. (2004). Mutations in the Desmosomal Protein Plakophilin-2 are Common in Arrhythmogenic Right Ventricular Cardiomyopathy. *Nature Genetics*, 36: 1162-1164.

Gough, S., & Simmonds, M. (2007). The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Current Genomics*, 8(7):453-465.

Green, R., Green, J., Buehler, S., Robb, J., Daftary, D. G., McLaughlin, J., et al. (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.

Heiss, N., Knight, S., Vulliamy, T., Klauck, S., Wiemann, S., Mason, P., et al. (1998). X-Linked Dyskeratosis Congenita is Caused by Mutations in a Highly Conserved Gene with Putative Nucleolar Functions. *Nature*, 19: 32-38.

Hodgson, U., Laitinen, T., & Tukiainen, P. (2002). Nationwide Prevalence of Sporadic and Familial Idiopathic Pulmonary Fibrosis: Evidence of Founder Effect Among Mulitplex Families in Finland. *Thorax*, 57(4): 338-342.

Howe, P., Draetta, G., & Leof, E. (1991). Transforming Growth Factor Beta1 Inhibition of p34cdc2 Phosphorylation and Histone H1 Kinase Activity is Associated with G1-S Phase Growth Arrest. *Molecular Cell Biology*, 11:1185-1194.

Huang, D., Sherman, B., & Lempicki, R. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nature*, 4: 44-57.

Huang, F., & Chen, Y. (2012). Regulation of TGF-Beta Receptor Activity. *Cell and Bioscience*, 2:9.

Hubbard, R., Lewis, S., Richards, K., Johnston, I., & Britton, J. (1996). Occupational Exposure to Metal or Wood Dust and Aetiology of Cryptogenic Fibrosing Alveolitis. *Lancet*, 347: 284-289.

Javaherii, S., Leclerer, D., Pella, J., Mark, G., & Levine, B. (1980). Idiopathic Pulmonary Fibrosis in Monozygotic Twins: The Importance of Genetic Predisposition. *Chest*, 78: 591-594.

Ju, Y., Kim, J., Kim, S., Hong, D., Park, H., Shin, J., et al. (2011). Extensive Genomic and Transcriptional Diversity Identified through Massively Parallel DNA and RNA Sequencing of Eighteen Korean Individuals. *Nature Genetics*, 43: 745-752.

Kamel, F. (2010). Determining the Genetic Etiology of Familial Pulmonary Fibrosis in Six Newfoundland Families. *Masters thesis, Memorial University*.

Khalil, N., & Greenberg, A. (1991). The Role of TGF-Beta in Pulmonary Fibrosis. *Ciba Foundation Symposium*, 157: 194-207.

Khalil, N., O'Conner, R. F., & Unruh, H. (1996). TGF-Beta 1, but not TGF-Beta 2 or TGF-Beta 3, is Differentially Present in Epithelial Cells of Advanced Pulmonary Fibrosis: an Immunohistochemical Study. *American Journals of Respiratory Cell and Molecular Biology*, 14: 131-138.

King, T. J., Costabe, l. U., Cordier, J., doPico, G., du Bois, R., Lynch, D., et al. (2000). Idiopathic Pulmonary Fibrosis: Diagnosis and Treatment. *American Journal of Respiratory Critcal Care Medicine*, 161: 646-664.

Kirwan, M., & Dokal, I. (2009). Dyskeratosis Congenita, Stem Cells and Telomeres. *Biochimica and Biophysica Acta*, 1792(4): 371-379.

Kropski, J., Mitchell, D., Markin, C., Polosukhin, V., Choi, L., Johnson, J., et al. (2014). A Novel Dyskerin (DKC1) Mutation is Associated with Familial Interstitial Pneumonia. *Chest*.

Lacerte, A., Korah, J., Roy, M., Yang, X., Lemay, S., & Lebrun, J. (2008). Transforming Growth Factor Beta Inhibits Telomerase through SMAD3 and E2F Transcription Factors. *Cellular Signalling*, 20: 50-59.

Leask, A., & Abraham, D. (2004). TGF-Beta Signaling and the Fibrotic Response. *FASEB Journal*, 18: 816-827.

Lee, H., Ryu, J., Wittmer, M., Hartman, T., Lymp, J., Tazelaar, H., et al. (2005). Familial Idiopathic Pulmonary Fibrosis: Clinical Features and Outcome. *Chest*, 127(6): 2034-41.

Leong, P., Muhammad, R., Ibrahim, N., Cheong, S., & Seow, H. (2011). HLA-A and Breast Cancer in West Peninsular Malaysia. *Medical Oncology*, 28(1):51-56.

Li, A., Wang, D., Feng, X., & Wang, G. (2004). Latent TGF-Beta1 Overexpression in Keratinocytes Results in a Severe Psoriasis-Like Skin Disorder. *EMBO*, 23: 1770-1781.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078-2079.

Li, H., Xu, D., Li, J., Berndt, M., & Liu, J. (2006). Transforming Growth Factor Beta Supresses Human Telomerase Reverase Transcriptase (hTERT) by SMAD3 Interactions with c-Myc and the hTERT Gene. *The Journal of Biological Chemistry*, 281(35): 25588-25600.

Li, W., Wu, C., & Luo, C. (1984). Nonrandomness of Point Mutation as Reflected in Nucleotide Substitutions in Pseudogenes and its Evolutionary Implications. *Journal of Molecular Evolultion*, 21:58–71.

Litvinov, I., Bizet, A., Binamer, Y., Jones, D., D, S., & Philip, A. (2011). CD109 Release from the Cell Surface in Human Keratinocytes Regulates TGF-β Receptor Expression, TGF-β Signalling and STAT3 Activation: Relevance to Psoriasis. *Experimental Dermatology*, 20(8): 627-632.

Liu, T., Ullenbruch, M., Young-Choi, Y., Yu, H., Ding, L., Xaubet, A., et al. (2013). Telomerase and Telomere Length in Pulmonary Fibrosis. *American Journal of Respiratory Cell and Molecular Biology*, 49(2): 260-268.

Man, X., Finnson, K., Baron, M., & Philip, A. (2012). CD109, a TGF-β Co-receptor, Attenuates Extracellular Matrix Production in Scleroderma Skin Fibroblasts. *Arthritis Research Therapy*, 14(3): 144.

Marshall. (2000). Adult Familial Cryptogenic Fibrosing Alveolitis in the United Kingdom. *Thorax*, 55(2):143-146.

Martin, P. (1997). Wound Healing. Aiming for Perfect Skin Regeneration. *Science*, 276: 75-81.

Mathai, S., Schwartz, D., & Warg, L. (2014). Genetic Susceptibility and Pulmonary Fibrosis. *Current Opinion Pulmonary Medicine*, 20:000-000.

Mathieson, I., & McVean, G. (2012). Differenital Confounding of Rare and Common Variants in Spatially Structured Populations. *Nature Genetics*, 44(3): 243-246.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20:1297-1303.

Merner, N., Hodgkinson, K., Haywood, A., Connors, S., French, V., Drenckhahn, J., et al. (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.

Møller, P., Clark, N., & Mæhle, 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.

Nakerakanti, S., & Trojanowska, M. (2012). The Role of TGF-Beta Receptors in Fibrosis. *The Open Rheumatology Journal*, 156-192.

Nalysnyk, L., Cid-Ruzafa, J., Rotella, P., & Esser, D. (2012). Incidence and prevalence of idiopathic pulmonary fibrosis: a review of the literature. *European Respiratory Review*, 21(126): 355-61.

Ng, P., & Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes that Affect Protein Function. *Nucleic Acids Research*, 31(13):3812-3814.

Ng, S., Buckingham, K., Lee, C., Bigham, A., Tabor, H., Dent, K., et al. (2010). Exome Sequencing Identifies the Cause of a Mendelian Disorder. *Nature*, 42: 30-35.

Ng, S., Turner, E., Robertson, P., Flygare, S., Bigham, A., Lee, C., et al. (2009). Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes. *Nature*, 461: 272-276.

Nicholson, A., Florio, R., Hansell, D., Bois, R., Wells, A., Hughes, P., et al. (2006). Pulmonary Involvement by Niemann- Pick Disease. A Report of Six Cases. *Histopathology*, 48(5): 596-603.

Nogee, L., Dunbar, A., & Wert, S. (2001). A Mutation in the Surfactant Protein C Gene Associated with Familial Interstitial Lung Disease. *New England Journal of Medicine*, 344: 573-579.

Online Mendelian Inheritance in Man, O. (2013, 07 2013). *Idiopathic Pulmonary Fibrosis (OMIM #178500).* Baltimore, MD.: Johns Hopkins University.

Ortiz, L., Lasky, J., Hamilton, R., Holian, A., Hoyle, G., Banks, W., et al. (1998). Expression of TNF and the Necessity of TNF Receptors in Bleomycin-Induced Lung Injury in Mice. *Experimental Lung Research*, 24: 721-743.

Pirzada, A. (2012). *Inherited Predispostion to Idiopathic Pulmonary Fibrosis in the Newfoundland Population.* Memorial University: Masters Thesis.

Polakoff, P., Horn, B., & Scherer, O. (1979). Prevalence of Radiographic Abnormalities Among Northern California Shipyard Workers. *Annals of the New York Academy of Sciences*, 330: 293-311.

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., et al. (2014). The Pfam Protein Families Database. *Nucleic Acids Research*.

Raddatz-Sikkema, B., Johansson, L., de Boer, E., Almomani, R., Boven, L., van den Berg, M., et al. (2013). Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. *Human Mutation*, 1-8.

Raghu, G., Collard, H., Egan, J., Martinez, F., Behr, J., K, B., et al. (2011). An Offical ATS/ERS/JRS/ALAT Statement: Idiopathic Pulmonary Fibrosis: Evidence-Based Guidelines for Diagnosis and Management. *American Journal of Respiratory and Critical Care Medicine*, 183: 788-824.

Rahman, P., Jones, A., Curtis, J., Bartlett, S., Peddle, L., Fernandez, B., et al. (2003). The Newfoundland Population: A Unique Resource for Genetic Investigation of Complex Diseases. *Human Molecular Genetics*, 12: 167-172.

Rashid, R., Liang, B., Baker, D., Youssef, O., He, Y., Phipps, K., et al. (2006). Crystal Structure of a Cbf5-Nop10-Gar1 Complex and Implications in RNA-Guided Pseudouridylation and Dyskeratosis Congenita. *Molecular Cell*, 21(2): 249-260.

Robinson, C., Neary, R., Levendale, A., Watson, C., & Baugh, J. (2012). Hypoxia-Induced DNA Hypermethylation in Human Pulmonary Fibroblasts is Associated with Thy-1 Promoter Methylation and the Development of a Pro-Fibrotic Phenotype. *Respiratory Research*, 13: 74.

Rosenkranz, S. (2004). TGF-Beta1 and Angiotension Networking in Cardiac Remodeling. *Cardiovascular Research*, 63: 423-432.

Sarovich, D., & Price, E. (2014). SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. *BMC Research Notes*, 7:618.

Sato, T., Murakumo, Y., Hagiwara, S., Jijwa, M., Suzuki, C., Yatabe, Y., et al. (2007). High-Level Expression of CD109 is Frequently Detected in Lung Squamous Cell Carcinomas. *Pathology International*, 57: 719-724.

Schurawitzki, H., Stiglbauer, R., Graninger, W., Herold, C., Pölzleitner, D., Burghuber, O., et al. (1990). Interstitial Lung Disease in Progressive Systemic Sclerosis: High-Resolution CT Versus Radiography. *Radiology*, 176: 755-759.

Schwartzentruber, J. (2012). *McGill University and Genome Quebec*. Retrieved December 3, 2013, from Tutorials - FORGE Canada's exome sequencing analysis: http://gqinnovationcenter.com/services/bioinformatics/tutorials/tutorials.aspx?l=e

Seibold, M., Wise, A., Speer, M., Steele, M., Brown, K., Loyd, J., et al. (2011). A Common MUC5B Promoter Polymorphism and Pulmonary Fibrosis. *The New England Journal of Medicine*, 364: 1503-1512.

Sellick, G., Barker, K., Stolte-Dijkstra, I., Fleischmann, C., Coleman, R., Garrett, C., et al. (2004). Mutations in PTF1A Cause Pancreatic and Cerebellar Agenesis. *Nature Genetics*, 36(12): 1301-1305.

Selman, M., & Pardo, A. (2014). Revealing the Pathogenic and Aging-Related Mechanisms of Enigmatic Idiopathic Pulmonary Fibrosis. An Integral Model. *American Journal of Respiratory Critical Care Medicine*, 189(10):1161-72.

Selman, M., Lin, H. M., Jenkins, A., Estrada, A., Lin, Z., Wang, G., et al. (2003). Surfactant Protein A and B Genetic Variants Predispose to Idiopathic Pulmonary Fibrosis. *Human Genetics*, 113: 542-550.

Semczuk, A., Zakrzewski, P., Forma, E., Cygankiewicz, A., Semczuk-Sikora, A., Brys, M., et al. (2013). TGF-Beta Pathway is Down-Regulated in a Uterine Carcinosarcoma: A Case Study. *Pathology- Research and Practice*, 344-348.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., et al. (2001). dbSNP: the NCBI database of genetic variation. . *Nucleic Acids Research*, 29(1):308-11.

Society, A. T. (2000). Idiopathic Pulmonary Fibrosis: Diagnosis and Treatment. International Consensus Statement. American Thoracic Society (ATS) and European Respiratory Society (ERS). *American Journal of Respiratory and Critical Care Medicine*, 161: 646-664.

Solomon, J., Olson, A., Fischer, A., Bull, T., Brown, K., & Raghu, G. (2013). Scleroderma Lung Disease. *European Respiratory*, 22(127): 6-19.

Steele, M., & Schwartz, D. (2013). Molecular Mechanisms in Progressive Idiopathic Pulmonary Fibrosis. *Annual Review of Medicine*, 64: 265-276.

Stitziel, N., Kiezun, A., & Sunya, S. (2011). Computational and Statistical Approaches to Analyzing Variants Identified by Exome Sequencing. *Genome Biology*, 12: 227-237.

Strieter, R. (2002). Con: Inflammatory Mechanisms are not a Minor Component of the Pathogenesis of Idiopathic Pulmonary Fibrosis. *Antioxid Redox Signal*, 10: 287-301.

Stuckless, S., Green, J., Dawson, L., Barrett, B., Woods, M., Dicks, E., et al. (2013). Impact of Gynecological Screening in Lynch Syndrome Carriers with an MSH2 Mutation. *Clinical Genetics*, 83(4):359-364.

Talbert, J., & Schwartz, D. (2005, January 21). *Pulmonary Fibrosis, Familial*. Retrieved 2015, from GeneReviews: http://www.ncbi.nlm.nih.gov/books/NBK1230/

Tatler, A., & Jenkins, G. (2012). TGF-Beta Activation and Lung Fibrosis. *Proceedings of American Thoracic Society*, 9(3): 130-136.

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56-65.

Thomas, A., Lane, K., Phillips, J., & al, e. (2002). Heterozygosity for a Surfactant Protein C Gene Mutation Associated with Usual Interstitial Pneumonitis and Cellular Nonspecific Interstitial Pneumonitis in One Kindred. *American Journal of Respiratory and Critical Care Medicine*, 165: 1322-1328.

Toya, S., & Tzelpis, G. (2009). The Many Faces of Scleroderma Sine Scleroderma: A Literature Review Focusing on Cardiopulmonary Complications. *Rheumatology International*, 29(8): 861-868.

van Moorsel, C., van Oosterhout, M., Barlo, N., de Jong, P., van der Vis, J., Ruven, H., et al. (2010). Surfactant Protein C Mutations are the Basis of a Significant Portion of Adult Familial Pulmonary Fibrosis in a Dutch Cohort. *American Journal of Respiratory and Critical Care Medicine*, 182: 1419-1425.

Varga, J., & B, P. (2009). Transforming Growth Factor Beta is a Therapeutic Target in Systemic Sclerosis. *Nature Reviews Rheumatology*, 5: 200-206.

Vorstenbosch, J., Al-Ajmi, H., Winocour, S., Trzeciak, A., Lessard, L., & Phillip, A. (2013). CD109 Overexpression Ameliorates Skin Fibrosis in a Mouse Model of Bleomycin-Induced Sceleroderma. *Arthertis and Rheumatism*, 65(5): 1378-1383.

Vulliamy, T., & Dokal, I. (2006). Dyskeratosis Congenita. *Seminars in Hematology*, 43: 157-166.

Vulliamy, T., Marrone, A., Goldman, F., Dearlove, A., Bessler, M., Mason, P., et al. (2001). The RNA Component of Telomerase is Mutated in Autosomal Dominant Dyskeratosis Congenita. *Nature*, 413: 432-435.

Walne, A., Vulliamy, T., & Marrone, A. e. (2007). Genetic Heterogeneity in Autosomal Recessive Dyskeratosis Congenita with one Subtype due to Mutations in the Telomerase-Associated Protein NOP10. *Human Molecular Genetics*, 16(13): 1619-1629.

Wang, K., Mingyao, L., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acid Research*, 38(16): e164.

Wang, Y., Kuan, P., Xing, C., Cronkhite, K., Torres, F., Rosenblatt, R., et al. (2009). Genetic Defects in Surfactant Protein A2 are Associated with Pulmonary Fibrosis and Lung Cancer. *American Journal of Human Genetics*, 84:52-59.

Warburton, D., Shi, W., & Xu, B. (2013). TGF-Beta- Smad3 Signalling in Emphysema and Pulmonary Fibrosis: An Epigenetic Aberration of Normal Development. *American Physiological Society*, 304(2): 83-85.

Weedon, M., Cebola, I., Patch, A., Flanagan, S., De Franco, E., Caswell, R., et al. (2014). Recessive Mutations in a Distal PTF1A Enhancer Cause Isolated Pancreatic Agenesis. *Nature Genetics*, 46: 61-64.

Wei, R., Li, C., Zhang, M., Jone-Hall, Y., Myers, J., Noth, I., et al. (2014). Association between MUC5B and TERT Polymorphisms and Different Interstitial Lung Disease Phenotypes. *Translational Research*, 163(5): 494-502.

Wells, A., Desai, S., Rubens, M., Goh, N., Cramer, D., Nicholson, A., et al. (2003). Idiopathic Pulmonary Fibrosis: A Composite Physiologic Index Derived from

Disease Extent Observed by Computed Tomography. *Americal Journal of Respiratory Critical Care Medicine*, 277(33):34017-34023.

Wijayawardhana, D. (1999). *Map of the Island of Newfoundland*. Retrieved October 21, 2013, from Newfoundland and Labrador Heritage Web Site: http://www.heritage.nf.ca/nfld_fullmap.html

Wilkes, M., Mitchell, H., Penheiter, S., Dore, J., Suzuki, K., Edens, M., et al. (2005). Transforming Growth Factor-Beta Activation of Phosphatidylinositol 3-Kinase is Independent of Smad2 and Smad3 and Regulates Fibroblast Responses via p21-Activated Kinase-2. *Cancer Research*, 65(22):10431-10440.

Wilson, M., & Wynn, T. (2009). Pulmonary Fibrosis: Pathogenesis, Etiology and Regulation. *Mucosal Immunology*, 2: 103- 121.

Winocour, S., Vorstenbosch, J., Trzeciak, A., Lessard, L., & Phillip, A. (2014). CD109, a Novel TGF-β Antagonist, Decreases Fibrotic Responses in a Hypoxic Wound Model. *Experimental Dermatology*, doi: 10.1111/exd.12439.

Wojtyla, A., Gladych, M., & Rubis, B. (2011). Human Telomerase Activity Regulation. *Molecular Biology Reports*, 38(5): 3339-3349.

Wright, J. (2004). Host Defense Functions of Pulmonary Surfactant. *Biology of the Neonate*, 85(4):326-32.

Wuyts, W., Agostini, C., Antoniou, K., Bouros, D., Chambers, R., Cottin, V., et al. (2013). The Pathogenesis of Pulmonary Fibrosis: A Moving Target. *European Respiratory Journal*, 41:1207–1218.

Xaubet, A., Marin-Arguedas, A., Lario, S., Ancochea, J., Morell, F., Ruiz-Manzano, J., et al. (2003). Transforming growth factor-beta1 gene polymorphisms are associated with disease progression in idiopathic pulmonary fibrosis. *American Journal of Respirology Critical Care Medicine*, 168(4):431-435.

Xie, Y., Zheng, H., Leggo, J., Scully, M., & D, L. (2002). A Founder Factor VIII Mutation, Valine 2016 to Alanine, in a Population with an Extraordinarily High Prevalence of Mild Hemophilia A. *Thrombosis Haemostatsis*, 87(1): 178-179.

Xu, Y., Hua, J., Mui, A., O'Conner, R., Grotendorst, G., & Khalil, N. (2003). Release of Biologically Active TGF-Beta 1 by Alveolar Epithelial Cells Results in Pulmonary Fibrosis. *American Journal of Physiology*, 285: L527-L539.

Yang, L., Chen, Y., Cui, T., Knosel, T., Zhang, Q., Albring, K., et al. (2012). Desmoplakin acts as a Tumor Suppressor by Inhibition of the Wnt/β-Catenin Signaling Pathway in Human Lung Cancer. *Carcinogenesis*, 33(10): 1863-1870.

Young, T., Penny, L., Woods, M., Parfrey, P., J, G., Hefferton, D., et al. (1999). A Fifth Locus for Bardet-Biedl Syndrome Maps to Chromosome 2q31. *American Journal of Genetics*, 64(3):900-904.

Zhang, A., Zheng, C., Hou, M., Lindvall, C., Li, K., Erlandsson, F., et al. (2003). Deletion of the Telomerase Reverse Transcriptase Gene and Haploinsufficiency of Telomere Maintenance in Cri du Chat Syndrome. *American Journal of Human Genetics*, 72(4): 940-948.

Zhang, J., Xu, D., Wu, B., Zheng, M., Chen, J., & Huang, J. (2012). HLA-A and HLA-B Gene Polymorphism and Idiopathic Pulmonary Fibrosis in a Han Chinese Population. *Respiratory Medicine*, 106(10): 1456-1462.

Zhang, Y., Noth, I., Garcia, J., & Kaminski, N. (2011). A Variant in the Promoter of MUC5B and Idiopathic Pulmonary Fibrosis. *New England Journal of Medicine*, 364(16): 1576-1577.

Appendices

Appendix A: High Resolution Commuted Tomography Scoring Rubric for the Diagnosis of Interstitial Lung Disease

HRCT Scoring Chart

Family _____
Patient Name_____ **DOB**_____ **Study # NFPF**_____
Radiologist _____ **CT Scan Date**_____
High Resolution - Yes ( ) No ( ) Appropriate Images Yes ( ) No ( )
Comments_____

*** Each film must be evaluated at the following 5 levels***

1. **The origin of the great vessels**
2. **The mid arch of the aorta**
3. **The main carina**
4. **The pulmonary venous confluence**
5. **1 cm above the higher of the 2 hemidiaphragms (usually the right)**

I. Estimation of the total extent of disease at each level to the nearest 5%
(combination of extent of reticular disease and ground glass)

II. Coarseness of reticular abnormality - Grade coarseness (0-3)

1. **Predominantly fine intralobular thickening**
2. **Predominantly microcystic honeycombing (comprising air spaces < 4mm in diameter)**
3. **Predominantly macrocystic honeycombing (comprising air spaces > 4mm in diameter)**

III. Overall Extent of Emphysema

- **destruction of lung parenchyma resulting in decreased attenuation without visible walls (to nearest 5%) for same 5 levels**

Level	Proportion of abnormal lung (%)	Coarseness of reticular abnormality - Grade coarseness (0-3)	Extent of Emphysema
1			
2			
3			
4			
5			
Total	Overall Proportion (%) _____ (average of the 5 levels)	Total (max 15) _____	% Overall _____

IV. Ratio of reticular disease / ground glass (eg. 70/30, 50/50) _____

V. Airtrapping Yes ( ) degree _____ **No ( )**
VI. Overall CT appearance

Appendix B: Promega Deoxyribonucleic Acid Extraction from Blood

Purpose: To extract highly purified DNA from whole blood

Materials:

1) Blood collected in EDTA tube
2) Laminar flow Biosafety Containment Cabinet.
3) Centrifuge
4) Vortex
5) 50ml sterile centrifuge tubes
6) 1.5 ml microcentrifuge tubes
7) Wizard® Genomic Purification Kit Cat #PRA-1620
8) Isopropanol room temperature
9) 70% Ethanol

Method: For 12-16ml whole blood

1) In the Laminar Flow Biosafety Containment Cabinet, add 30 ml of Cell Lysis Solution to one 50 ml sterile centrifuge tube.
2) Gently rock the tubes of blood until mixed thoroughly. Add 12 ml of whole blood to the 50 ml sterile centrifuge tube containing the 30 ml of Cell Lysis Solution. Invert the tube 5-6 times
3) Incubate the mixture at room temperature for 10 minutes (invert 5-6 times half way through incubation).
4) Centrifuge the mixture at 2000 x g for 10 minutes at room temperature.
5) Remove and discard as much supernatant as possible without disturbing the pellet at the bottom of the tube.
6) Vortex and add 10 ml of Nuclei Lysis Solution and vortex 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 x g for 10 minutes at room temperature.
9) Add 10 ml of Isopropanol to a fresh, sterile 50 ml centrifuge tube.
10) Gently pour the supernatant into the 50 ml centrifuge containing 10 ml of Isopropanol.
11) Centrifuge at 2000 x g for 2 minutes at room temperature.
12) Wash with 70% ethanol and repeat step 11.
13) Let air dry.
14) Dissolve DNA into 400 µl of Rehydration Buffer overnight.
15) Mix sample and quantify concentration.
16) DNA sample can be stored at 4ºC or frozen for an extended period.

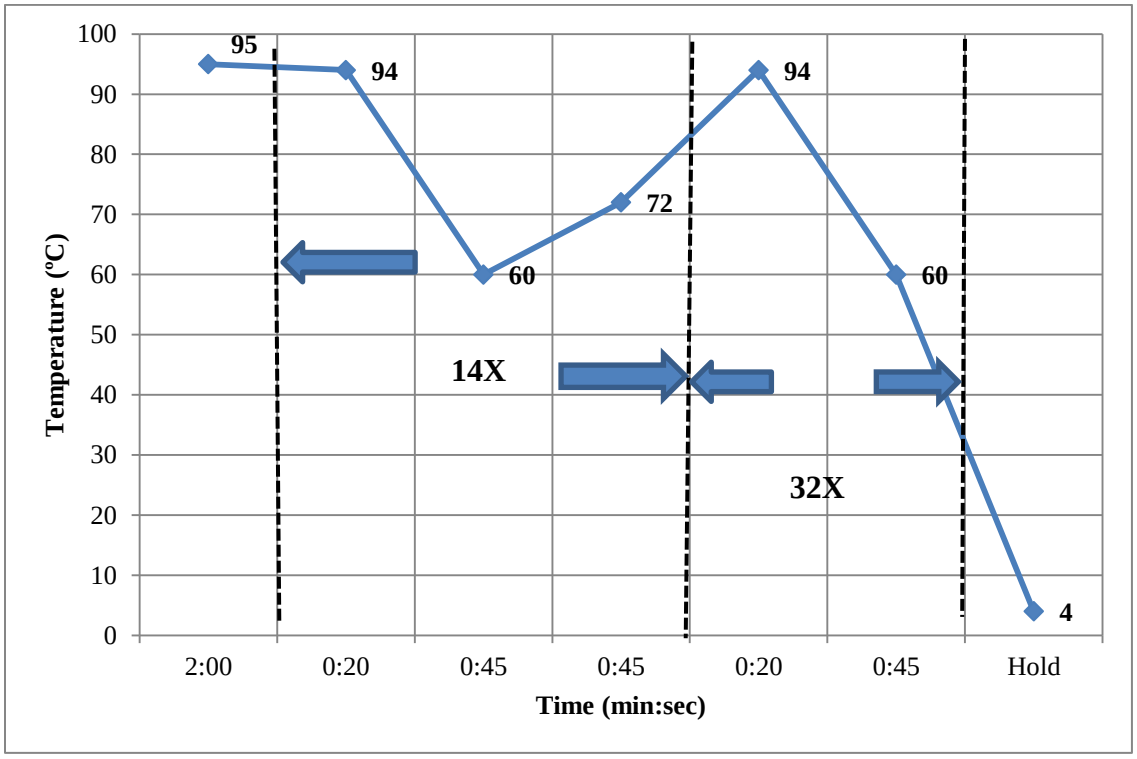
Appendix C: Primer Sequences and Thermocycler Protocols for All Genes Sequenced

Gene	Exon	Forward Primer	Thermocycler Program
CD109	*CD109*_Exon_4_F *CD109*_Exon_4_R	ttgcaattaaaacatttttgagg caagcagaacttcccagagg	Touchdown
	*CD109*_Exon_5_F *CD109*_Exon_5_R	ggaatgctgcaaggcattat attttcccccacctgaccta	Standard @ 65ºC
	*CD109*_Exon_6_F *CD109*_Exon_6_R	tgcaagtaagtaagcaagcaaaa aaacaccagaattttaactccattg	Standard @ 65ºC
	*CD109*_Exon_12_F *CD109*_Exon_12_R	tgaaaatccaatgtctggtga ctgcacatgtaccccagaac	Multiplex
	*CD109*_Exon_13_F *CD109*_Exon_13_R	ggtaatgcttacaattcacttgaga ctgcacatgtaccccagaac	Multiplex
	*CD109*_Exon 14_F *CD109*_Exon 14_R	tttgaagctggtttaattcaaga ttctaaccagtttggcttaatga	Multiplex
	*CD109*_Exon_15_F *CD109*_Exon_15_R	tgagccaaggcaaatgtaga ttgacttacacatgcaaaggaac	Touchdown
	*CD109*_Exon_18_F *CD109*_Exon_18_R	cctgatactctcactggcaca aacattagaaagcgggcaaa	Touchdown
	*CD109*_Exon19_F *CD109*_Exon19_R	ttgaaaagttgggatttactgtttt ggactttgcacatttagcagttt	Touchdown
	*CD109*_Exon_20_F *CD109*_Exon_20_R	tcacaggttttataaagattgcattt cccaaatgcccagatactaca	Touchdown
	*CD109*_Exon_21_F *CD109*_Exon_21_R	cagatttgacagtacttaaactggtga gccaacaaatttaaccaagagaa	Standard @ 63ºC
	*CD109*_Exon_23_F *CD109*_Exon_23_R	ggaagtctctttgccaccag acaccgtgtagcccatcact	Standard @ 65ºC
	*CD109*_Exon_24_F *CD109*_Exon_24_R	tctctgcccatatttcctcaa ggaaaaagcaacctcccagt	Standard @ 65ºC
	*CD109*_Exon_25_F *CD109*_Exon_25_R	tcttctgatgggggaaaaag ccacaaatgcaaagcaaaag	Touchdown
	*CD109*_Exon_26_F *CD109*_Exon_26_R	tggttgattttcattgcctagt tcaatttcaaacctggctaaaa	Touchdown
	*CD109*_Exon_27_F *CD109*_Exon_27_R	acaacttggaccagagtaggaa tgaaactgccaagtaattttatga	Touchdown
	*CD109*_Exon_28_F *CD109*_Exon_28_R	tgaatatccagcaacaaaatactact cagcataaagaattgtctacactgat	Standard @ 63ºC
	*CD109*_Exon_29_F *CD109*_Exon_29_R	tttcctaatgagggataggactg tgcttaagtcacagatgtaaatcatta	Touchdown
	*CD109*_Exon_30_F *CD109*_Exon_30_R	tcttactgggtccatccaaag tgcataactgaattctttgttcaat	Touchdown

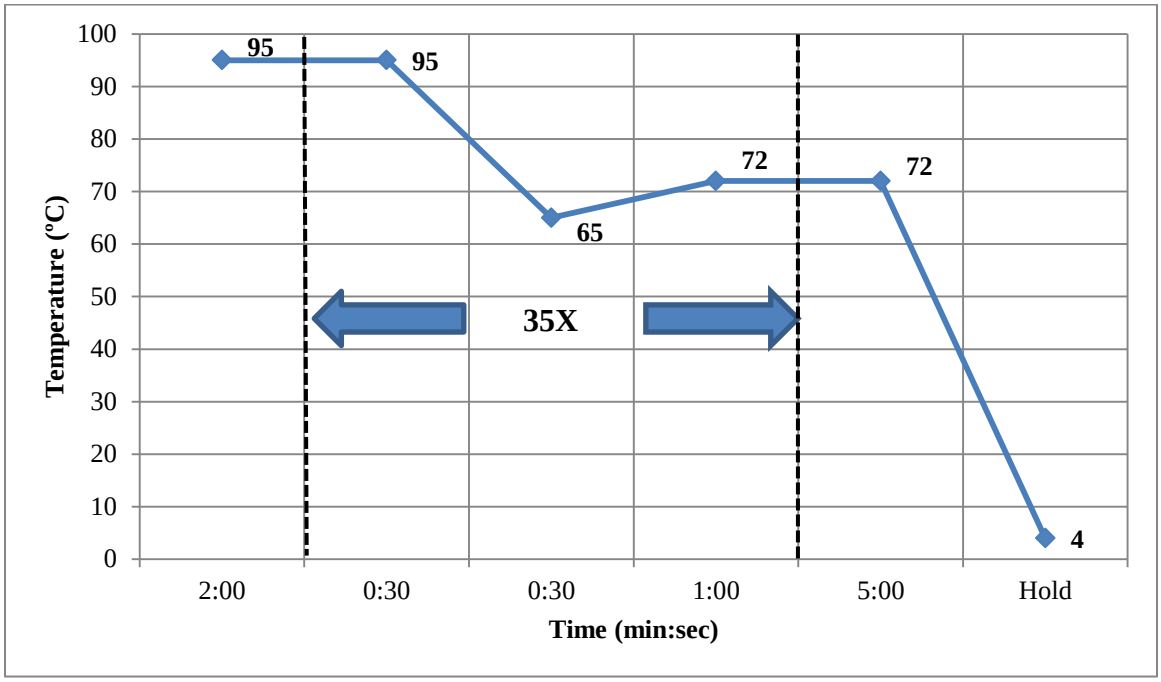
	*CD109*_Exon_31_F *CD109*_Exon_31_R	tttggtgggtttgatttttgt tgccatgttaaactcattcagg	Touchdown
	*CD109*_Exon_32_F *CD109*_Exon_32_R	tccctaagtgttagtctctgtttcc accaatccaatgtgtggtca	Standard @ 65ºC
	*CD109*_Exon_33_F *CD109*_Exon_33_R	ttggacatttcagttgtttcttg ttctacgaaaacaaaacaatcaca	Standard @ 63ºC
CIAPIN 1	*CIAPIN1*_Exon_9_F *CIAPIN1*_Exon_9_R	catttccttcaggccactgt cccatgtcaggaacctccta	Standard @ 65ºC
DSP	*DSP*_Exon_1_F *DSP*_Exon_1_R	gtagcgagcagcgacctc ccgctggtcacctcgtag	GC Rich
FGFR4	*FGFR4*_Intron_2_F *FGFR4*_Intron_2_R	gcagcatgtgtggtataagca gacacccacacagccacat	Touchdown
HLA-A	*HLA-A*_Exon_4_F *HLA-A*_Exon_4_R	ctgactcttcccgtcagacc cttaccccatctcagggtga	Touchdown
IL-32	*IL32*_Exon_7_F *IL32*_Exon_7_R	ctggggagagcttttgtgac ctccgcgatcatggtatctc	Touchdown
SFTPA2	*SFTPA2*_Intron_2_F *SFTPA2*_Intron_2_R	gctccccagagctccttacct aacatctcccccactgtgct	Touchdown
TEP1	*TEP1*_Exon_29_F *TEP1*_Exon_29_R	ttcctccacactgtcctcct cactgaccactccgtgtgac	Standard @ 65ºC
TERF1	*TERF1*_Exon_1_F *TERF1*_Exon_1_R	gaggaggaggaggaggagga acgtagggaaccagccctct	Standard @ 69ºC

Appendix D: Thermocycler Programs

"Touchdown" Thermocycler Program

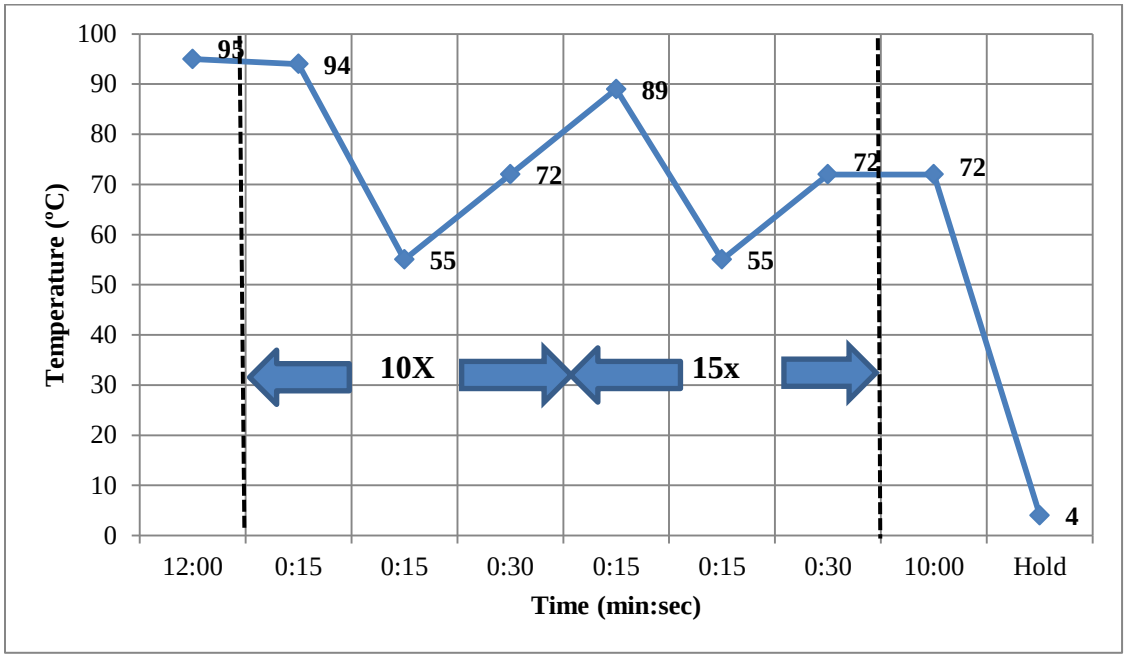


“Standard” Thermocycler Program

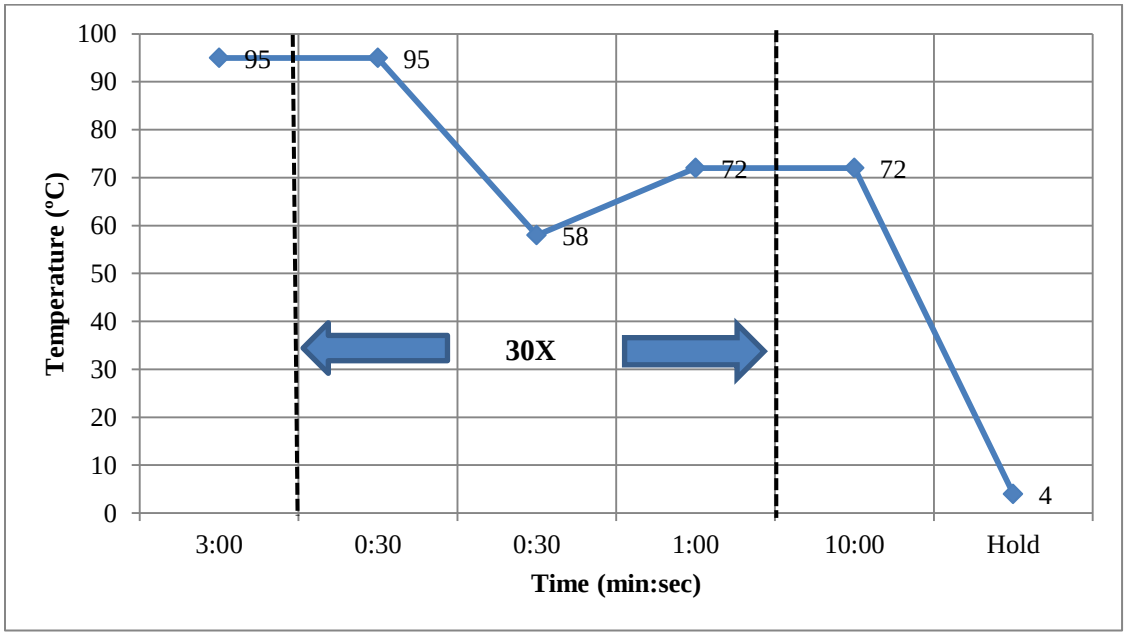


For each “Standard” Thermocycler Program used, the 55 ºC temperature was changed according to the specific PCR protocol, as shown in Appendix C. Each “Standard” Thermocycler Program was entitled based on the temperature used. For example, the “Standard” Thermocycler Program Standard 65 used an annealing temperature of 65ºC, whereas Standard63 used an annealing temperature of 63ºC.

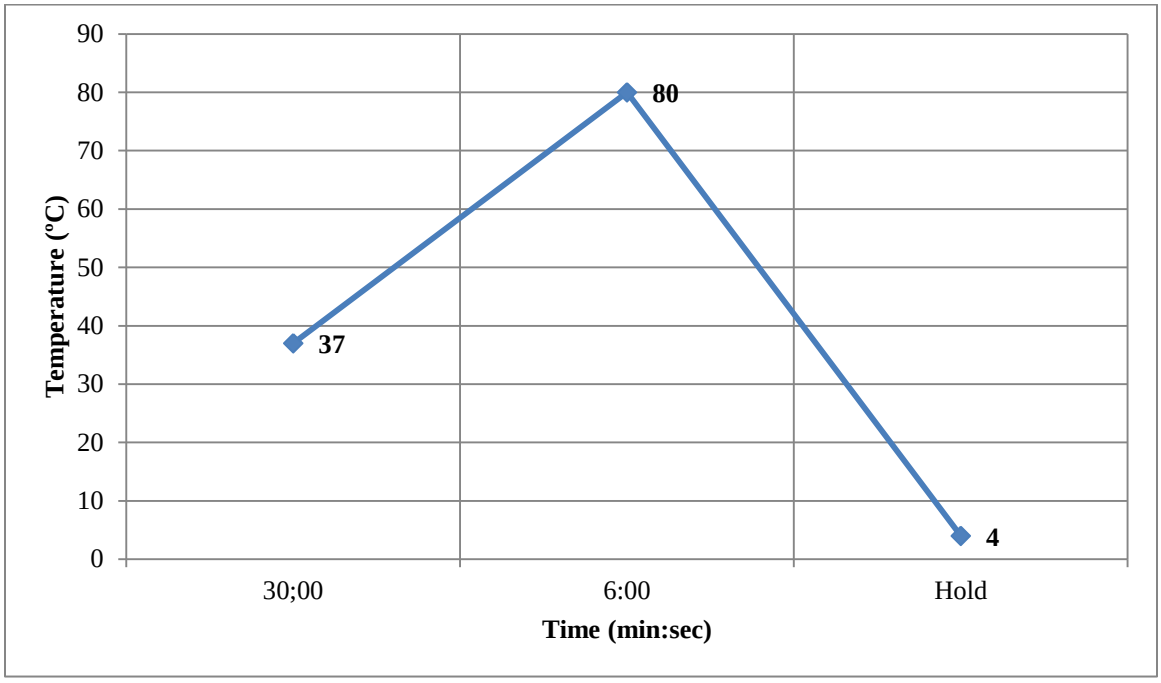
“Multiplex” Thermocycler Program



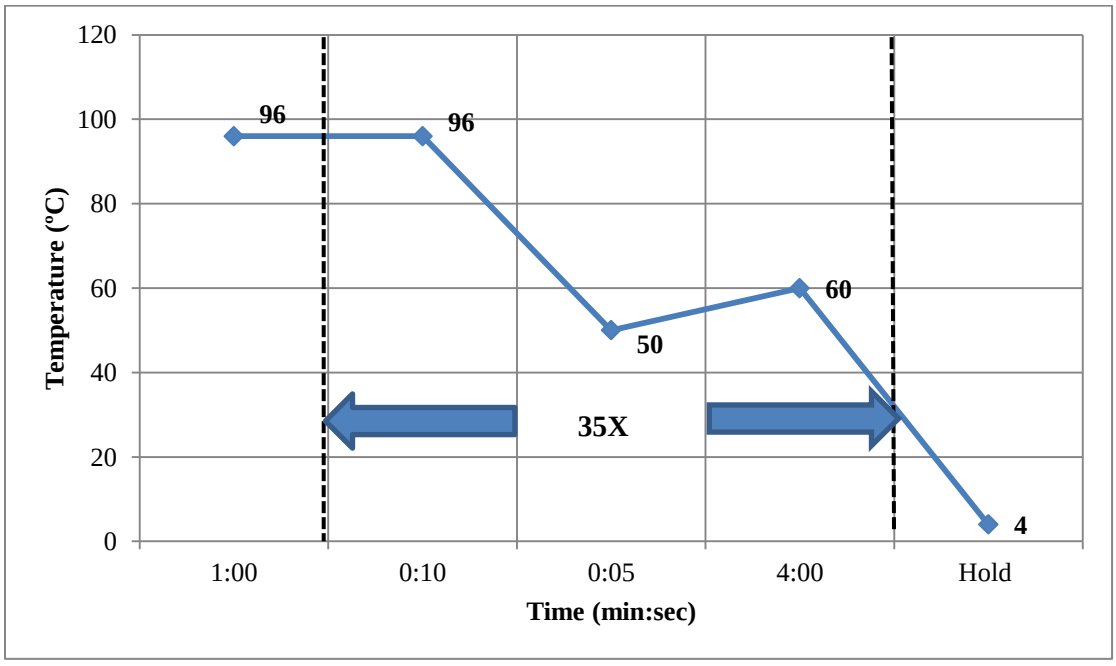
“GC Rich” Thermocycler Program



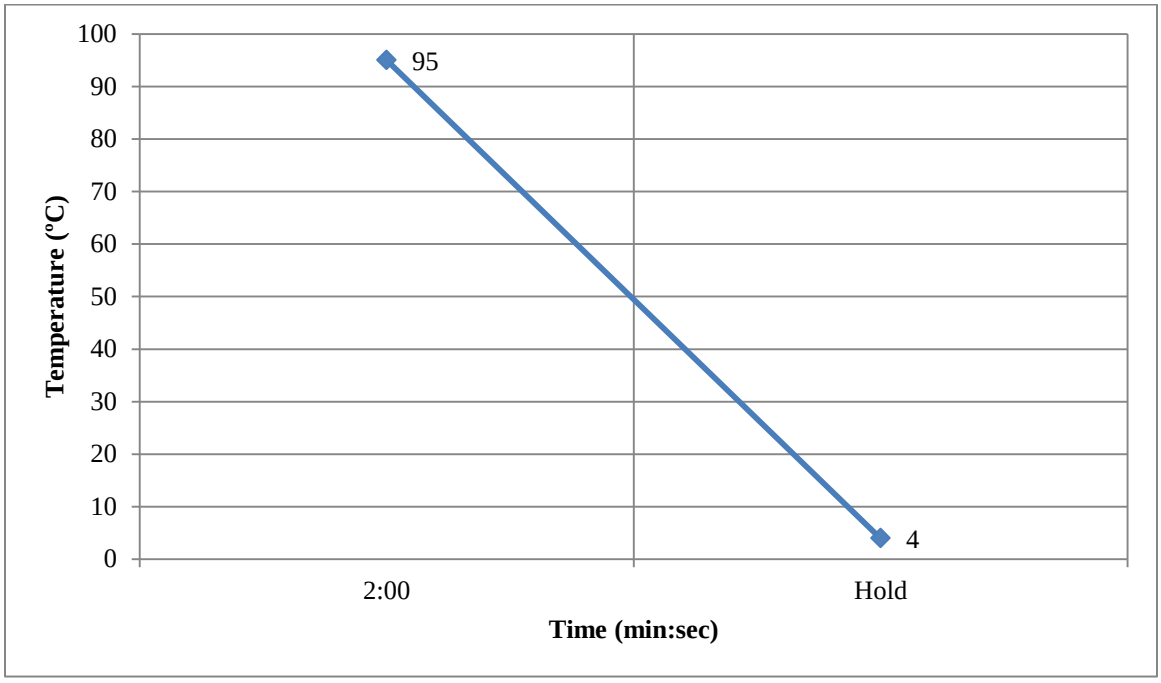
“Exosap” Thermocycler Program



“ABIseq” Thermocycler Program



“Denature” Thermocycler Program



Appendix E: Publisher's Permission to use Copyright Materials

This is a License Agreement between Robyn L Byrne ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Robyn L Byrne
Customer address	300 Prince Phillip Drive St. John's, NL A1B 3V6
License number	3372021510034
License date	Apr 18, 2014
Licensed content publisher	Elsevier
Licensed content publication	The Lancet
Licensed content title	Idiopathic pulmonary fibrosis
Licensed content author	Talmadge E King,Annie Pardo,Moisés Selman
Licensed content date	3–9 December 2011
Licensed content volume number	378
Licensed content issue number	9807
Number of pages	13
Start Page	1949
End Page	1961
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of your thesis/dissertation	THE UTILIZATION OF WHOLE EXOME SEQUENCING IN DETERMINING NOVEL GENETIC VARIANTS IN PULMONARY FIBROSIS SUSCEPTIBILITY GENES IN NEWFOUNDLAND FAMILIES
Expected completion date	Sep 2014
Estimated size (number of pages)	130
Elsevier VAT number	GB 494 6272 12

This is a License Agreement between Robyn L Byrne ("You") and Nature Publishing Group ("Nature Publishing Group"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

Get the printable license.

License Number	3374831499329
License date	Apr 23, 2014
Licensed content publisher	Nature Publishing Group
Licensed content publication	Mucosal Immunology
Licensed content title	Pulmonary fibrosis: pathogenesis, etiology and regulation
Licensed content author	M S Wilson and T A Wynn
Licensed content date	Jan 7, 2009
Type of Use	reuse in a dissertation / thesis
Volume number	2
Issue number	2
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	THE UTILIZATION OF WHOLE EXOME SEQUENCING IN DETERMINING NOVEL GENETIC VARIANTS IN PULMONARY FIBROSIS SUSCEPTIBILITY GENES IN NEWFOUNDLAND FAMILIES
Expected completion date	Sep 2014
Estimated size (number of pages)	130

This is a License Agreement between Robyn L Byrne ("You") and Nature Publishing Group ("Nature Publishing Group"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

Get the printable license.

License Number	3452640327371
License date	Aug 19, 2014
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Exome sequencing as a tool for Mendelian disease gene discovery
Licensed content author	Michael J. Bamshad, Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson, Jay Shendure
Licensed content date	Sep 27, 2011
Type of Use	reuse in a dissertation / thesis
Volume number	12
Issue number	11
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Box 1 | Workflow for exome sequencing
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	THE UTILIZATION OF WHOLE EXOME SEQUENCING IN DETERMINING NOVEL GENETIC VARIANTS IN PULMONARY FIBROSIS SUSCEPTIBILITY GENES IN NEWFOUNDLAND FAMILIES
Expected completion date	Sep 2014
Estimated size (number of pages)	130
Total	0.00 USD

MEMORIAL UNIVERSITY

Byrne, Robyn <v25rlb@mun.ca>

Copyright permission

ATS Permission Requests <permissions@thoracic.org> Tue, Jul 29, 2014 at 11:36 AM
To: "Byrne, Robyn" <v25rlb@mun.ca>

Hello Robyn,

Thank you for your interest in the American Thoracic Society journals.

Your request is granted for no charge. Please be sure to include the below wording. Thank you.

Reprinted with permission of the American Thoracic Society. Copyright © 2014 American Thoracic Society.

Cite: Author(s)/Year/Title/Journal Title/Volume/Pages.

Official Journal of the American Thoracic Society.

Best,

Lan Vay

Journal Program Coordinator

American Thoracic Society

25 Broadway, 18th Floor

New York, NY 10004-1012

http://www.atsjournals.org

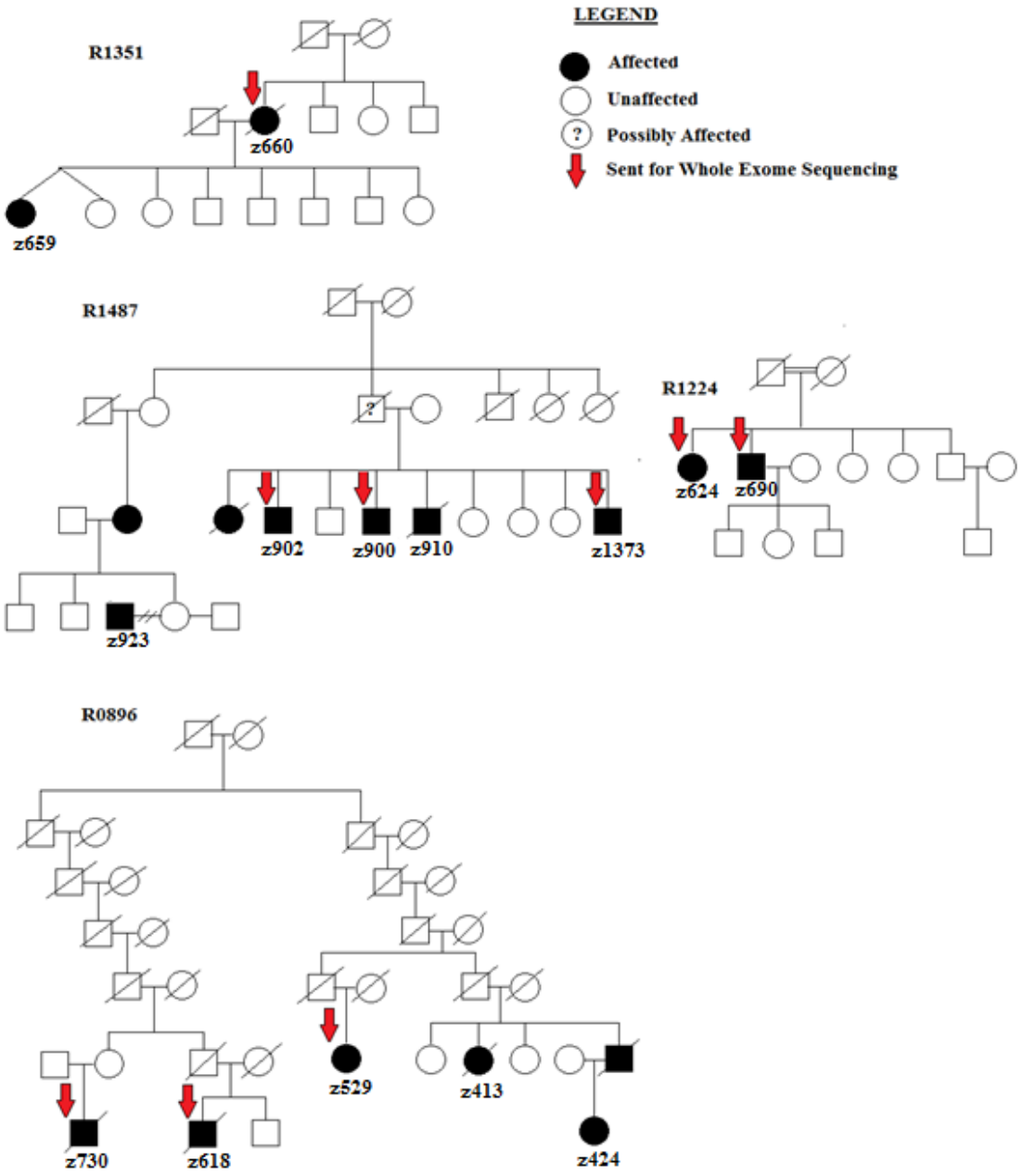
lvay@thoracic.org

Phone: 212-315-6440

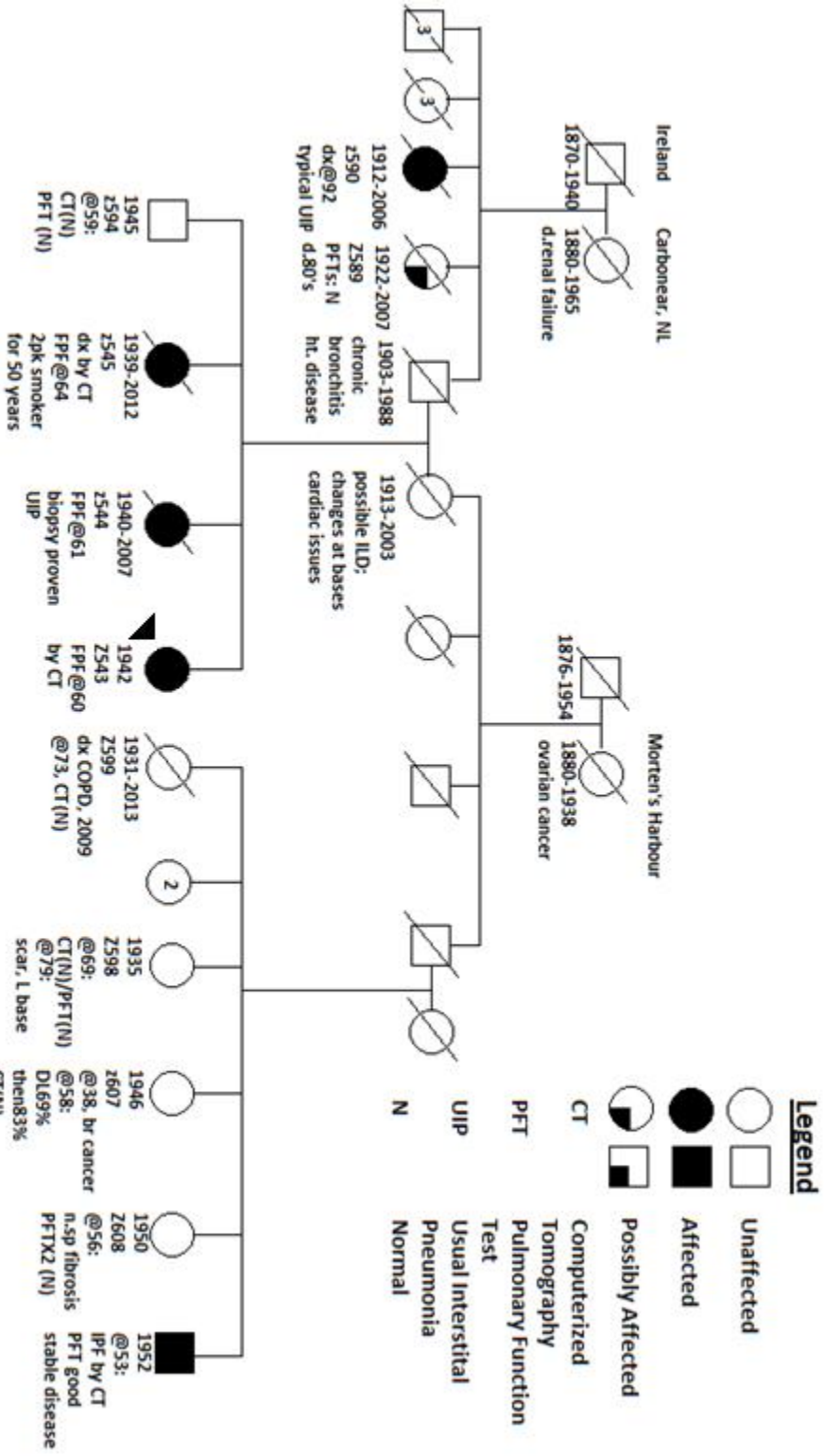
Appendix F: Variants Investigated in Newly Associated Idiopathic Pulmonary Fibrosis Loci

Gene	Variant	Segregation	MAF	Bioinformatics	Eliminated
DSP	c.-1insA	No R1136, 2/3	0.19	Polyphen: Benign GERP: Conserved	Yes
FAM13A	c.2305C>T	Yes R1136, 3/3	0.26	Polyphen: Benign GERP: Moderately Conserved	Yes
OBFC1	c.743C>G	No R1136, 1/3	0.17	Polyphen: Benign GERP: Moderately Conserved	Yes
SFTPA1	c.56T>C	No R1136, 1/3	0.11	Polyphen: Benign SIFT: Benign	Yes
SFTPA2	c.0.667G>T	No R0942, 2/3	0.25	Polyphen: Benign GERP: Non-Conserved	Yes
SFTPA2	c.26G>T	No R0942, 1/3 No R1136, 1/3	N/A	Polyphen: Benign SIFT: Benign	Yes
SFTPA2	c.271G>C	No R1136, 1/3	0.15	Polyphen: Benign GERP: Moderately Conserved	Yes
SNRNP48	c.807G>A	No R0942, 2/3 Yes R1136, 3/3	0.29	Polyphen: Unknown GERP: Non-conserved	Yes
SSR1	c.83T>C	No R0942, 2/3	0.33	Polyphen: Benign GERP: Conserved	Yes

Appendix G: Pedigrees of Four of Fourteen Families sent for Whole Exome Sequencing



Appendix H: Clinical Pedigree of Family R1136



Appendix I: Variants Uncovered in the High Impact Filtering List that Passed Filtering Criteria for All Families

Genes of Interest from High Impact List in R0896

Gene	Chromosome	Position	Gene Information	Variant tion
ADAMTSL1	9p22.3	18826261	Disintegrin and metalloproteinase-like protein (lacks metalloproteinase and disintegrin-like domain); Involved in ECM	Del TT Intron 21-22 *rs75054093* MAF: N/A
CAPN11	6p21.1	44145310	Calcium dependant cysteine proteases; Involved in cytoskeleton remodeling, altering subcellular localization;	C>T Intron 12-13 *rs4714765* MAF: 0.21 (A)
DEFB126	20p13	126310	Important in immunologic response to invading microorganisms	Del CC Exon 2 *rs140685149* MAF: N/A
DEFB126	20p13	126155		Del CAAA Exon 2 *rs140685149* MAF: N/A
DUOX1	15q21.1	45457127	Involved in antimicrobial defense at the mucosal surface; Prominent in airway epithelial cells; Dual oxidase 1; Associated with MS and sarcoidosis; mRNA greatly expressed in fibrofatty lesions; Critical role in mucin expression in airway epithelial; Regulated expression of DUOX1 during alveolar maturation; Role in microbial defense	C>T Exon 35 *rs1769193* MAF:0.31

EGFR	7	55214348	Epidermal growth factor receptor; Ubiquitously expressed; associated with non-small cell lung carcinoma; may play a role in the malformation of pulmonary airways	C>T; Synonymous variant Exon 4 *rs2072454* MAF:0.45
EMR1	19p13.3	6897464	Epidermal growth factor-like module receptor; May be involved in cell-cell interaction; May be involved in mucin interaction	C>G Exon 5 *rs330880* MAF: 0.47
GAB4	22q11.1	17469049	Growth factor receptor bound protein	C>A Nonsense variant Exon 3 *rs28502153* MAF: 0.34
HLA-DRB5	6p21.3	32487426	Protein component in antigen presenting complex; Major histocompability complex; Involved in immune response; Found in extracelleular space	C>A/G Exon 3 *rs1071751* MAF: 0.51
LRRK1	15q26.3	101601367	Forms a protein complex with EGFR	Del TTAC Intron 29-30 *rs148929418* MAF: N/A
MREIIA	11q21	94225807	Involved in telomere length maintenance, homologous recombination	C>T *rs496797* MAF: 0.47

MUC3A	7q22.1	100552738	Mucin protein; Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung, involved in epithelial structure maintenance	C>T *rs79874934* MAF: NA
MUC19	12q12	40820208	Mucin protein; Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung, involved in epithelial structure maintenance	G>A Intron 10-11 *rs11176575* MAF: 0.43
NLRC5	16q13	57095842	Role in cytokine response and antiviral immunity through inhibition of NF-kappa B activation	Del GAAA Intron 32-33 *rs142124514* MAF: N/A

Genes of Interest from High Impact List in R1351

Gene Name	Chromosome	Position	Gene Information	Variant Information
C5	9	123787728	Immune inflammation; absence results in scarring; associated with liver fibrosis and asthma	Ins ACACACACAC Predicted splice site variant (intronic) *rs71370618* MAF:N/A
AOAH	7	36552793	Acyloxyacyl hydrolase; removes secondary acyl chain in lipid A bacterial liopolysaccharides; immune response; associated with asthma	Ins T Exon 27 Not previously reported
COL6A1	21	47420240	Collagen gene; associated with multiple sclerosis; stimulator in epithelial cell proliferation during wound healing	C>CT Not previously reported
COL13A1	10	71647208	Involved in cell-matrix, cell-cell adhesion; may be involved in linking muscle fiber to basement membrane; role in branching morphogenesis in lung; widely expressed in ocular tissue	Del T Intronic splice site variant *rs201577179*
SON	21	34948686	SON DNA binding protein; indispensible growth factor; might protect cells from apoptosis; widely expressed (classified under lung neoplasia in DAVID)	Ins A Exonic splice site variant *rs34377180* MAF:N/A

MMRN1	4	90875430	Multimerin: found in platelets and endothelial vessels; may act as a carrier protein for platelet factor V; Functions in ECM and adhesion; associated with autosomal bleeding disorder	Del T Not previously reported
MUC21	6	2302447	Mucin protein; highly expressed in the lung	T>C Not previously reported

Genes of Interest from High Impact List in R1487

Gene	Chromosome	Position	Gene Information	Variant Information
ADAMTSL1	9p22.3	18826261	Disintegrin and metalloproteinase-like protein (lacks metalloproteinase and disintegrin-like domain); Involved in ECM	Del TT Intron 21-22 *rs75054093* MAF: N/A
CLEC1	12p13.31	9885707	Type II Transmembrane; May act as a T-cell co-stimulatory molecule that enhances interleukin-4 production and may become involved in the regulation of the immune system and response; Involved in cell-cell interaction	INS TAAGT Not reported
DEFB126	20p13	126310	Important in immunologic response to invading microorganisms	Del CC Exon 2 *rs140685149* MAF: N/A
DUOX1	15q15.3	45457127	Dual oxidase 1; Associated with MS and sarcoidosis; mRNA greatly expressed in fibrofatty lesions; Critical role in mucin expression in airway epithelial; Regulated expression of DUOX1 during alveolar maturation; Role in microbial defense	C>T Exon 3 *rs1769193* MAF:0.31
EPDR1	7p14.1	37960262	May be involved in calcium dependant cell adhesion	Del AGCAGGC AGTGGC Exon 1 *rs201513905* MAF:0.21

GSDMB	17	38064469	Gasdermin B; Localization in apical region of gastric chief cells and colonic surface mucous cells; Polymorphisms may contribute to childhood asthma	T>C Intron 4-5 *rs11078928* MAF 0.33
GZMB	14q11.2	25103414	Encodes enxyme responsible in cell lysis during cell-mediated immune response; apoptosis execution	G>A *rs22738414* MAF:0.28
HYDIN	16q22.2	71100838	Gene encodes protein involved in cilia motility	T>C Intronic variant *rs74361942* MAF:0.28
HLA-A	6	29911240	Protein component in antigen presenting complex; Major histocompability complex; Involved in immune response; Found in extracelleular space	T>A/G
HLA-DRA	6	32411035		A>C
HLA-J	6	29977145		G>C
HLA-P	6	29759885		G>A
IL17RB	3p21.1	53899276	Cytokine receptor; Mediates activation of NF-kappa; May be upregulated during intestinal inflammation; Immunoregulartory activity	C>T Exon 11 *rs1043261* MAF: 0.11
ITGB2	21.q22.3	46328099	Integrin protein; involved in cell adhesion	C>T; *rs760462* MAF: 0.14
ITIH1	3p21.1	52821992	May serve as a carrier for hylauran, protein involved in extra cellular matrix	Del T Intron 16-17 *rs68094128* MAF: 0.37
LAIR2	19q13.4	55019261	Member of the immunoglobulin superfamily; Thought to help modulate mucosal tolerance	C>T

LCE1D	1q21.3	152770613	Encodes the precursor to cornified envelope of the stratum corneum (outer layer of epidermis)	T>G Exon 2 *rs41268500* MAF: 0.05
MREIIA	11q21	94225807	Involved in telomere length maintenance, homologous recombination	C>T *rs496797* MAF:0.47
MUC19	12q12	40820208	Mucin protein; Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung, involved in epithelial structure maintenance	G>A Intron 10-11 Splice site acceptor variant *rs11176575* MAF: 0.43
MUC3A	7q22.1	100552738		C>T *rs79874934* MAF: NA;
MUC19	12q12	40837264		T>C
NOTCH2	1p13-p11	120612006	Contains structural similarities to extracellular domain proteins; EGF repeats; plays a role in cell fate; Involved in remodelling	G>A Not reported
NOTCH4	6p21.3	32191658		Del CAG
NPSR1	7p14.3	34889222	Member of G-coupled receptor, plasma membrane protein; Increased expression associated with asthma	T>C Synonymous variant Exon 9 *rs10275028* MAF: 0.36
NPSR1	7p14.3	34917740		C>T *rs7809642* MAF: 0.2
PAPLN	14q24.2	73730754	Encodes papilin, an extracellular matrix glycoprotein	T>C Not reported
PCSK5	9q21.3	78790207	Proprotein convertase; processes integrins, type1 –matrix metalloproteinase	C>A Intronic variant *rs10124596* MAF: NA

PCSK5	9q21.3	78790217		A>C Not reported
TMPRSS11A	4q13.2	68829109	Expressed in trachea, localized to ECM; involved in proteolysis, cell cycle senescence and cycle arrest	C>T *rs977728* MAF: 0.16
TP53INP1	8q22	95951916	Response to cell stress, anti-proliferative and pro-apoptotic	Del ACTG Not reported

Genes of Interest from High Impact List in R1224

Gene	Chromosome	Position	Gene Information	Variant tion
CAPN11	6p21.1	44145310	Calcium dependant cysteine proteases; Involved in cytoskeleton remodeling, altering subcellular localization;	C>T Intron 12-13 *rs4714765* MAF: 0.21 (A)
DEFB126	20p13	126310	Important in immunologic response to invading microorganisms	DelCC Exon 2 *rs140685149* MAF: N/A
DEFB126	20p13	126155		DelCAAA Exon 2 *rs140685149* MAF: N/A
EGFR	7	55214348	Epidermal growth factor receptor; Ubiquitously expressed; associated with non-small cell lung carcinoma; may play a role in the malformation of pulmonary airways	C>T; Synonymous variant Exon 4 *rs2072454* MAF:0.45
HLA-A	6	29911240	Protein component in antigen presenting complex; Major histocompability complex; Involved in immune response; Found in extracellular space	T>A/G *rs9260156* MAF: 0.27
HLA-A	6	29912028		DelG *rs66729206* MAF: 0.37
LRRK1	15q26.3	101601367	Forms a protein complex with EGFR	Del TTAC Intron 29-30 *rs148929418* MAF: N/A
MREIIA	11q21	94225807	Involved in telomere length maintenance, homologous recombination	C>T *rs496797* MAF: 0.47

MUC3A	7q22.1	100552738	Mucin protein; Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung, involved in epithelial structure maintenance	C>T *rs79874934* MAF: NA
MUC6	11p15.5	1017041	Mucin protein; Glycoprotein component of mucus gels; Provides protective barrier against infectious particles; associated with ulcerative colitis, arthritis and renal carcinogenesis; not specific to lung, involved in epithelial structure maintenance	G>T Not reported
PCSK5	9q21.3	78790207	Proprotein convertase; processes intgerins, type1 –matrix metalloproteinase	C>A Intronic variant *rs10124596* MAF: NA
PCSK5	9q21.3	78790217		A>C Not reported